Distributional Semantic Models of Attribute Meaning in Adjectives and Nouns

Von der Neuphilologischen Fakultät der Ruprecht-Karls-Universität Heidelberg
zur Erlangung der Würde eines Dr. phil. genehmigte Abhandlung

Vorgelegt von

Matthias Hartung

aus Aschaffenburg

Hauptberichter:	Prof. Dr. Anette Frank
Mitberichter:	Prof. Dr. Sebastian Padó
Tag der mündlichen Prüfung:	11. Dezember 2015

Institut für Computerlinguistik
Ruprecht-Karls-Universität Heidelberg

2015

Contents

Abstract

Attributes such as SIZE, WEIGHT or COLOR are at the core of conceptualization, i.e., the formal representation of entities or events in the real world. In natural language, formal attributes find their counterpart in attribute nouns which can be used in order to generalize over individual properties (e.g., *big* or *small* in case of SIZE, *blue* or *red* in case of COLOR). In order to ascribe such properties to entities or events, adjective-noun phrases are a very frequent linguistic pattern (e.g., *a blue shirt*, *a big lion*). In these constructions, attribute meaning is conveyed only implicitly, i.e., without being overtly realized at the phrasal surface.

This thesis is about modeling attribute meaning in adjectives and nouns in a **distributional semantics** framework. This implies the acquisition of meaning representations for adjectives, nouns and their phrasal combination from corpora of natural language text in an unsupervised manner, without tedious handcrafting or manual annotation efforts. These phrase representations can be used to predict implicit attribute meaning from adjective-noun phrases – a problem which will be referred to as **attribute selection** throughout this thesis.

The approach to attribute selection proposed in this thesis is framed in structured distributional models. We model adjective and noun meanings as distinct semantic vectors in the same semantic space spanned by attributes as dimensions of meaning. Based on these word representations, we make use of vector composition operations in order to construct a phrase representation from which the most prominent attribute(s) being expressed in the compositional semantics of the adjective-noun phrase can be selected by means of an unsupervised selection function. This approach not only accounts for the linguistic principle of **compositionality** that underlies adjective-noun phrases, but also avoids inherent sparsity issues that result from the fact that the relationship between an adjective, a noun and a particular attribute is rarely explicitly observed in corpora.

The attribute models developed in this thesis aim at a reconciliation of the conflict between **specificity** and **sparsity** in distributional semantic models. For this purpose, we compare various instantiations of attribute models capitalizing on pattern-based and dependency-based distributional information as well as attribute-specific latent topics induced from a weakly supervised adaptation of Latent Dirichlet Allocation. Moreover, we propose a novel framework of **distributional enrichment** in order to enhance structured vector representations by incorporating additional lexical information from complementary distributional sources. In applying distributional enrichment to distributional attribute models, we follow the idea to augment structured

representations of adjectives and nouns to centroids of their nearest neighbours in semantic space, while keeping the principle of meaning representation along structured, interpretable dimensions intact.

We evaluate our attribute models in several experiments on the attribute selection task framed for various attribute inventories, ranging from a thoroughly confined set of ten core attributes up to a **large-scale set of 260 attributes**. Our results show that large-scale attribute selection from distributional vector representations that have been acquired in an unsupervised setting is a challenging endeavor that can be rendered more feasible by restricting the semantic space to confined subsets of attributes. Beyond quantitative evaluation, we also provide a thorough analysis of performance factors (based on linear regression) that influence the effectiveness of a distributional attribute model for attribute selection. This investigation reflects strengths and weaknesses of the model and sheds light on the impact of a variety of **linguistic factors** involved in attribute selection, e.g., the relative contribution of adjective and noun meaning.

In conclusion, we consider our work on attribute selection as an instructive showcase for applying methods from distributional semantics in the broader context of **knowledge acquisition from text** in order to alleviate issues that are related to implicitness and sparsity.

Acknowledgements

"In Anbetung kniee ich vor mir."[1]
Dieter Hildebrandt (1927-2013)

Here you go. Hildebrandt's quote is the precise answer to all those who have been asking how it feels after having completed your dissertation. On second thought, however, after a moment of reflection and back on my feet again, I realize that this thesis has been influenced by many more people than just myself. Over the last years, they provided me with encouragement, inspiration and support in various ways.

Anette Frank has supervised this thesis by granting me the freedom to identify and pursue my own research interests, showing an immense personal dedication to my work, being an endless source of ideas, and occasionally pushing and pulling me into the direction of constant progress. Admittedly, it was sometimes hard to meet her quality standards; on the other hand, her strong commitment to scientific rigour has shaped this thesis considerably and will remain a guideline for me.

Sebastian Padó served on my doctoral committee and influenced many of my perspectives on distributional semantics in particular and various topics from computational linguistics in general during numerous inspiring discussions we had while sharing the fate of commuting on the train between Heidelberg and Stuttgart.

Christiane Fellbaum provided invaluable help by motivating her students at Princeton to annotate the HeiPLAS data set. Without their huge efforts, a thorough evaluation of my experiments would have been so much harder, if not impossible. The same holds for all annotators from Heidelberg who contributed to the creation of various other data sets: **Carina Silberer**, **Sascha Fendrich**, **Johannes Knopp** and **Wolodja Wentland**. **Patrick Simianer**, **Hiko Schamoni** and **Markus Kirschner** did an excellent job in administering and maintaining the ICL compute cluster environment, which saved me a lot of computing time during my experiments.

In the Computational Linguistics Colloquium at Heidelberg and on various other occasions, I was lucky to receive very instructive feedback and comments from **Stefan Riezler**, **Michael Strube**, **Mirella Lapata**, **Ed Hovy**, **Ido Dagan**, **Rainer Osswald** and **Ian Witten**, to name just a few. Also, it was a great pleasure to see many of **the students in my seminars** taking an active interest in distributional semantics and engaging in lots of lively discussions. I always felt their interest as a strong encouragement for my own work.

[1] My own rough translation: *In adoration I kneel down before myself.*

After joining his Semantic Computing Group at Bielefeld, **Philipp Cimiano** granted me an enormous freedom to finish up this thesis, at the same time considering me a "quasi-postdoc" from the very beginning of our collaboration. Philipp never exerted any pressure, but showed an extremely rare mixture of everlasting patience, trust and sympathy in all matters of scientific and real life.

Philipp was also among the first people, back when I was still undergrad, who encouraged me in my vague and wacky ideas to pursue a dissertation, as well as **Peter Hellwig**, **Rolf Kailuweit**, **Klaus-Peter Konerding** and **Markus Demleitner**.

Nils Reiter was my first colleague at ICL at that time and almost instantaneously became a like-minded friend. During the years we have spent together, Nils brightended so many of my days (and evenings!) with humorous and musical interludes in his famous *Rockbüro*[2], deep thoughts in order to revolutionize semantics (*ontospiel*!), countless sessions of our not-so-scientific *Doktorandenstammtisch*, the *Kulturkalender* and – I don't know where to end.

Likewise, I have wonderful memories of the regular gatherings "on the sundeck", where all the folks from ICL who carried semantics in their mind (or, some of them, secretly in their heart) used to meet over coffee in order to discuss all the bigger and smaller questions in a Ph.D. student's daily life. **Michael Roth**, **Sascha Fendrich**, **Britta Zeller**, **Eva Sourjikova**, **Gerhard Kremer** and **Hiko Schamoni** were the most regular participants in this circle. Anyway – some members of the **Coliktiv** were (literally) always available for fun and relaxation and made Heidelberg such a particular, familial place to dive into computational linguistics. *Always Coli!*

In regular meetings of the "Herrenrunde" (and yet all too seldom), **Tobias Federwisch**, **Lars Vogel**, **Hendrik Ehrhardt** and **Matthias Heise** offered me interesting shifts of perspective into the social sciences and humanities, which opened me an entire universe of new inspiration and... – well, most importantly, they are just great friends.

My parents, in their particular way, always gave me the feeling of taking a lively interest in my work. Almost naturally, they managed so many things behind the scenes and were always there when I needed them.

Finally, there is one person who always understood – when I felt only one more nightshift away from the breakthrough (once again!), and when I hid behind walls of self-irony and grumpiness the day after, who calmly took all my mockeries of syntax, and who sometimes even managed to distract me from work: **Sina**, sine qua non.

I would like to express my sincerely felt gratitude to all of you.

[2]Even more famous than the *stylebureau* which I intermittently shared with **Simone Ponzetto**!

1 Introduction

This thesis aims at automatically interpreting adjective-noun phrases with respect to their attribute meaning. Using a distributional semantic framework, we acquire meaning representations for adjectives, nouns and their phrasal combination from corpora of natural language text in an unsupervised manner. These distributional representations are designed as to give insight into attribute meanings that are implicitly conveyed by the compositional semantics of adjective-noun phrases (e.g., the attribute COLOR being conveyed by the phrase *blue shirt*, or STRENGTH being conveyed by *strong lion*). Thus, they facilitate automatic exploitation of the rich source of hidden attribute knowledge that is provided by the ubiquity of adjective-noun phrases in natural language.

1.1 Life Cycle of Knowledge in Natural Language Processing

Knowledge creators and knowledge consumers in NLP. The high relevance of ontological knowledge for applications in Natural Language Processing has prevailed even after the statistical turn and the machine-learning turn , as can be seen from examples in several fields as diverse as question answering (Unger et al., 2012; Cimiano et al., 2014), information retrieval (Fernández et al., 2011), word sense disambiguation (Ponzetto and Navigli, 2010; Agirre et al., 2014) or coreference resolution (Ponzetto and Strube, 2006; Rahman and Ng, 2011), among others.

From this spectrum, a bidirectional interaction between NLP applications and ontological knowledge can be construed, regarding NLP systems as *knowledge consumers* or *knowledge creators* (Lenci, 2010). The former refers to knowledge-rich systems directly responding to a user's information need. Depending on the complexity of the particular task, such systems may integrate a variety of different knowledge sources, as has been recently demonstrated by IBM's *Watson* system for "deep question answering" (Brown et al., 2013; Ferrucci et al., 2013). On the other hand, knowledge creators perform NLP tasks in order to generate knowledge for these knowledge consumers. The division between creators and consumers is not always clear-cut, as knowledge creators may depend on initial knowledge themselves. Therefore, the issue well-known as the *knowledge acquisition bottleneck* in Artificial Intelligence (Gruber and Cohen, 1987) applies to both of them: *Where does the knowledge come from?*

Overcoming the knowledge acquisition bottleneck. Given that knowledge continuously accumulates and changes over time, tedious and time-consuming handcraft-

ing is clearly not a viable solution for creating and maintaining domain-specific and up-to-date knowledge resources. Consequently, as argued by Lenci (2010), *automatic* knowledge acquisition methods should be given preference, as they bear the potential to establish a "life cycle of knowledge" by tightly interlinking knowledge creators and knowledge consumers. To this end, large amounts of text are used as the raw material from which knowledge resources are automatically created and populated. Subsequently, these knowledge resources are available to be used in practical NLP applications, i.e., knowledge consumers or linguistically enhanced knowledge creators.

1.2 Knowledge Induction from Text

Turning natural language text into machine-interpretable knowledge requires a commitment to an "explicit specification of a shared conceptualization of a domain of interest" in terms of a formal ontology (Gruber, 1993). Defining the common vocabulary in which shared knowledge may be formally represented, ontologies provide a mapping from possibly ambiguous natural language terms to formal concepts (Gruber, 1993). In response to the knowledge acquisition bottleneck, the last years have seen the emergence of *ontology learning from text* (Cimiano, 2006). At the core of ontology learning is the task of relation extraction, i.e., detecting semantically meaningful relations between previously determined or known (e.g., by referring to an already existing ontology) domain-specific concepts.

Relation extraction. There is a wide range of approaches to relation extraction. At one pole, pre-defined individual semantic relations are harvested using pattern-based techniques in the tradition of Hearst (1992)[1]. In these approaches, ontological relations of interest have to be specified in advance by manually providing possible linguistic manifestations (e.g., *is born in* for the BIRTHPLACE relation holding between persons and locations, or *works for* for the EMPLOYEE relation between persons and companies). Pattern-based approaches offer high precision for an often considerable amount of manual specification efforts. At the other end of the spectrum, *open information extraction* approaches (Banko et al., 2007; Etzioni et al., 2008) do not rely on pre-defined extraction schemas, but extract all occurrences of particular syntactic configurations in natural language text that are likely to convey different types of semantic relations (e.g., transitive verb constructions). Thus, open information extraction clearly overcomes the rigidity of pattern-based approaches, at the expense of a lack of ontological grounding.

Drawbacks of surface-related approaches. Sticking closely to the linguistic surface, both of these paradigms suffer from two major problems, viz. from (i) sparsity and (ii)

[1] In the seminal work of Hearst (1992), so-called *is-a patterns* (e.g., *car is a vehicle*) have been used in order to detect taxonomic relations. In more recent work, pattern-based approaches have been extended to non-taxonomic relations as well (Lin and Pantel, 2001; Pantel and Pennacchiotti, 2006).

a lack of semantic generality beyond individual surface patterns.

By the term *sparsity* we refer to a problem that is very common in corpus-based modeling of natural language phenomena, i.e., the large number of rare or even unobserved events (Manning and Schütze, 1999). As a consequence of sparsity, purely surface-related approaches to relation extraction are notoriously incapable of acquiring all possible instantiations of a semantic relation.

The *generality* problem is mainly due to lexical and syntactic alternations at the levels of the information need specified by the user and the textual surface (e.g., *kill* vs. *murder*; active vs. passive voice). Being out of the scope of pattern-based or open information extraction methods, both these issues have only been addressed by grouping instances of semantically similar surface patterns together (Wang et al., 2011; Lewis and Steedman, 2013). Their clusters or latent variables, respectively, do not provide an obvious link to information needs explicitly specified by users, though.

In this thesis, we will explore ways of pursuing knowledge acquisition within a distributional semantic framework that allows for various adaptations in order to deal with sparsity in particular. To this end, we will propose distributional semantic models that incorporate aspects of compositionality and integrate latent variable models.

1.3 Attribute Knowledge in Knowledge Consumers and Knowledge Creators

Abstraction in knowledge-consuming information systems. As a desideratum for future developments in relation extraction, we envisage a semantic layer of *abstraction* over individual relations that is accessible for intuitive specification by users, based on abstract natural language terms. As a knowledge-consuming application in this direction, we anticipate *attribute-based information retrieval*: Imagine a user being interested in evidence on whether or not HIV/AIDS is a dangerous disease or whether Heidelberg is an attractive city to live in, for instance. These information needs may be presented to the envisaged information system in terms of abstract queries as in (1):

(1) a. DANGER of the disease HIV/AIDS
    b. ATTRACTIVENESS of the city Heidelberg

We argue that *attribute nouns* such as *danger* or *attractiveness* are of high value for this purpose, as they are often ontologically grounded and provide a linguistic interface to applications addressing information needs specified by users in natural language.

(2) *...interferes more and more with the immune system...*
...making the patient much more likely to get infections inlcuding tumors...
...no cure or vaccine...
...treatment reduces risk of death...

(3) *...situated on the River Neckar...*
...fifth-largest town in Baden-Württemberg...
...popular tourist destination...
...romantic and picturesque cityscape...
...baroque-style Old Town...

In (2) and (3), we list several relevant results[2] responding to these queries. Note that none of these results contains an explicit mention of the query terms *danger* or *attractiveness*; they rather consist of individual semantic relations which can be subsumed under the attributes provided in the query. It is the task of the underlying knowledge acquisition methods to induce a link between the textual surface and the semantic layer of abstraction that is referred to in the query in terms of attribute nouns.

A similar problem is encountered in *aspect-based sentiment analysis* (Pontiki et al., 2014) aiming at the identification of aspects of given target entities and the sentiment expressed for each of them. While these aspects often denote ontological attributes of the target entities (e.g., the TASTE of a meal, the PRICE of a car, etc.), textual descriptions rarely mention them explicitly in terms of their corresponding attribute nouns.

A showcase life cycle of attribute knowledge. The relevance of attribute knowledge is not limited to knowledge-consuming applications. In fact, we can imagine a full-fledged *life cycle of attribute knowledge* in various sub-fields of NLP. For illustration, we consider coreference resolution as an example case:

(4) [The box]$_{m_1}$ could not be moved. Due to [its heavy weight]$_{m_2}$, it remains behind the door.

(5) There are [two boxes]$_{m_1}$ in the room. The [red box]$_{m_2}$ is under the chair. The [blue box]$_{m_3}$ has a yellow ribbon.

On the one hand, coreference resolution systems may contribute to the creation of attribute knowledge, in that coreferent entities within a discourse enable the acquisition of more meaningful and accurate attribute profiles. In (4), this is demonstrated for the relation between the concept *box* and the attribute WEIGHT, which can only be established if the coreference between m_1 and m_2 is successfully resolved.

On the other hand, knowledge-rich coreference resolution systems (Rahman and Ng, 2011) may benefit from a specific constraint rooted in attribute knowledge[3] given in (6):

(6) If two entities exhibit different values for the same attribute in the same discourse, they are unlikely to be coreferent.

This is illustrated in (5), where both m_2 and m_3 are coreferent with m_1. However, m_2 and m_3 are not coreferent with each other, as *red box* and *blue box* denote different values

[2] Manually extracted from the Wikipedia pages `http://en.wikipedia.org/wiki/HIV/AIDS` and `http://en.wikipedia.org/wiki/Heidelberg`, respectively; both last accessed on October 28, 2014.

[3] Thanks to Michael Strube (p.c.) for pointing this out.

for the same attribute COLOR. Taken together, these examples illustrate the potential of knowledge-centered NLP architectures to enable a continuous life cycle based on the entanglement of knowledge creators and knowledge consumers. Given their abstraction potential and their ubiquity in language and ontology, a vital role in all stages of this life cycle may be played by attributes.

1.4 Attribute Meaning

Attributes are important in the fields of cognitive science, knowledge representation and linguistics. We briefly review the basic notion of *attribute* in each of these fields here, with an emphasis on the particular aspects that will be of relevence for the work presented this thesis.

Cognitive science. In cognitive science, there is broad consensus that conceptual representations in the human mind are not *atomic*, but involve a mechanism of decomposition of concepts into *properties* or *features* (Baroni et al., 2010). Properties may denote any characteristics that qualify an instance of the respective concept. Properties describing members of the concept *bird*, for example, may include *feathers*, *flies*, *lays eggs* and *nest*, among many others (Barsalou, 1992).

The internal structure of concepts, however, is the cause of an intense, long-standing debate. At one end of the spectrum, there are theories assuming that concept descriptions merely consist of *feature lists* (Rosch and Mervis, 1975) with no structural difference being postulated between the singular properties in the list. The example properties given above for *bird*, for instance, comprise parts, activities and associated concepts, without accounting for these semantic differences (cf. Poesio and Almuhareb, 2005).

In contrast, Barsalou (1992) assumes a *frame structure* being superimposed on the properties describing a concept, arguing that "rather than categorizing entities solely on the basis of specific values, people often categorize them on the basis of more abstract attributes". The building blocks of frames are sets of *attributes* and their possible *values*, with values corresponding to such properties. Consider, for example, a possible frame representation of the concept *bird* that might contain the attributes COLOR and SIZE, among others, together with *yellow* and *small* as their corresponding values.

Linguistics. Natural language refers to attributes in terms of *attribute nouns* such as *color*, *size* or *shape*. According to Guarino (1992), attribute nouns are ambiguous with respect to functional and sortal readings in the sense that they may either denote "conceptual components of something as well as concepts on their own". In the latter case, the noun has to be given a *sortal* reading as in (7b), contrary to the functional interpretation in (7a).[4]

[4]These examples are due to Löbner (2013).

(7) a. The color of the potato is purple.
b. Purple is a color.

In order to disambiguate these readings in a particular linguistic context, Guarino (1992) suggests the following rule that he refers to as *Woods' linguistic test*[5]:

(8) Y is a value of the attribute A of X if we can say that Y is an/the A of X.

Attribute nouns may denote sets (Löbner, 2013) or scales of values (Sheinman et al., 2013; Hatzivassiloglou and McKeown, 1993; Levinson, 1983); hence they can be used as "abstraction terms" in order to generalize over individual properties of entities or events. Notably, attribute meaning in unrestricted natural language text is most often conveyed by adjectives in predicative or attributive syntactic configurations, as illustrated in the following examples:

(9) The car is blue. (COLOR of the car)

(10) This is an old man. (AGE of the man)

Alternatively, attributes may be conveyed in noun compounds or by means of stative verbs (Gamerschlag, 2008):

(11) This is a stone wall. (MATERIAL of the wall)

(12) These shoes cost 100 Euro. (PRICE of the shoes)

Note that in all these constructions, attribute meaning is conveyed only *implicitly*, i.e., without being overtly realized at the linguistic surface. In order to render implicit attribute meaning explicit in an automatic corpus-based approach, lexico-syntactic patterns similar to Woods' test in (8) can be used. However, such patterns are very sparse in natural language text, which is one of the main motivations for the distributional approach to modeling attribute meaning taken in this thesis.

Knowledge representation. To some extent, attribute information is contained in existing semantic resources commonly used in NLP – most notably WordNet (Fellbaum, 1998), but also SUMO (Pease et al., 2002; Niles, 2003) or Cyc (Lenat, 1995; Matuszek et al., 2006). However, these resources face limitations in that they cover attribute information on the lexical or conceptual level only: In WordNet and SUMO, attribute information is restricted to (attributes of) adjectives, i.e., no attribute knowledge is provided for nouns. Cyc features attribute information on the concept level, but lacks lexical coverage as the only existing mapping between Cyc and WordNet merely accounts for a small subset of WordNet synsets (Scheffczyk et al., 2006).

As a consequence, neither of the existing resources is capable of assigning attribute meaning to adjectives and nouns in context. This is clearly a shortcoming, given that attribute meaning in natural language is most often implicitly conveyed in adjective-noun phrases as in (9) and (10).

[5]Note that this test yields a positive result only for functional readings as in (7a), whereas sortal readings as in (7b) can be determined only by exclusion.

Main topics of this thesis. In this thesis, we treat the acquisition of attribute knowledge as a special case of knowledge induction from text, with a focus on adjective-noun phrases as ubiquitous sources of attribute knowledge in natural language. We will follow an unsupervised induction approach that is embedded into a distributional semantic framework. This enables us to acquire attribute knowledge for adjectives, nouns and their phrasal combinations from text corpora without tedious handcrafting or manual annotation efforts. We will show that distributional methods effectively alleviate issues in attribute acquisition that are related to implicitness and sparsity.

Our methodology can be flexibly adapted to various domains with varying inventories of attributes. We argue that this latter point is of major importance both from a cognitive and a practical perspective: As Barsalou (1992) points out, attribute formation is a very productive process: "People are highly creative in their construction of attributes, often producing new ones relevant to specific contexts". This clearly underlines the importance of an automated acquisition approach beyond static lexical or ontological resources.

The approach followed in this thesis does not rely on a particular existing ontology as backbone. Being primarily motivated from the linguistic abstraction potential of attributes, it remains close to the linguistic surface, assuming a collection of attribute nouns as its only requirement to be specified in advance. In the interest of being able to explore a broad, large-scale inventory of attribute meaning, we rely on attribute nouns as provided by WordNet (Fellbaum, 1998). Thus, we avoid translation efforts between ontology and language (Gruber, 1993) during acquisition. Linking the acquired attribute knowledge to a particular ontology is left as a subsequent task in this work, which may be achieved manually or automatically, using techniques from entity linking or word sense disambiguation, for instance Moro et al. (2014).

1.5 Thesis Overview

The thesis is structured as follows. Subsequently to this introduction, Chapter 2 outlines the foundations of distributional semantics most relevant in order to put the models proposed in this thesis into a broader context. This includes conceptual and notational foundations and different variants of distributional semantic models along with their particular assumptions. Most importantly, we will focus on the difference between structured vs. unstructured distributional models and introduce the notions of *specificity* and *sparsity* which relate to conflicting goals in distributional modeling.

Chapter 3 reviews previous work related to the most important aspects of this thesis, i.e., attribute learning, adjective classification, structured distributional models, topic models in distributional semantics, distributional models of phrase meaning and distributional enrichment of structured models.

Chapter 4 summarizes the most important research questions being addressed in this thesis and its major contributions. In this chapter, we define the task of *attribute*

selection from adjective-noun phrases, which constitutes the core of our approach to the acquisition of attribute knowledge from text, and we outline our methodology in addressing this task.

Chapter 5 restricts the subject matter to be investigated in this thesis by confining attribute selection to a subclass of adjectives, i.e., property-denoting lexical types. To this end, we propose an adjective classification scheme that is empirically validated in a corpus annotation study and an automatic classification experiment.

Chapter 6 presents our attribute selection models in full technical detail, including different configurations and parameters. This includes pattern-based and dependency-based distributional models as well as distributional attribute models incorporating attribute-specific latent topics as induced from weakly supervised variants of Latent Dirichlet Allocation.

In Chapter 7, the different attribute selection models are subjected to a contrastive experimental evaluation against specifically created data sets. These reflect two scenarios of attribute selection being carried out (i) on a confined set of core attributes and (ii) on a large scale.

In Chapter 8, the best-performing attribute model as identified in these experiments undergoes a thorough performance analysis based on linear regression models. This study assesses the impact of various linguistic variables on attribute selection, exposing strengths of our model as well as potentials for improvement.

In Chapter 9, we present a framework for *distributional enrichment* that aims at improving structured distributional representations. The framework is evaluated in an experiment in order to assess its capabilities in enhancing the performance of distributional attribute models in large-scale attribute selection.

Chapter 10 concludes the thesis by summarizing its main contributions and highlighting perspectives for future work.

2 Foundations of Distributional Semantics

2.1 Distributional Hypothesis

Distributional modelling of lexical meaning is based on the fundamental assumption known as the *distributional hypothesis*: "You shall know a word by the company it keeps." (Firth, 1957)

Following the late Wittgenstein's *meaning as use* conception (Wittgenstein, 2001), the distributional hypothesis states that the meaning of a word is largely characterized by its observed use in language, i.e., the words in the context of which it frequently occurs. As an important implication of the distributional hypothesis, words that occur in similar contexts tend to be similar in meaning (Harris, 1954; Firth, 1957; Deerwester et al., 1990).

In a more generalized form, "statistical patterns of human word usage can be used to figure out what people mean" (Turney and Pantel, 2010). In fact, human speakers are most often capable of guessing the meaning of a word solely based on its usage (Erk, 2012), which confirms the plausibility of the distributional hypothesis. This holds even for artificial non-words (not existing in the vocabulary of a particular language), as demonstrated by Baldwin (2006):

(13) a. *groyter*

b. *specialist, cause, patient, severity, biopsy*

(14) a. *hoonger*

b. *steep hoonger, hoonger top, climb a hoonger, wooded hoonger*

Here, the words in (13a) and (14a) are the target words whose meaning is to be determined. Assume that *groyter* is frequently observed to occur in documents also mentioning the words given in (13b), while *hoonger* frequently occurs in contexts as given in (14b). Even though neither *groyter* nor *hoonger* exist as real words of English, human speakers almost effortlessly infer that the former is meant to denote a disease and the latter a mountain or a hill.[1]

[1] I recommend to present these examples to students unfamiliar with distributional semantics, as I recurrently did in all the courses I have taught on the topic. Both examples worked smoothly for all participants (most of them non-native speakers of English), which made them immediately believe.

	drive	*park*	*engine*	*flower*
car	15	7	11	0
truck	13	4	16	0
tree	0	5	0	9

Figure 2.1: Vector representations for the target words *car*, *truck* and *tree*, using *drive*, *park*, *engine* and *flower* as context words. The numbers in the individual fields of each vector (being made up for this example) can be interpreted as weights denoting the strength of the relationship between the target and the respective context word.

Arguably, this inference process is due to a human capacity denoted as *similarity-based generalization* in cognitive psycholinguistics (Yarlett, 2008) that empowers language learners to acquire meaning representations for unknown words based on their contextual similarity to already learned ones, which is a direct reflection of the distributional hypothesis.

2.2 Meaning Representation in Distributional Semantic Models

In computational linguistics, the distributional hypothesis is prominent in *distributional semantic models* or *vector space models* (Turney and Pantel, 2010) which can be seen as implementations of automated "discovery procedures" (Sahlgren, 2008) for representations of word meaning. Contrary to formal semantic approaches in the Fregean tradition (cf. Portner, 2005), distributional representations do not rely on abstract logical forms (e.g., *car'* as a predicate for representing the meaning of *car*) which have to be resolved against tediously hand-crafted models (Carnap, 1947), but on purely corpus-based contextual cues for representing lexical meaning. The distributional approach towards modeling the meaning of a target word would be to automatically collect all context words (i.e., words contextually related to all mentions of the target in large text corpora) and store them in a vector. See Figure 2.1 for an example.

Context words may be seen as empirically determined features or meaning components contributing to the intensional meaning of a target word, similar to lexical descriptions developed in componential analysis (Dowty, 1979). However, this analogy to generativist approaches to lexical semantics should not be overstressed since distributional features, given their origin in the language system itself, cannot be claimed to be abstract and universal semantic primitives (Wierzbicka, 1996).

Within distributional semantics, the internal structure of word meaning (Pulman, 2005) is usually abstracted from, and so is the question as to whether contextual features reflect such an internal structure or not. More importantly, the word vectors constructed through the distributional "discovery procedure" can be given a geometrical

interpretation (Widdows, 2004) as points in a semantic space that is spanned by the context words as its dimensions of meaning. This perspective licenses to equate *topological proximity* with *semantic relatedness* (Baroni and Lenci, 2011), i.e., words that are closely adjacent to each other in semantic space are considered semantically strongly related[2]. Thus, in the example in Figure 2.1, *car* is much more closely related to *truck* than to *tree*.

The primary focus of distributional semantic models is on acquiring densely populated vector representations which give rise to meaningful, reliable relatedness estimates for various target words. For most applications making use of vector space models (e.g., word clustering, query expansion or document classification, to name just a few prominent ones[3]), it is of primary importance that the dimensions of meaning as spanned by the context words are sufficiently discriminative in order to facilitate meaningful clusters of target words within the semantic space. In contrast, the linguistic properties of the context words are often disregarded in models addressing these tasks, which means that assessing the *degree* of similarity (or relatedness, respectively) between target words is given preference over linguistic insights into its *source* (cf. Hartung and Frank, 2011a) or *type* (Padó and Lapata, 2003). In the next section, we outline a spectrum of different variants of distributional semantic models, reflecting a range of linguistic layouts and purposes of application.

2.3 Variants of Distributional Semantic Models

2.3.1 Conceptual and Notational Foundations

Our notation follows Thater et al. (2010). Assuming sets of *target words* W and *context words* C, we define the most general variant of a Euclidean vector space V (Jänich, 1994) being spanned by the set of orthonormal basis vectors $\{\vec{e}_c | c \in C\}$. Given that any vector in V can be represented as a linear combination of its basis vectors, a *target vector* representing the meaning of a word $w \in W$ in V is defined as follows:

$$\vec{w} = \sum_{c \in C} \omega(w, c) \cdot \vec{e}_c \tag{2.1}$$

In this definition, $\omega : W \times C \rightarrow \mathbb{R}$ is a function that assigns a weight to each component of a vector, i.e. to each pair consisting of a target $w \in W$ and a context word $c \in C$. Throughout this thesis, we will refer to ω as the *component weighting function*. Depending on the type of vector space model and the envisaged tasks, its result can be raw

[2] In the distributional semantics literature, the terms *relatedness* and *similarity* are sometimes used interchangeably, which we consider a lack of terminological rigidity. We adhere to the notion of *semantic relatedness* as introduced by Budanitsky and Hirst (2006) who regard semantic similarity as a special case of semantic relatedness (which subsumes a variety of lexical semantic relations such as hyponymy, meronymy, antonymy etc.). Citing an example from Resnik (1995), they consider *car* and *gasoline* more closely related than *car* and *bicycle*, whereas the latter pair is more similar.

[3] See Turney and Pantel (2010) for a more exhaustive survey.

frequencies as extracted from corpora or their transformations to association scores or probabilities, for instance.

As an example, consider $\vec{w}$, the vector representation of a target word w in a vector space with dimensions $C = \{drive, park, engine, flower\}$ corresponding to base vectors $\vec{e}_{drive} = (1,0,0,0)$, $\vec{e}_{park} = (0,1,0,0)$, $\vec{e}_{engine} = (0,0,1,0)$ and $\vec{e}_{flower} = (0,0,0,1)$. Assuming we extracted 15 occurences of w in the context of *drive*, 7 occurrences of w with *park* and 11 occurrences of w with *engine* from the underlying corpus, a purely frequency-based instantiation of the context weighting function yields:

$$\omega(w, drive) = 15 \tag{2.2}$$
$$\omega(w, park) = 7 \tag{2.3}$$
$$\omega(w, engine) = 11 \tag{2.4}$$
$$\omega(w, flower) = 0 \tag{2.5}$$

By substitution into Equation (2.1), we have:

$$\vec{w} = 15 \cdot (1,0,0,0) + 7 \cdot (0,1,0,0) + 11 \cdot (0,0,1,0) = (15,7,11,0) \tag{2.6}$$

Note that setting $w = car$ in this example leads to the target vector $\vec{car}$ as contained in Figure 2.1; target vectors $\vec{truck}$ and $\vec{tree}$ can be constructed analogously.

Vector representations within one vector space can be compared with regard to their *spatial proximity* which is considered as a distributional correlate of *semantic similarity*[4] (Widdows, 2004). A common metric for assessing the degree of similarity between two vectors $\vec{w}_1$ and $\vec{w}_2$ is given in (2.7):

$$sim(\vec{w}_1, \vec{w}_2) = \frac{\vec{w}_1 \cdot \vec{w}_2}{||\vec{w}_1||\ ||\vec{w}_2||} \tag{2.7}$$

This metric is based on the inner product of $\vec{w}_1$ and $\vec{w}_2$ in the numerator. Normalizing both vectors by their magnitude in the denominator reduces them to unit length. Thus, $sim(\vec{w}_1, \vec{w}_2)$ is equivalent to the cosine of the angle enclosed by the two vectors (Widdows, 2004). The metric is commonly referred to as as *cosine similarity* and widely used within distributional semantics, mostly for its inherent normalization capacities which abstracts from frequency effects and measures the semantic concordance of two vectors only from their spatial orientation instead (Widdows, 2004). For instance, *car* and *truck* are much more similar than *truck* and *tree* according to the example given in Fig. 2.1 ($sim(\vec{car}, \vec{truck}) = 0.96$ vs. $sim(\vec{truck}, \vec{tree}) = 0.09$), because their vector representations have more dimensions of meaning in common, not because their overall frequency profiles are similar. An overview of alternative similarity metrics and their properties is given by Weeds et al. (2004).

[4] Being rooted in linear algebra rather than in linguistics, the interpretation of spatial proximity in vector space models we discuss here does not account for the linguistically motivated differences between semantic similarity and semantic relatedness as introduced above.

In the following, we introduce several variants for constructing distributional semantic models, leading to models of different layouts which will be of importance throughout this thesis.

2.3.2 Structured vs. Unstructured Models

We first discuss two different modes of *context selection* (Padó and Lapata, 2007) which result in the distinction between *structured* and *unstructured* distributional models. The main difference between both types of models is that they impose different constraints on the co-occurrence relation between target and context words.

Unstructured models. Our terminology largely follows Baroni and Lenci (2010) who state that unstructured distributional models "do not use the linguistic structure of texts to compute co-occurrences, and only record whether the target occurs in or close to the context element, without considering the type of this relation". The most important parameter of such a model concerns the size of the contextual region[5] surrounding a target word (Sahlgren, 2006), often also denoted as the *context window*. Using five tokens on either side as context region has become a popular setting (Church and Hanks, 1990), even though considerably larger windows are also found useful for specific purposes (Niwa and Nitta, 1994; Schütze and Pedersen, 1997; Schütze, 1998). At the extreme end of the spectrum, entire documents are considered as context regions. Approaches of this kind are usually adopted in information retrieval (Salton et al., 1975). Due to their unordered nature and lack of internal structure, the context words selected by unstructured distributional models are often denoted as *bags of words*.

Structured models. In contrast, structured models are based on the assumption that specific linguistic contexts contribute to different aspects in the representation of word meaning (Padó and Lapata, 2007; Lin, 1998). In the linguistic literature, syntactic patterns of co-occurrence are found to be particularly important cues to lexical meaning (Levin, 1993). Therefore, structured distributional models explicitly rely on linguistic structure in order to select the context words used as features for representing target words.[6] Our notion of structured distributional models throughout this thesis subsumes both *dependency-based* and *pattern-based* strategies[7] for context selection: In the former case, co-occurrence relations between target and context words are established

[5] Additionally, the context words extracted from a particular context region are often filtered by eliminating stop words (Turney and Pantel, 2010) and/or function words (Dagan et al., 1993), and lemmatized (Karlgren and Sahlgren, 2001).

[6] In our notion of structured distributional models, we focus on the aspect of using *linguistic structure* for context selection, whereas aspects of structured representation and their relationship to the expressive power of distributional models (Baroni and Lenci, 2010) are abstracted from.

[7] For slightly different approaches also categorized as structured distributional models, see Erk (2012) and Erk and Padó (2008, 2009).

	my	*new*	*sports*	*drives*	*very*	*fast*	*the*	*has*	*a*	*red*	*color*
car	1	1	1	1	2	2	1	1	1	1	0

(a) Unstructured model

	my	*new*	*sports*	*drives*	*very*	*fast*	*the*	*has*	*a*	*red*	*color*
car	0	0	0	1	0	0	0	1	0	0	0

(b) Dependency-based model

	my	*new*	*sports*	*drives*	*very*	*fast*	*the*	*has*	*a*	*red*	*color*
car	0	0	0	0	0	0	0	0	0	0	1

(c) Pattern-based model

Figure 2.2: Comparison of structured and unstructured distributional semantic models constructed from toy corpus in Example (15).

along syntactic dependency paths, in the latter case by matching lexico-syntactic patterns.

(15) *My new sports car drives very fast. The car has a red color.*

Specificity and sparsity of distributional models. Considering the toy corpus in (15), we compare three different settings of constructing a distributional semantic model for representing the lexical meaning of the target word *car*:

(a) an **unstructured model**, using a context window of three tokens on each side of a target word (ignoring punctuation)

(b) a structured **dependency-based model**, extracting all head verbs linked to a target word by a `sbj` relation

(c) a structured **pattern-based model**, extracting all context words matching the lexico-syntactic pattern: `the TARGET has a JJ CONTEXT`[8]

The distributional representations resulting from these settings are depicted in Figures 2.2a–2.2c. Despite being highly fragmentary, these vectors reveal interesting differences pointing to important analytical concepts for comparing distributional semantic models of different types.

The first difference concerns the aspect of *sparsity*: Vector representations generated from an unstructured model (cf. Figure 2.2a) tend to be rather *dense*, i.e., to contain

[8]Here, `TARGET` stands for the target word, `CONTEXT` for a context word, and `JJ` denotes arbitrary adjectives as defined by the Penn Treebank tagset (Marcus et al., 1993).

a large number of vector components populated with positive values, given that the bag-of-words assumption yields a large number of target-context pairs. In contrast, the pattern-based model (cf. Figure 2.2c) suffers from high sparsity, with only one vector component above zero due to the fact that the lexico-syntactic pattern used for context selection can be expected to yield only a small number of corpus matches. The dependency-based model (cf. Figure 2.2b) is in an intermediate position.

The second difference concerns the aspect of *specificity*. As discussed above, the degree of semantic relatedness between target words as predicted by a vector space model spanning multiple dimensions of meaning does not necessarily allow a conclusion with regard to the type of semantic relation holding between these targets (Baroni and Lenci, 2011).[9] For instance, adopting an example by the same authors, it is fairly plausible to imagine a distributional model rating *dog* and *animal* as highly related terms, as well as *dog* and *tail*. Such a model clearly lacks the ability to discriminate different types of semantic relations. As noted by Padó and Lapata (2003), this is one of the major criticisms put forward against distributional models in general, as it may limit the practical usability of such models in many NLP tasks that require this capability.

We argue, in turn, that criticisms of this kind are too broad as they ignore recent trends in the field to (i) develop multi-purpose distributional models that can be adapted to specific tasks (Baroni and Lenci, 2010) or (ii) tailor distributional models to particular semantic relations in the first place. The latter can be achieved by intelligent combinations of structured methods of context selection (Padó and Lapata, 2007; Baroni et al., 2010) and similarity metrics (Weeds et al., 2004; Michelbacher et al., 2011).

Throughout this thesis, we refer to the ability of a distributional model to tailor its context selection method to one particular semantic relation as its *specificity*. The example in Figure 2.2c, for instance, illustrates a highly specific pattern-based model[10] that selects only attribute terms as contexts for representing noun meaning. In comparison, the dependency-based model in Figure 2.2b selects a group of functionally related contextual terms resembling qualia roles (Pustejovsky, 1995) which are difficult to nail down to one particular semantic relation, though. Therefore, dependency-based distributional models can be seen as representatives of medium specificity. Unstructured models, as exemplified in Figure 2.2a, usually induce a wide variety of relations between targets and context words ranging from properties (*car–new*, *car–fast*, *car–red*) to loose topical associations (*car–sports*), due to the lack of meaningful linguistic constraints imposed on their extraction. Thus, unstructured models are sufficiently rich in order to cover the entire spectrum of semantic relatedness, at the same time being the most unspecific ones among the three types of models compared here.

[9] According to Sahlgren (2008), this problem can be traced back to the distributional hypothesis being "a strong methodological claim with a weak semantic foundation. It states that *differences* of meaning correlate with *differences* of distribution, but it neither specifies *what kind* of distributional information we should look for, nor *what kind* of meaning differences it mediates."

[10] Apart from the showcase introduction provided here for ease of presentation, please refer to Section 3.3 for a more thorough discussion of previous work on structured distributional models.

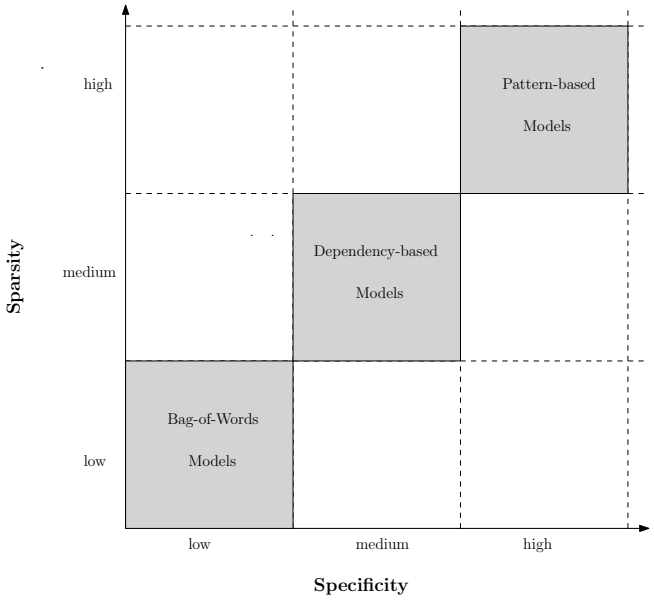


Figure 2.3: Different types of distributional models in a spectrum of specificity and sparsity

Setting specificity and sparsity in relationship to each other, we observe a reciprocal pattern that gives rise to the following hypothesis: The more specific a model, the more liable it is to effects of sparsity, and vice versa. This relationship is graphically depicted in Figure 2.3. It follows directly from this hypothesis that specificity and density are conflicting goals in distributional models of lexical meaning. This leads to the important research question as to whether distributional semantics is inherently limited to capturing unspecific semantic relatedness, thus opposing its suitability for NLP applications that require more specific semantic knowledge, or how a practical compromise between specificity and sparsity can be effectively achieved. In this thesis, this question is investigated with regard to the attribute relation between adjectives and nouns.

2.3.3 Syntagmatic vs. Paradigmatic Models

Syntagmatic and paradigmatic distributional models owe their names to the functional relation – *syntagmatic* or *paradigmatic* – that holds between the target elements being represented within the respective model and the context words used to describe them.

Roots in structuralism. According to the structuralist school in linguistics (dating back to de Saussure (1916)), syntagmatic relations hold between words[11] that co-occur in sequential configurations, whereas paradigmatic relations hold between words that

[11] In fact, structuralist theory can be (and has been) analogously applied to more basic units of language such as morphemes and phonemes as well, but we are merely concerned with words here.

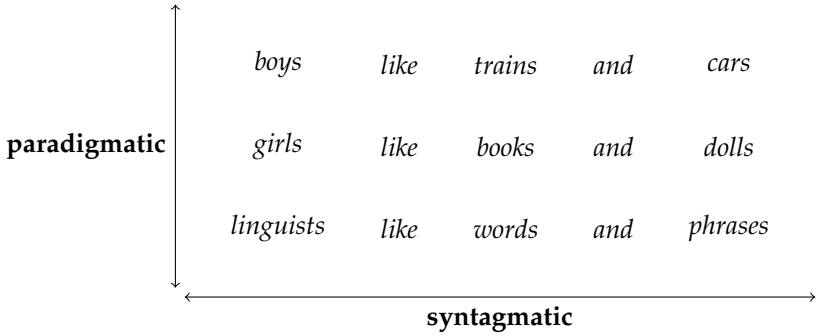


Figure 2.4: Syntagmatic and paradigmatic relations

	trains	*cars*	*books*	*dolls*	*words*	...
boys	1	1	0	0	0	...
girls	0	0	1	1	0	...
linguists	0	0	0	0	1	...

(a) Syntagmatic model

	boys	*girls*	*linguists*	*trains*	*cars*	...
boys	0	1	1	0	0	...
girls	1	0	1	0	0	...
linguists	1	1	0	0	0	...

(b) Paradigmatic model

Figure 2.5: Fragments of distributional models based on the toy corpus from Fig. 2.4

do not co-occur syntagmatically, but are substitutable by one another as they occur in the same (or similar) contexts at different times. Thus, syntagmatic relations can be seen as functional relations *in praesentia*, paradigmatic ones as relations *in absentia* (Sahlgren, 2006). For illustration, consider Figure 2.4, where three toy sentences were arranged in a tabular structure in order to highlight their functional equivalences and differences: All words in the same line stand in a syntagmatic relation, while all words in the same column share a paradigmatic relation. These examples clearly demonstrate that the semantic relations derived from syntagmatic and paradigmatic structural patterns are both meaningful, but obviously different in type (e.g., *linguist–book*, *linguist–word* vs. *linguist–boy* or *linguist–girl*).

Constructing syntagmatic distributional models. Based on these definitions, syntagmatic distributional models can straighforwardly be constructed from text corpora by (i) selecting sets of target elements and context words and (ii) extracting all instances of syntagmatic relations between the members of these sets such that only those vector components linking syntagmatically related words obtain a positive value. Applying this procedure to the toy corpus from Fig. 2.4 yields the syntagmatic distributional model displayed in Fig. 2.5a.

Constructing paradigmatic distributional models. Paradigmatic distributional models aim at discovering members of paradigmatic semantic relations, e.g., fillers of a predicate's argument position. These are often "semantically coherent" in that they can be subsumed under a particular ontological category and/or share common properties (Erk et al., 2010; Schulte im Walde, 2010). An idealized example of a paradigmatic model is shown in Fig. 2.5b. According to this model, based on the toy data from Fig. 2.4, *boys* are paradigmatically related to *girls* and *linguists*, while being unrelated to *trains* and *cars*, for instance. Given that paradigmatic relations are functional relations *in absentia*, the most common approach to inducing paradigmatic distributional models automatically from corpora is based on (i) constructing vector representations for all target words of interest from appropriate syntagmatic contexts, and (ii) applying a similarity metric (e.g., the cosine measure as introduced in Section 2.3.1 above) in order to identify highly similar vectors. Thus, pairs of target words with a high proportion of shared syntagmatic contexts are considered paradigmatically related[12]. Consequently, the selection of syntagmatic contexts has an immediate impact on the prospects of acquiring meaningful paradigmatic relations. Following up on our previous discussion on structured vs. unstructured distributional models, paradigmatic models require a thoroughly designed compromise between fostering specificity and avoiding sparsity in order to be effective.

2.3.4 First-order vs. Second-order Models

The distinction between first-order and second-order distributional models is primarily due to differences in their underlying notions of context representation: While first-order models represent the meaning of a target word along directly co-occurring, syntagmatically related context words, second-order models capitalize on *context words of context words* (Schütze, 1998; Purandare and Pedersen, 2004).

Most intuitively, these types of distributional models can be understood in graphical terms as being constructed from contextual paths between target elements and context words, where the path length k determines the order of the model ($k = 1$ yielding a first-order model, $k = 2$ a second-order model). For illustration, compare Figs. 2.6 and 2.7, showing first- and second-order contextual paths for the target word *tractor*. In both figures, the target word is surrounded by a rectangle, while all words entering the resulting model as a context word are marked by an ellipse. As can be seen, the first-order representation of *tractor* is solely based on paths with $k = 1$, i.e., each context word is reached by following one arc (cf. Fig. 2.6). Second-order representations are obtained recursively from computing contextual paths on top of all first-order context words, thus extending the first-order paths by one additional arc (cf. Fig. 2.7). Note that only *end points* of second-order contextual paths are used as context words in the

[12] In practice, either clustering techniques (Prescher et al., 2000; Purandare and Pedersen, 2004) or similarity thresholds (McNamara et al., 2007; Corley and Mihalcea, 2005) are applied in order to discriminate members of a valid paradigmatic relation from noise.

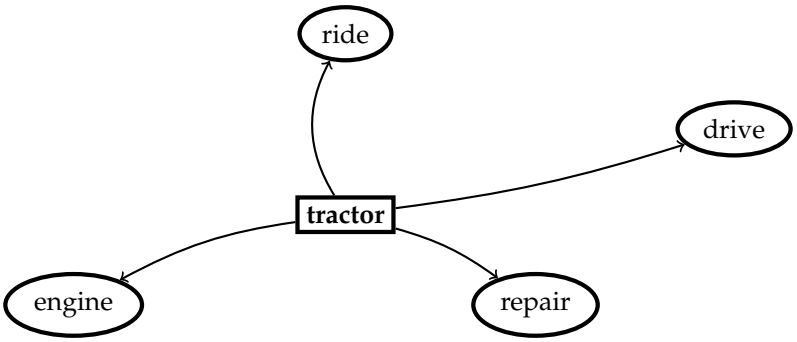


Figure 2.6: First-order contextual paths for *tractor*

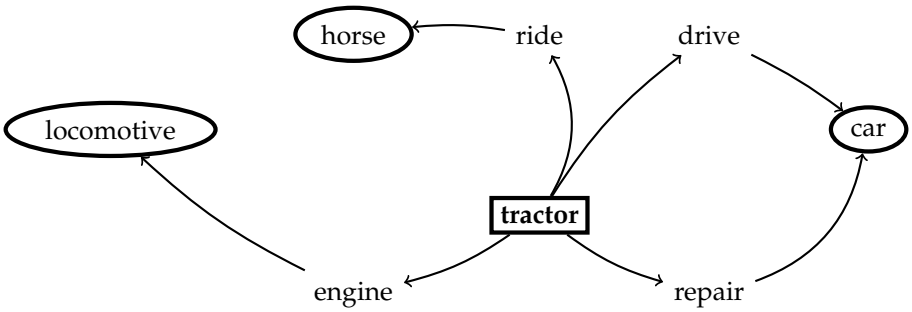


Figure 2.7: Second-order contextual paths for *tractor*

resulting second-order vector representation of the target word.

Following our definition from Chapter 2, originally based on Thater et al. (2010), we give formal definitions of first-order and second-order vectors as follows. A first-order vector $\vec{w}$ representing a target word w is defined as:

$$\vec{w} = \sum_{w' \in C} \omega(w, w') \cdot \vec{e}_{w'} \tag{2.8}$$

As can be seen from the definition in (2.8), the backbone for populating the components of the first-order vector $\vec{w}$ is the contextual relation between w and each member w' of a set of context words C.

A second-order vector representing w is defined as:

$$\vec{w} = \sum_{w'' \in C} \left(\sum_{w' \in C} \omega(w, w')\ \omega(w', w'') \right) \cdot \vec{e}_{w''} \tag{2.9}$$

This definition clearly emphasizes that second-order representations are constructed from first-order contextual paths by recursion on w'. Thus, w' merely serves as a hinge between w and w'', without being explicitly contained in the resulting second-order vector: The components of $\vec{w}$ are determined by the relation (w, w'') that is computed by summing up all paths from w to w'', generalizing over w' (cf. Thater et al., 2010).

Essentially, the two strategies of vector construction just discussed are reflected by different inventories of features that are available in order to describe the meaning of a target word in a vector representation: A target is either described in terms of first-order context words (e.g., a *tractor drives*, has an *engine*, etc.) or second-order context words sharing the same first-order contexts (e.g., *tractors* are closely related to *cars* along the shared context *drive*, and also to *locomotives* along the shared context *emit fumes*). Note that the second-order approach increases the chance of observing a particular target-context pair (w, w'') by considering multiple paths along different intermediate contexts w', whereas first-order models are dependent on observing (w, w'') directly. This aspect is often referred to as an advanced *density* of second-order vector representations compared to first-order ones (cf. Schütze, 1998). Therefore, second-order distributional models can be expected to be particularly effective in paradigmatic settings aiming at the acquisition of pairs of target words that are members of the same semantic category.

2.4 Meaning Representation beyond the Word Level

For a long time, distributional semantic models have been merely applied to represent the meaning of individual words. Only recently, the linguistic principle of *compositionality*, stating that the meaning of a complex linguistic expression is a function of its constituents and their syntactic combination (Frege, 1892), has found its way into distributional modelling. Being deeply rooted in formal semantics since Montague (1974), this principle is at the core of all computational approaches to *semantics construction*,

i.e., the attempt of assigning formal meaning representations to natural language expressions.

Many of these approaches rely on predicate logic as target formalism, using the *lambda calculus* for constructing logical representations of phrase and sentence meaning (Blackburn and Bos, 2005; Bos, 2008). The formal system of the lambda calculus is based on logical predicates as elementary representations of word meaning, from which complex formulae are composed by means of the lambda operator used for binding arguments to predicates or combining predicates.

In order to lift distributional semantic models beyond the word level, a similar idea is applied, following Mitchell and Lapata (2010): Individual word vectors (as in Fig. 2.2, for instance) are taken as elementary representations of lexical meaning; as a composition operator, the two vector operations $\oplus$ (vector addition) and $\odot$ (vector multiplication) are used. These operations are defined as component-wise addition and multiplication, respectively, as given in Equations (2.10) and (2.11):

$$\begin{aligned} \vec{w_1} \odot \vec{w_2} &= \sum_{a \in A} \omega(w_1, a) \cdot \vec{e_a} \odot \sum_{a \in A} \omega(w_2, a) \cdot \vec{e_a} \\ &= \sum_{a \in A} (\omega(w_1, a) \cdot \omega(w_2, a)) \cdot \vec{e_a} \end{aligned} \tag{2.10}$$

$$\begin{aligned} \vec{w_1} \oplus \vec{w_2} &= \sum_{a \in A} \omega(w_1, a) \cdot \vec{e_a} \oplus \sum_{a \in A} \omega(w_2, a) \cdot \vec{e_a} \\ &= \sum_{a \in A} (\omega(w_1, a) + \omega(w_2, a)) \cdot \vec{e_a} \end{aligned} \tag{2.11}$$

This particular approach to compositionality in distributional semantic models has two important implications: First, elementary and composed vectors live in the same semantic space, i.e., one and the same model is used to represent lexical and phrasal (or even sentential) meaning. While there are good reasons for assuming that this causes problems on the level of sentence meaning (Erk and Padó, 2008; Baroni et al., 2014), we consider it a practical assumption for the attribute selection task that is in focus of this thesis. Second, the lambda operator and the vector operations introduced cannot be seen as full equivalents, because vector representations – contrary to logical predicates – usually do not encode syntactic properties of a target[13]. As a consequence, complex vectors constructed by means of vector addition or vector multiplication do not reflect the internal syntactic structure of their constituents[14]. In an attempt to overcome this obvious violation of an inherent aspect of the principle of compositionality, Baroni et al. (2014) propose to derive compositional distributional representations of phrase mean-

[13] See Erk and Padó (2008, 2009), Grefenstette and Sadrzadeh (2011) or Grefenstette et al. (2014) for notable exceptions.

[14] Being commutative operations, vector addition and vector multiplication yield identical vector representations for the phrases *dog bites man* and *man bites dog*, for instance.

ing based on linear mappings[15] closely replicating the notion of functional application from formal semantics (cf. Heim and Kratzer, 1998).

Again, we argue that the shortcomings of vector-based composition just discussed do not pose a major issue for the particular application of distributional semantic models envisaged in this thesis. A more thorough discussion of these aspects is deferred until Chapter 6.

[15] Please refer to Section 3.5 for a more detailed discussion of compositional distributional models based on linear mappings.

3 Related Work

Previous work that is related to the topics investigated in this thesis comprises *adjective classification* and various aspects of *distributional modeling of attribute meaning* in adjectives and nouns. Classification of adjective types is an important prerequisite for the corpus-based acquisition of attribute meaning from adjectives, as adjectives can be grouped into different semantic classes according to their ontological foundations. A major distinction concerns the difference between property-denoting and relational adjective types, among which only the former exhibit semantic characteristics that can be leveraged for attribute learning.

3.1 Adjective Classification

Adjective classification schemes similar to the one we apply in this thesis have been presented by Boleda (2006) for Catalan and Raskin and Nirenburg (1998) for English. Their goal was the creation of a large-scale adjective lexicon for NLP tasks. The most fundamental difference between the work of Raskin and Nirenburg and ours is that they created their resource manually. In contrast, we aim at automatic classification, as effective automatic methods have the advantage that they can be applied to novel, specialized domains and possibly to other languages. Boleda (2006) made use of clustering techniques to automatically establish adjective classes in Catalan. She obtained various sets of clusters that were evaluated against a human-annotated gold standard, yielding up to 73% accuracy. A strict comparison of the two approaches will not be possible due to the different languages considered and divergences regarding the selected target classes. In fact, her approach included several language-specific features so that it is unclear whether it can be transferred to English. Since our aim is the targeted acquisition and classification of adjectives for the purpose of ontology learning, we opt for a classification approach that allows us to pre-specify (and possibly refine and extend) appropriate target classes for concept learning – which is not possible within a clustering approach.

Amoia and Gardent (2008) present a (manual) classification of adjectives that relies on logical properties of adjectives in the tradition of Montague (1974). While this perspective is orthogonal to our work, their work might be useful to supplement our approach by providing further adjective classes that may be sorted out as being neither property-denoting nor relational.

Methodologically, our approach to adjective classification is related to a great body of

work in automatic verb classification (e.g., Miyao and Tsujii (2009)), going back to the empirical work of Levin (1993). Although in this field the number of target classes is by far greater and aimed at a conceptual semantic classification, the common denominator between verb semantic classes and the adjective classes considered here is that certain distributional properties on the type level are constitutive for class membership, while the full range of these properties is not observable on the token level. In line with this strand of work on Levin-style verb classification, our classification approach will operate on the type level.

3.2 Attribute Learning

Manual efforts. The importance of attribute information for many NLP applications is underlined by manual efforts to construct ontological resources such as the *Extended Named Entity Ontology* (Sekine and Nobata, 2004; Sekine, 2008). This ontology contains about 200 classes relevant for named entity recognition. It is built around attribute information as the main source to distinguish these classes and arrange them in a hierarchy. This information has been manually compiled from dictionaries and encyclopedias.

Learning from structured and semi-structured sources. Due to the tediousness and notorious incompleteness of these manual attempts, other approaches have focused on learning attribute information from semi-structured textual sources such as tables in web pages (Tokunaga et al., 2005; Yoshinaga and Torisawa, 2007), Wikipedia infoboxes (Wu et al., 2008; Bing et al., 2013), query logs (Alfonseca et al., 2010; Pasca, 2011) or even existing knowledge bases or lexical resources (Lee et al., 2013; Bakhshandeh and Allen, 2015). These approaches underline the relevance of attribute knowledge for a variety of practical applications; our work focusses on attribute knowledge acquisition from unstructured textual sources, though.

Learning from unstructured text. Using adjectives for attribute learning has first been proposed by Almuhareb and Poesio (2004) and Cimiano (2006). Cimiano's work on this particular task is based on the investigation of adjective-noun phrases from corpora. For every adjective modifying a noun, its possible attributes are extracted from WordNet (Fellbaum, 1998) and associated with the respective noun. As this approach depends on an external lexical resource, it is obviously limited in coverage.

Almuhareb (2006) aims at learning this information by means of a pattern-based approach that operates on large web-based corpora. The outcome of his work on this task, however, is considerably affected by the lack of a separation between property-denoting and relational adjectives, such that a large number of adjectives is erroneously identified by his system as denoting a property. In Chapter 5 of this thesis, we present a classification approach for distinguishing these classes automatically. In Chapter 7,

we provide a comparative evaluation of our distributional methods for attribute acquisition against Almuhareb's pattern-based approach.

In the *Generative Lexicon* (Pustejovsky, 1995), the internal lexical structure of nouns is modeled as consisting of *qualia roles* which are grouped into FORMAL, CONSTITUTIVE, AGENTIVE and TELIC qualia. Each qualia role subsumes particular types of properties: The FORMAL role contains properties that are used to distinguish the referent of the noun from related concepts; the other roles contain properties involved in bringing the referent about, its use, or its constituitve parts, respectively. Hence, qualia roles can be seen as further abstractions on top of attributes. Cimiano and Wenderoth (2007) aim at automatically inducing ranked qualia structures from the web, using manually designed patterns specifically created for each role in order to launch queries against a web search engine. Their obtained qualia roles are rated as largely plausible by human judges in an a posteriori evaluation; however, their approach achieves rather low performance in both precision and recall in an experimental evaluation against a manually created gold standard, which underlines the sparsity issues inherent to pattern-based acquisition methods. Following a similar approach to their acquisition, Katrenko and Adriaans (2008) demonstrate that qualia structures may be of use for automatic categorization of concrete nouns.

Tandon et al. (2014) propose a semi-supervised method for populating a large knowledge base with triples of nouns, attributes and adjectives, as given in (16):

(16) a. ⟨botanical-plant, hasColor, green-color⟩
b. ⟨power-plant, hasQuality, green-environmental⟩

Domain and range of these triples are acquired from predicative and attributive adjective-noun phrases as occurring in the Google 5-grams corpus (Brants and Franz, 2006). Their system also includes word sense disambiguation, i.e., ambiguous adjectives and nouns as in *green plant* are automatically mapped to their word senses according to WordNet, as illustrated in (16a) and (16b). Apart from this disambiguation aspect, their goal is similar to ours in that attribute knowledge that remains implicit in adjective-noun phrases is made explicit. Importantly, however, they follow a semi-supervised label propagation approach in order to determine the most appropriate attribute for a given phrase: In a graph connecting adjectives, nouns and word senses, a set of monosemous adjectives is labeled with the correct attribute as provided by WordNet. Afterwards, these labels are propagated across the entire graph. Thus, their approach crucially depends on the existence of a lexical resource providing initial mappings between adjectives and attributes.

Likewise in a semi-supervised setting, Probst et al. (2007) induce attributes from product descriptions. Their work is based on highly product-specific attributes (e.g., *hitting surface* or *construction* for *baseball bats*). Using attribute profiles in order to compare products from different vendors for recommendation purposes, they demonstrate the relevance of attributes as a powerful source of knowledge for practical applications.

3.3 Structured Models in Distributional Semantics

For this discussion, we distinguish two types of structured distributional models: *dependency-based* models and *pattern-based* models.

Dependency-based models. Prominent examples for the former type of models are Padó and Lapata (2007), Lin (1998) and Rothenhäusler and Schütze (2009), among others, who use syntactic dependencies as contexts for constructing the meaning dimensions of semantic spaces. The authors show that the higher degree of specificity of their models compared to simple bag-of-words representations (cf. Section 2.3.2) may be beneficial for various tasks, among them synonymy detection, word sense ranking and concept categorization.

Extending the distributional hypothesis beyond individual words, Lin and Pantel (2001) assume that dependency paths which frequently link the same set of words tend to express similar meanings. Based on this assumption, their *DIRT* system discovers semantic equivalences between dependency paths such as *X wrote Y* and *X is the author of Y* which can be seen as paraphrases of each other and are highly useful for NLP tasks such as question answering.

As an important difference between the models discussed so far, we note that Padó and Lapata (2007) restrict themselves to rather short dependency paths (paths of length 1 plus a selection of additional paths covering the internal structure of noun phrases), whereas Lin and Pantel (2001) and Rothenhäusler and Schütze (2009) also consider more complex syntactic configurations beyond the clause level. In our work, we also follow the latter strategy in order to design dependency paths meaningful for attribute selection.

Baroni and Lenci (2010) criticize distributional models as being mostly tailored to one particular task. As a consequence, they propose the *Distributional Memory* which has to be regarded as the most general distributional model currently available. Overcoming the assumption underlying all prior work in distributional semantics that distributional patterns emerge from binary relations between target and context words, their framework represents all corpus data in terms of triples of target words, context words, and the link between them being made explicit. These triples are extracted from dependency parses and stored in a three-dimensional tensor. This resource provides a unique source of distributional information which can be exploited for numerous tasks by slicing the tensor into individual matrices providing a particular view on the data that is most suitable for the respective task. Despite its generality, however, the Distributional Memory does not contain ready-to-use information for the attribute selection task.

Erk and Padó (2008, 2009) extend dependency-based approaches to models capable of capturing predicate-argument structure: In their model, transitive verbs, for instance, are represented as triples of vectors comprising one *lexical vector* and two vectors encoding their selectional preferences in each of their syntactic argument slot. The model generates contextualized meaning representations by combining lexical vectors

of predicates with preference vectors of arguments reflecting the particular syntactic relation that holds between predicate and argument. For vector composition, they also use component-wise multiplication and addition. Hence, this work is also representative for a second trend, i.e., the representation of meaning aspects going beyond the word level in distributional models. As pointed out by the authors themselves, however, their system should not be considered as producing compositional meaning representations: "Rather than yielding a single, joint vector for the whole expression, [this] procedure for computing meaning in context results in one context-adapted meaning representation per word, similar to the output of a WSD system."

Thater et al. (2010, 2011) address the same problem in a more compositional manner, i.e., by computing one phrase vector for a verb to be contextualized and one of its arguments. In their approach, second-order dependency vectors are used to represent predicate meaning, while arguments are still represented by first-order vectors. Thus, the authors overcome a problem faced by many vector-based approaches to compositional semantics, i.e., that verbs and their arguments have different syntactic neighbors, which is why their first-order vectors are not easily interoperable in order to yield a compositional representation of a verb-argument-pair (cf. Thater et al., 2010).

Pattern-based models. The *Strudel* system (Baroni et al., 2010) is the most prominent example of a vector space model constructed from lexico-syntactic patterns for specifying the relation between target and context words. Strudel highlights two strengths of VSMs that incorporate interpretable dimensions of meaning: cognitive plausibility and effectiveness in concept categorization tasks. Concepts induced by Strudel are characterized in terms of salient properties and relations (e.g., *children* have *parents*, *wolves* live in *forests*, *grass* is *green*). However, their approach is restricted to nouns. Open questions are (i) whether it can be extended to different word classes (adjectives, in particular) and (ii) whether the interpreted meaning layers are interoperable across word classes, in order to cope with compositionality.

In Hartung and Frank (2011a), we extend pattern-based distributional modeling to an interpretable, compositional vector space model that is applied to adjective-noun composition with attributes providing shared dimensions of meaning. Moreover, to our knowledge, this is the first attempt to expose such a model to a pairwise similarity judgement task on the level of adjective-noun phrases.

Pattern-based distributional models can also be tailored to simulating complex cognitive tasks such as analogy identification, as demonstrated in Turney (2008). At the core of his *Latent Relation Mapping Engine* is a pair-pattern matrix representing pairs of words (e.g., *stone:mason* or *wood:carpenter*) along dimensions of meaning constructed from lexico-syntactic patterns. Following the distributional hypothesis, two pairs are considered as forming an analogy if they share a substantial proportion of patterns linked to them.

Pattern-based approaches in Relation Extraction and Ontology Learning. The work presented in this thesis is also influenced by other pattern-based approaches to modelling lexical meaning whose underlying methodology clearly reflects distributional principles, without being geared towards meaning representation in semantic spaces. This comprises work in relation extraction and ontology learning, as summarized by Frank and Padó (2012): In her seminal work on mining hypernym-hyponym pairs from corpora by means of *is-a* patterns, Hearst (1992) laid the cornerstone for automated pattern-based extraction of fine-grained semantic or ontological relations. In subsequent years, her methodology has been followed, among others, by Girju et al. (2006) covering part-whole relations, Pantel and Pennacchiotti (2008) covering causality and Girju et al. (2009) classifying a variety of fine-grained semantic and ontological relations.

Integration of word spaces and lexico-syntactic patterns. Lexico-syntactic patterns are sometimes used for classifying or labeling the output of unspecific distributional semantic models (i.e., clusters of semantically related words), typically built from bag-of-words contexts (Lin et al., 2003; Pantel and Ravichandran, 2004). Apart from this, the only research we are aware of that integrates pattern-based extractions with distributional semantic models is Mirkin et al. (2006) who aim at inducing word pairs exhibiting lexical entailment. Based on the observation that lexico-syntactic patterns and distributional similarities (as provided by a dependency-based model in this case) offer some complementarity (cf. Section 2.3.2), the authors propose to generate entailment candidates from both sources first, then construct a feature set for these candidates based on both extracted patterns and distributional similarity scores and finally select only those candidates considered valid from both perspectives. The latter step, however, implies supervised classification, which clearly goes beyond distributional semantics as an inherently unsupervised methodology.

3.4 Topic Models in Distributional Semantics

Recently, Latent Dirichlet Allocation (Blei et al., 2003; Steyvers and Griffiths, 2007) has found its way into lexical semantics. Originally designed for tasks such as text classification and document modeling, LDA provides an unsupervised generative probabilistic approach to the decomposition of document collections into latent *topics*, i.e., probability distributions over words. These topics can be used (i) as low-dimensional representations of the contents of individual documents, and (ii) in order to characterize the meaning of individual words within the collection.[1]

Topic modelling on pseudo-documents. The latter aspect, in particular, qualifies LDA for being used in lexical semantics. Ritter et al. (2010) and Ó Séaghdha (2010), for in-

[1] More technical details of LDA will be given in Section 6.3.1.

stance, model selectional restrictions of verb arguments by inducing topic distributions that characterize mixtures of topics observed in verb argument positions. As a basis for LDA modelling, they collect *pseudo-documents*, i.e., bags of words that co-occur in syntactic argument positions. Thus, they are able to generalize over sets of individual words observed in these syntactic positions, towards distributional approximations of semantic roles.

Topics as dimensions. Mitchell and Lapata (2009, 2010) were the first to integrate LDA and distributional semantic models by using topics induced from bag-of-words context representations as dimensions in semantic spaces. The resulting model is evaluated in a similarity prediction task on pairs of adjective-noun, noun-noun and verb-object phrases. In comparison to standard vector space models over bag-of-word contexts, LDA induces topic-based meaning representations which turn out inferior for adjective-nouns and noun-nouns. Only in case of verb-object pairs, the topic space enables more accurate similarity judgements than the word space model.

Comparison. In our work, we adopt an approach similar to Mitchell and Lapata (2010) by embedding latent variable information from LDA into a distributional semantic model representing adjective-noun meaning. However, our method differs from theirs in two important aspects: First, we induce topics from attribute-specific pseudo-documents rather than bags of context words. As a result, the topics induced by our approach will be specifically tied to attribute meaning. Thus, they offer an interpretable semantics[2] and can be used as probabilistic indicators for the prediction of attributes as semantic target categories in adjective-noun composition. Second, with respect to the benefits of dimensionality reduction in semantic spaces, Mitchell and Lapata (2010) exploit only part of the generative process underlying LDA, limiting themselves to word-topic probabilities as the only source of probabilistic information being transfered from LDA into the distributional model. In our approach, the use of attribute-specific pseudo-documents licenses to consider probability distributions of topics over attributes as well. This bears the potential of inferring smooth probability estimates for the relation between target words and attributes to be injected into the distributional model in order to alleviate sparsity problems.

3.5 Distributional Models of Phrase Meaning

In this section, we review previous work on distributional semantic models for representing meaning aspects beyond the word level. As this topic has received considerable

[2]The lack of interpretability is a notorious difficulty of plain LDA models (Chang et al., 2009) and other methods for dimensionality reduction in distributional models such as Latent Semantic Analysis (Deerwester et al., 1990; Landauer et al., 1998).

attention in the literature throughout the last years (see Erk (2012) for an overview), we focus on work involving adjective-noun compositionality here.

Vector Mixture Models. Mitchell and Lapata (2010) propose a general framework for distributional *vector mixture models*, considering compositional phrase meaning in distributional models as a function of two individual word vectors. In their study, they compare a variety of vector composition operators such as vector addition, vector multiplication (cf. Section 2.4 of this thesis), tensor product, circular convolution or dilation, among others. In order to assess the linguistic plausibility of these operators and their performance, they are used to generate composed vector representations for adjective-noun, noun-noun and verb-object phrases. Comparing these phrase representations in a similarity judgment task, the authors find that vector multiplication yields the best correlation with human judgments (Spearman's $\rho = 0.46$, compared to a human upper bound of $\rho = 0.55$) for adjective-noun phrases. This result is confirmed in noun-noun phrases ($\rho = 0.49$, equal to the upper bound), which suggests that vector multiplication is a very plausible and robust choice for modeling vector composition in modification contexts.

Linear Mappings. Based on the insight from theoretical linguistics that intersective approaches to adjective-noun composition fail in cases of subsective or intensional adjectives (Kamp, 1975), Baroni and Zamparelli (2010) as well as Guevara (2010) argue that vector mixture models are not sufficiently expressive in order to provide a general means for constructing compositional distributional representations of adjective-noun phrases. Adopting the formal semantics approach taken by Montague (1970), they stipulate that adjectives should be treated as functions from the meaning of a noun onto the meaning of a modified noun, such that the original meaning of the noun is not necessarily preserved in the compositional representation (as in phrases involving intensional adjectives, e.g., *fake gun*).

Transferring this idea to linear algebra, the authors model adjectives as matrices encoding linear mappings between noun vectors. The weights populating these matrices are estimated by partial least squares regression. In this process, noun vectors are taken into account as independent variables, while observed vectors of adjective-noun phrases become the dependent variable. Vectors of the latter kind are constructed by collecting bag-of-words contexts for a particular phrase in the same way as for individual words. Hence, adjective meaning in these models can be seen as a mapping from the contexts observed for an individual noun, on the one hand, to the contexts of the same noun when being modified by the adjective in question, on the other. Considering *fake gun* as an example once again, a very small overlap would be expected between the contexts observed with *gun* and those observed with *fake gun*, which is the typical situation for intensional adjectives.

As a notable difference, Guevara (2010) learns *one* generic linear map for all adjectives

in the corpus (thus analytically falling back behind Kamp (1975) again), while the linear maps induced by Baroni and Zamparelli (2010) are specific to *individual* adjectives (thus lacking generality).

Functional Application in Distributional Models. Baroni et al. (2014) generalize these ideas to a general "program for compositional distributional semantics" that is centered around the notion of *functional application* in semantic spaces. Their work can be seen as the most comprehensive generalization of the insight that different linguistic phenomena require to be modeled in corresponding algebraic structures, using different composition operators as available for these structures (cf. Widdows, 2008; Grefenstette and Sadrzadeh, 2011; Grefenstette et al., 2014).

Relying on categorial grammar (Montague, 1970) as syntactic backbone of their model, Baroni et al. (2014) postulate a correspondence between syntactic categories and semantic types. The semantic type then determines the algebraic structure chosen to encode the compositional behaviour of the members of a syntactic category. While nouns, being of the elementary category N, are still represented as vectors, adjectives belong to the complex category N/N mapping the meaning of a noun to a modified noun. This mapping is encoded as a matrix. Transitive nouns, being of the complex category $(S\backslash NP)/NP$ (mapping a noun phrase NP to another complex category taking an NP itself in order to yield a fully specified sentence S), have to be encoded in terms of a third-order tensor.[3] Crucially, these mappings ensure that the algebraic structures used to represent word and phrase meaning in distributional models are no longer forced to be objects of the same space as in purely vector-based models. Hence, the proposal of Baroni et al. (2014) adds a considerable degree of flexibility to distributional models of compositionality and empowers them to deal with an unprecedented variety of linguistic phenomena.

For the time being, the empirical evidence put forward to support functional application models is encouraging indeed: For instance, apart from Baroni and Zamparelli (2010) (as discussed above), Vecchi et al. (2011) separate unseen adjective-noun phrases into plausible and ill-formed ones. Their results suggest that composition by functional application has an important role to play in distributional semantics. However, as admitted by Baroni et al. (2014) themselves, simple multiplicative vector mixture models also "perform fairly well across the board".

Boleda et al. (2012) subject various composition methods for computing distributional representations of adjective-noun phrases to an empirical test motivated from formal semantics, i.e., their capability of discriminating three types of adjectives: intersective, subsective and intensional ones (cf. Kamp, 1975; Amoia and Gardent, 2008). Varying patterns of cosine similarities between phrase and individual word representations as generated by the composition functions under investigation are used as indi-

[3] With regard to the corpus-based induction of these mappings, Baroni et al. (2014) stick to the linear regression approach from Baroni and Zamparelli (2010) discussed above.

cators for differences between the three classes. Their results show that multiplicative vector composition and functional application based on adjective-specific linear maps as proposed by Baroni and Zamparelli (2010) are most appropriate for this task. As an important qualification, it has to be mentioned that the data set used in this study is clearly not representative for the entire spectrum of intersective and subsective adjectives, as it contains only *color-denoting* adjectives in these classes and a relatively small number of adjective types overall.

In fact, relying on a broader data set and framing the task as a two-class problem of discriminating intensional vs. non-intensional adjectives, Boleda et al. (2013) do not observe any difference across the different composition strategies: Neither vector mixture models nor functional application by linear maps are capable of separating the two classes. Interpreting this negative result which is counterintuitive from a theoretical perspective, the authors hypothesize that (i) bags-of-words contexts might not be accurate for modelling the differences between these two classes of adjectives, or (ii) using differences in the cosine similarities between phrase and individual word representations might be inappropriate as an experimental approach for this task.

Comparison. While both these shortcomings apply to vector mixture and functional application models equally, we would like to point out another possible explanation which may challenge a fundamental assumption of the functional application paradigm: From our view, it is questionable whether contextual differences between adjectives and observed adjective-noun pairs are, in general, indicative of the compositional contribution of the adjective to the phrase meaning. Just as much as it seems plausible that the observed contexts of, e.g., *alleged murderer* exhibit very little overlap with the contexts of the intensional adjective *alleged*, we expect the contexts of *red car* or *fast boat* to reveal very little about the meaning of the non-intensional adjectives *red* and *fast*.

For these reasons, and because we anticipate that vector mixture models lean themselves directly to intersective compositional processes underlying attribute meaning in adjective-noun phrases, our approach to capturing this aspect of phrasal semantics will be entirely based on word-based vector representations and their composition.

3.6 Distributional Enrichment of Structured Models

Precursors of distributional enrichment. Distributional semantic models have long been a natural choice for leveraging lexical coverage issues in other models, either knowledge-rich or data-driven, in various application domains such as word sense discovery (Pantel and Lin, 2002) and disambiguation (Miller et al., 2012), selectional preference modeling in computational psycholinguistics (Erk et al., 2010), dependency parsing (Wang et al., 2005) or named entity recognition (Jonnalagadda et al., 2012). The overarching idea in these approaches is that lexical items for which only a sparse or otherwise insufficient feature representation can be provided may be enriched by more

informative features gained from distributionally similar items.

On the other hand, in order to overcome sparsity issues in distributional semantic models themselves, *dimensionality reduction* techniques such as Singular Value Decomposition or Latent Semantic Analysis (Deerwester et al., 1990; Martin and Berry, 2007) have become a de-facto standard in distributional modeling (Turney and Pantel, 2010). These approaches have been proven widely effective in distributional modeling. However, they come at the expense of linguistic intransparency, as the dimensions of the resulting semantic space are no longer semantically interpretable (Hu et al., 2007). Therefore, dimensionality reduction cannot be straightforwardly applied in distributional tasks or settings like ours that require interpretable dimensions of meaning.

Recently, Padó et al. (2013) coined the term *derivational smoothing* for their work on overcoming sparsity issues in a structured distributional model by enriching sparse word representations with information about derivationally related words (e.g., the word vector representing *oldish* is enriched by the representation of *old* such that both of them will be assigned a high semantic similarity to *ancient*). Previous approaches to use morphological information for enhancing distributional models have been presented by Bergsma et al. (2008) and Allan and Kumaran (2003).

Bootstrapping. Besides, distributional enrichment has a close relation to bootstrapping approaches in the tradition of Riloff and Jones (1999), mostly being applied in pattern-based relation extraction approaches. The underlying idea is to acquire, from a limited set of manually approved seed instances, a large amount of additional instances sharing semantic properties with the seeds, thus instantiating the same semantic relation. Bootstrapping is an iterative, semi-supervised process based on the duality of patterns and extracted instances (Brin, 1999): Each iteration of the system yields new candidates, the most reliable ones of which enter the next iteration as additional seeds, or new extraction patterns which can be used further for candidate extraction, respectively. This procedure requires a compromise between the antagonistic goals of *generalization* (for the sake of acquiring a large number of additional instances) and *consistency* (in order to avoid semantic drifts). Confidence metrics being applied in order to ensure consistency range from the proportion of positive and negative extractions generated by a pattern (Riloff and Jones, 1999; Agichtein and Gravano, 2000; Rozenfeld and Feldman, 2006) to association scores between patterns and extracted instances based on mutual information (Pantel and Pennacchiotti, 2008) and probabilistic correlates of precision and recall in random walk processes (Fang and Chang, 2011). Generalization may be achieved by string alignment (Rozenfeld and Feldman, 2006), canonicalization (e.g., replacing named entities by their semantic category; Pantel and Pennacchiotti, 2008) or distributional similarity between a candidate instance and the centroid of previously encountered instances (Agichtein and Gravano, 2000).

Outside information extraction, bootstrapping has also been applied by Zhitomirsky-Geffet and Dagan (2009) who aim at improving the predictability of entailment relations

between words found distributionally similar. To this end, they propose a bootstrapping approach in order to promote those features in dependency-based word vectors that are most likely to enforce lexical entailment in word pairs exhibiting high similarity. In their approach, the compromise between generalization and consistency in representing the meaning of a target word is achieved by selecting, from the close vicinity of the target in the semantic space, those neighbours with the highest overlap in features assumed as indicative for entailment.

Representing words as regions. Erk (2009a,b) represents words as regions in semantic space, not primarily in order to increase density, but for performing word sense disambiguation for polysemous verbs and to support inferences such as hyponymy. Membership of a token vector[4] $\vec{x}$ to a region surrounding a word type vector $\vec{w}$ is predicted using a log-linear model that is trained in a self-supervised manner without recourse to labeled data: Token vectors of w are considered as positive training instances, token vectors of other words as negative ones. Spatial distances in the semantic space hosting $\vec{w}$ and $\vec{x}$ serve as features for classification.

Similarly, Schütze (1998) discriminates different word senses of a polysemous word w by determining the nearest *sense vector* for a *context vector* representation of w. Context vectors are constructed as centroids by summing over all words in the context of w; clustering several context vectors in close proximity yields a sense vector. Hence, sense vectors can be expected to provide a very dense and coherent representation of a region in the vector space and, thus, to have a better chance of matching the observed contexts of an ambiguous word.

Apart from being used for word sense disambiguation, the techniques proposed by Schütze (1998) and Erk (2009a,b) may serve a more general purpose of increasing density in distributional semantic models of different kinds. However, their approaches remain self-contained in the sense that they do not import complementary information from additional semantic sources.

Tensor factorization. Most recently, Zhang et al. (2014) presented a probabilistic tensor factorization approach (cf. Kolda and Bader, 2009) in order to combine semantic information of varying type and provenance. Their model integrates relatedness information from static lexical semantic resources (i.e., synonymy and antonymy relations) and distributional word vectors encoding distributional similarity in a third-order tensor representation. The three modes of the tensor correspond to (i) a lexical semantic relation being encoded for two target words, (ii) the degree of distributional relatedness between them, and (iii) the particular type of relatedness (e.g., relatedness along topical dimensions, taxonomic hierarchies or other ontological relations). Bayesian probabilis-

[4]Token vectors, in these models, represent a word in a particular context, e.g. *supersede knowledge*. They are computed by combining their constituent type vectors, i.e., *supersede* and *knowledge*, in this example (Erk, 2009a,b).

tic factorization methods are applied to decompose the tensor into latent word vectors which can afterwards be used for predicting relatedness scores for originally unobserved pairs of target words. For the task of answering antonymy questions from the GRE data set (Mohammad et al., 2008), the authors demonstrate that this approach is capable of almost doubling the recall (at a slight loss of precision) of detected antonyms by effectively enriching the highly precise, but very sparse lexical relations provided by a combination of WordNet and Roget's thesaurus.

Due to the generality of tensors as a data structure for representing relational data, Zhang et al.'s approach can be adapted to combining arbitrary semantic representations. Aiming at improving potentially sparse structured distributional vector representations by taking additional information from a complementary semantic perspective into account, our own approach to distributional enrichment follows a similar idea while remaining in a purely distributional framework and without relying on any hand-crafted lexical resources.

Comparison. None of the approaches reported in this section is fully compliant with our notion of distributional enrichment. In fact, our goal of improving *structured* distributional semantic models by *distributional means*, relying on *complementary information* from additional distributional models, to be outlined in Chapter 9, combines elements from three of the sources discussed in this section: The bootstrapping method used by Agichtein and Gravano (2000) in the *Snowball* system is similar to distributional enrichment in that they combine a pattern-based extraction approach with distributional candidate filtering using a bag-of-words model. We extend their approach by relying on an auxiliary distributional model that is tailored to complementary information of a more specific kind than that provided by bag-of-words representations. The aspect of taking complementary sources of semantic information into account is most prominent in Zhang et al. (2014). Being based on existing lexical resources, their implementation of this idea is clearly incompatible with our goal of an unsupervised corpus-based approach, though. Procedurally, our account to distributional enrichment is closely related to the construction of centroids from distributional neighbours (Erk, 2009a,b) in order to enhance sparsely represented noun meanings.

4 Distributional Models of Attribute Meaning

In this chapter[1], we develop the main research questions underlying this thesis (Section 4.1), followed by a high-level overview of the methods used to address them (Sections 4.2, 4.3 and 4.4) and the main contributions of the thesis (Section 4.5).

4.1 Research Questions

In proposing corpus-based semantic models of attribute meaning in adjectives and nouns, this thesis addresses the following research questions:

Learning Implicit Knowledge from Text. To what extent can attribute information, being partly in the realm of *world knowledge*, be learned from *textual sources*? World knowledge is of particular relevance for deep natural language understanding. However, it often remains implicit in natural language as it is assumed to be primarily acquired from situational context or perceptual input (Andrews et al., 2009; Barsalou, 2010; Thill et al., 2014) and thus taken for granted by individual speakers in their utterances (Frank and Padó, 2012; Saba, 2007). Therefore, attribute learning provides a challenging example for harvesting implicit knowledge automatically from text.

Large-scale Attribute Learning. With respect to the coverage of attributes and their nature, we raise the task of *attribute learning from text* to a *large scale*. Previous approaches have been limited to a small selective range of attributes (Almuhareb, 2006). Beyond such restricted experimental settings, it is an open question whether attribute learning from textual sources scales to larger inventories of attributes. Moreover, is it possible to fit an attribute model to a largely "universal" scheme of attributes such that it can be flexibly adapted to different attribute-related NLP tasks? Large-scale attribute learning implies a particular challenge as it runs the risk of including more abstract attribute concepts, which might further aggravate the textual acquisition bottleneck discussed above. In order to investigate these questions based on a broad range of attributes that does not commit to a particular domain of interest in the first place, our

[1] Parts of this chapter have been previously published in Hartung and Frank (2010b), Hartung and Frank (2011b) and Hartung and Frank (2014).

approach to large-scale attribute modeling relies on attribute nouns as provided by WordNet (Fellbaum, 1998).

Compositionality. Previous work has attempted to learn attribute knowledge from adjectives in isolation (Almuhareb, 2006; Cimiano, 2006). We argue that these approaches fall short of contextual shifts in adjective meaning, given that adjectives most frequently occur in language as syntactic dependents of nouns. In contrast, we will investigate as to what extent attribute meaning is the result of *compositional semantic processes* between an adjective and a noun when being syntactically combined in an adjective-noun phrase.

Degree of Supervision. Following recent trends in corpus-based modelling of semantic knowledge for specific domains, tasks and needs (Turney and Pantel, 2010), we aim at the development of models for attribute learning from adjectives and nouns that are *unsupervised* or *weakly supervised*.

Specificity and Sparsity in Structured Distributional Models. Our work is framed in *structured distributional models* that are built from syntactic dependencies or lexico-syntactic patterns. These models have the advantage of being able to address specific semantic relations of interest along naturally interpretable dimensions of meaning. Following the discussion in Section 2.3.2 above, we will explore whether it is feasible to tailor a structured model to attribute meaning, taking *compositionality* in adjective-noun phrases into account, while at the same time harmonizing the antagonistic principles of *specificity* and *sparsity*.

In order to address these questions, our approach follows three steps: We first identify attribute-denoting adjectives in an automated machine learning-based classification approach. Then, we broaden the perspective from the level of individual adjectives to the phrase level in order to investigate which attributes are evoked when adjectives and nouns are composed in adjective-noun phrases. Finally, we propose a novel framework for distributional enrichment that aims at alleviating sparsity issues of structured distributional models by integrating lexical semantic information about attribute meaning in adjectives and nouns from complementary distributional sources.

4.2 Identifying Attribute-denoting Adjectives

Background and Motivation. Previous work has developed a separation between attribute-denoting and relational adjectives (Boleda, 2006; Raskin and Nirenburg, 1998), which is highly relevant for knowledge representation and ontology learning. In a machine learning-based classification experiment capitalizing on distributional features extracted from corpus data, we show how this separation can be automatically carried

out in order to focus on relevant classes of adjectives for knowledge acquisition and particularly attribute learning.

Attribute learning, as initiated by Almuhareb and Poesio (2004) and Almuhareb (2006), aims at the automatic acquisition of concept representations in terms of attribute-value sets from natural language corpora. For example, from the co-occurrence of a noun and a **attribute-denoting adjective** in a phrase such as *red car*, we can infer that (i) members of the concept *car* have an attribute COLOR, that (ii) *red* is one of its possible values, and that (iii) the particular exemplar being referred to in the phrase has the value *red* for COLOR.

In relation learning tasks, the goal is to discover (non-taxonomic) relations between previously established concepts (Buitelaar et al., 2005). For this purpose, **relational adjectives** provide a valuable source of information. For instance, an adjective-noun phrase such as *agricultural equipment*, being composed of the relational adjective *agricultural* and a noun referring to the concept EQUIPMENT, is indicative of the semantic relation EQUIPMENT *to be used in* AGRICULTURE (Miller, 1998).

Adjective-noun phrases as introduced in the examples above are a particularly rich source for both learning attribute and relation learning, as they abound in natural language and can be easily detected in corpora without the need for deep syntactic analysis. On the downside, the distinction between attribute-denoting and relational adjectives is a critical prerequisite in order to (i) determine which adjectives are suitable for either attribute or relation learning and (ii) for appropriately encoding the acquired knowledge in ontologies, given that attributes and relations are formally disjoint components of an ontology (Cimiano, 2006) and, therefore, different formal requirements apply for their representation.

The formal distinction between attributes and relations is of importance for several NLP applications as well, e.g. question answering systems operating as interfaces between natural language queries and structured knowledge bases. Attribute-denoting adjectives and relational ones require different mappings from their natural language meaning into formal queries to the knowledge base (see recently Walter et al. (2014)).

Previous attempts to corpus-based attribute learning from adjectives (Almuhareb, 2006) have entirely neglected the distinction of different semantic types of adjectives. In our work on adjective classification (to be presented in Chapter 5.2), we aim at automatically separating attribute-denoting adjectives from relational ones, in order to provide a more reliable basis for attribute learning from text.

Methodology. Our classification approach involves two steps: First, we assess the validity of the classification scheme being adopted in a corpus study based on human annotations. Second, we present a machine learning approach for automatically classifying adjectives into attribute-denoting and relational lexical types.

In our annotation experiment, we observe that token-level annotation for lexical adjective types is time-consuming and difficult. At the same time, careful analysis of the

annotated corpus reveals that property-denoting and relational adjectives constitute stable classes of lexical types, with only few occurrences of class shifts observed at the token level. This ability of an adjective to change its class on the token level will be denoted as *class volatility*. A second observation is that features effectively that may be used to separate the two classes in a machine learning approach essentially operate on the level of lexical types, i.e., they focus on grammatical properties that may not be observable from single token occurrences in each individual context.

These insights suggest a type-based classification approach, similar to work in semantic verb classification by Miyao and Tsujii (2009). Based on the observed low class volatility, we opt for a weakly supervised training regime, using the token-level annotations from our annotated corpus as seeds for the acquisition of a large training set that is heuristically labeled by annotation projection.

The resulting classification model facilitates the identification of adjectives that are out of the scope of attribute learning. Thus, adjective classification also lays the foundation for our approach to the attribute selection task to be described in the next section.

4.3 Compositional Representations of Attribute Meaning in Adjective-Noun Phrases

Having separated adjectives into different semantic classes as discussed in the previous section, we aim at predicting the attribute(s) being evoked when attribute-denoting adjectives and nouns are composed in adjective-noun phrases. We refer to this problem as *attribute selection* and propose a structured distributional model which captures this aspect of the compositional semantics of adjectives and nouns.

Attribute Selection. We define *attribute selection* as the task of predicting the attribute meaning impliticly conveyed by a property-denoting adjective in composition with a noun without being overtly realized on the textual surface. Consider the following examples:

(17) *hot summer* → TEMPERATURE

(18) *hot debate* → EMOTIONALITY

(19) *hot soup* → TASTE/TEMPERATURE

The adjective *hot* may denote (a value of) attributes such as TEMPERATURE, TASTE or EMOTIONALITY. These adjectives can be combined with nouns such as *summer*, *soup* or *debate* which can be characterized in terms of attributes as well. For instance, *debate* might elicit attribute meanings such as EMOTIONALITY, DURATION, DEPTH or VOLUME, among others. It is by way of the composition of adjective and noun that specific attributes are *selected* (Pustejovsky, 1995) from the compositional semantics of the adjective and the noun, and lead to a disambiguation of the adjective and possibly the noun.

From a knowledge acquisition perspective, the examples in (17)–(19) clearly indicate the importance of a compositional approach to attribute selection. Purely adjective-based acquisition methods (Almuhareb, 2006; Cimiano, 2006) miss out the contribution of the noun. Thus, they do not exploit the disambiguation potential that is inherent in compositional semantics and do not take any knowledge into account that can be derived from co-occurrence statistics of the adjective and the noun. These shortcomings of purely adjective-based approaches would result in an overgeneration of knowledge base entries, as shown in (20) for the concept *summer*[2]:

(20) a. TEMPERATURE(*summer*)=*hot*
b. * EMOTIONALITY(*summer*)=*hot*
c. * TASTE(*summer*)=*hot*

Note, however, that adjective-noun phrases may still be ambiguous with regard to their implicit attribute meaning; (19) provides an example. Resolving ambiguities of this kind is out of the scope of our approach to attribute selection, as this requires additional context beyond the phrase to be taken into account. Instead, our model is designed to yield *all* attributes that are licensed by the compositional semantics of a particular phrase, irrespective of surrounding clausal or sentential context.

Roots in Formal Semantics. Attribute selection is rooted in formal semantics. One of the claims in the *Generative Lexicon* (GL) theory (Pustejovsky, 1995) is that the compositional semantics of intersective adjectives and nouns is brought about by a process denoted as *selective binding*, where the adjective selects one out of several possible roles or attributes from the noun. In the original statement of the theory, adjectives select properties which are part of a particular *qualia role*[3]. In our framework, these properties (as elicited by adjectives) are accessed from a more specific level of granularity, i.e., the attribute nouns under which they may be subsumed. This can be seen from the examples in (21) below:

(21) a. *cool summer*
b. *long summer*

While GL would assign both *cool* and *long* in these examples to the *formal* quale from the meaning of *summer*, our approach provides a more explicit semantics by predicting the attributes TEMPERATURE and DURATION, respectively. It is an open question whether attribute selection will be able to cover properties from all four qualia roles effectively. The experiments conducted in this thesis will provide first insights into this issue.

[2] And analogously for *debate* and *soup* from (18) and (19).

[3] Four different roles are offered by GL: a *formal*, *constitutive*, *telic* and *agentive* role. Pustejovsky (1995) does not provide a concrete implementation of the theory; i.e., in order to apply GL to semantic tasks, population of qualia roles with individual properties is an open issue which has been found very hard to address in a corpus-based induction approach (Cimiano and Wenderoth, 2007).

Methodology. Our approach to attribute selection as proposed in this thesis is fully framed in a structured distributional semantic model. Thus, it requires no manual work apart from prior specification of (i) the attribute inventory to be investigated and (ii) linguistic cues to the relation between adjectives and nouns to attributes. Given such cues, instantiations of these relations are extracted fully automatically.

As discussed in Section 2.3.2, tailoring a distributional model to a particular semantic relation (i.e., the attribute relation between adjectives and nouns in our case) requires reconciling the conflicting goals of specificity and sparsity. Ideally, attribute knowledge could be extracted from corpora by searching for patterns as in (22)[4]:

(22) the $\mathbf{color}_{\mathbf{ATTR}}$ of the $\mathbf{car}_{\mathbf{NOUN}}$ is $\mathbf{blue}_{\mathbf{ADJ}}$

However, linguistic patterns that explicitly relate nouns, adjectives and attributes are very rarely observed in corpora. We avoid these sparsity issues by reducing the triple

$$r = \langle noun, attribute, adjective \rangle$$

to tuples

$$r' = \langle noun, attribute \rangle \qquad \text{and} \qquad r'' = \langle attribute, adjective \rangle,$$

as suggested by Turney and Pantel (2010) for similar tasks. Both r' and r'', as instantiated by (23) or (24), for instance, can be observed much more frequently in text corpora than r.

(23) the $\mathbf{color}_{\mathbf{ATTR}}$ of the $\mathbf{car}_{\mathbf{NOUN}}$

(24) $\mathbf{blue}_{\mathbf{ADJ}}$ $\mathbf{color}_{\mathbf{ATTR}}$

Moreover, this enables us to model adjective and noun meanings as distinct semantic vectors in the same semantic space being spanned by attributes as dimensions. Based on these semantic representations, we make use of vector composition operations in order to reconstruct r from r' and r''. This, in turn, allows us to obtain composed vector representations for complete noun-attribute-adjective *triples* such as $\langle \text{car}, \text{COLOR}, \text{blue} \rangle$, from which the attribute(s) implicitly hidden in the phrase semantics can be selected in a fully unsupervised manner.

Hence, vector composition serves a double purpose in our models: On the one hand, it reflects the compositionality that is inherent to the attribute selection task, on the other hand, it provides a handle for overcoming sparsity issues. This general idea for resolving the antagonism between specificity and sparsity is implemented in two variants of structured distributional models for attribute selection: a pattern-based and a dependency-based model. The latter is extended to a topic-based distributional model by inducing attribute-specific latent topics from weakly supervised variants of Latent Dirichlet Allocation (Blei et al., 2003).

[4]Note that we are alternating between two levels here: While NOUN and ADJ refer to word classes, ATTR denotes a semantic category. In this thesis, we assume that attribute knowledge is always expressed via nouns. Therefore, the terms *attribute* and *attribute noun* will often be used interchangeably.

4.4 Distributional Enrichment

Background and Motivation. Structured distributional models are typically designed to extract distributional information that characterize specific semantic relations or properties. This focus on specificity usually goes along with sparsely populated vector representations if the target relation is either rarely observed in the corpus or tends to occur in a variety of surface realizations that are not comprehensively covered by the patterns or dependency paths used for co-occurrence extraction.

Apart from decomposing a complex target relation into more elementary ones as dicussed for the attribute relation in Section 4.3 above, we propose distributional enrichment as an additional strategy for alleviating sparsity issues in structured distributional models, while at the same time preserving their particular strengths with respect to specificity.

Methodology. Distributional enrichment aims at enhancing structured vector representations of individual target words by considering *complementary* distributional information in terms of their semantic neighbours. These are acquired from complementary distributional sources. The basic idea underlying distributional enrichment is formalized in Equation 4.1 in a slightly simplified form[5]:

$$\vec{w'_{attr}} = \vec{w_{attr}} \oplus \sum_{\vec{n} \in V_{aux}} \mu(\vec{n}) \cdot \lambda(\eta(\vec{w}), \vec{n}) \cdot \chi(\vec{n}) \tag{4.1}$$

Here, $\vec{w_{attr}}$ denotes the original structured vector representation of a target word w in an attribute-based distributional model, V_{attr}. The result of distributional enrichment, i.e., an enhanced version of this vector in the attribute model, is denoted as $\vec{w'_{attr}}$.

The enrichment process essentially works by computing a *centroid* of structured vector representations. The vectors taking part in the centroid are selected from an auxiliary distributional model, V_{aux}, which is designed to provide distributional information complementary to V_{attr}. Note that V_{attr} and V_{aux} are of different dimensionality; therefore, we use mapping functions $\eta(\cdot)$ and $\mu(\cdot)$ to map vector representations from V_{attr} to V_{aux} and vice versa.

For building the centroid (cf. the sum in Equation 4.1), we iterate over all vector representations $\vec{n}$ in V_{aux}, using $\mu(\vec{n})$ to retrieve the structured representation of the target word n in V_{attr}. This structured vector is weighted by (i) a scalar $\lambda(\eta(\vec{w}), \vec{n})$ that determines the strength of relatedness between w and n in V_{aux}, and (ii) an indicator function $\chi(\vec{n})$ which decides whether or not $\vec{n}$ is an appropriate semantic neighbour of $\vec{w}$ and becomes a member of the centroid.

In previous related work, supervised learning has been applied to centroid induction in one and the same distributional model (Erk, 2009a); in our approach to distributional

[5]For the sake of comprehensibility in this introductory context, we are abstracting from some additional parameters and constraints here. Their full specification is deferred until Section 9.3.

enrichment, we address the problem across complementary semantic spaces of different dimensionality and in an unsupervised manner.

We apply distributional enrichment to structured distributional attribute models as discussed above. In this scenario, two types of semantic neighbours are considered in order to enhance sparse attribute-based target vectors: (i) distributionally similar nouns, and (ii) adjectives being observed as modifiers of these sparse nouns. Thus, the principle of meaning representation along interpretable, attribute-based dimensions is kept intact, while at the same time the overall density of the attribute space will be increased.

4.5 Contributions of this Thesis

Summarizing this chapter, the major contributions of this thesis will be as follows:

1. We present empirical evidence for an **adjective classification scheme** for separating lexical adjective types into a **attribute-denoting** and a **relational** class so that adjectives to be used for attribute learning can be automatically distinguished from other types that are more suitable for relation learning tasks. This distinction has been largely neglected in prior work. We show that an effective identification of attribute-denoting adjectives has a substantial impact on attribute selection from adjectives.

2. We define a new task of **attribute selection from adjective-noun phrases**. This task consists in eliciting attribute(s) from a pre-defined inventory that are implicit in the compositional semantics of an adjective and a noun. We experiment with different inventories of attributes (of varying breadth and cardinality) that are compiled from the attribute nouns contained in WordNet.

3. We develop a novel class of structured distributional models for representing adjectives and nouns in a semantic space spanned by attributes as dimensions. These models make use of vector composition in order to reflect **compositional processes** in adjective-noun phrases and to alleviate the conflict between **specificity and sparsity** that is inherent in distributional models. Thus, contrary to prior work on attribute learning, our models enable attribute selection from phrasal contexts in an unsupervised manner.

4. The attributes harvested by our method are ontologically grounded, which situates our work at the interface of distributional semantics and **knowledge induction from textual sources**. We demonstrate that distributional semantics offers promising methods to address the challenges that ontological knowledge is very abstract in nature and usually remains **implicit in natural language**.

5. We compare various instantiations of attribute models built on pattern-based and dependency-based distributional information as well as attribute-specific topics induced from weakly supervised versions of Latent Dirichlet Allocation. These models will be evaluated on the attribute selection task framed for several inventories of attribute concepts, up to a **large-scale set of 260 attributes**. In practical application contexts, this inventory could be tailored to specific tasks by grouping individual attributes together, selecting subsets or mapping them to a particular target domain.

6. We provide a thorough performance analysis of our best-performing attribute selection model. This investigation reflects strengths and weaknesses of the model and sheds light on the impact of a variety of **linguistic factors** involved in attribute selection, e.g., the relative contribution of adjective and noun meaning.

7. We present a framework for **distributional enrichment** of structured distributional models. Its potential for alleviating sparsity issues inherent in such models is demonstrated by successfully applying distributional enrichment to distributional attribute models.

8. We release three annotated data sets for adjective classification and attribute selection used as gold standards throughout the experiments reported in this thesis, making them openly available to the research community.

5 Classification of Adjective Types for Attribute Learning

In this chapter, we empirically investigate the task of classifying adjectives into property-denoting vs. relational types. As only property-denoting adjectives are informative for corpus-based approaches to attribute learning, this distinction is highly relevant for the attribute models to be proposed in this thesis in that it facilitates candidate selection of adjectives that are useful for attribute learning. In previous work on attribute selection from adjectives (Almuhareb, 2006), the separation between property-denoting and relational adjectives has been entirely overlooked, with negative effects on performance. Therefore, our goal in the present study is to learn a classification model for identifying property-denoting adjectives at high levels of precision in order to be applied as a filter during attribute selection. Moreover, we are interested in analyzing which features are most important for this classification task.

Based on an adjective classification scheme that separates adjectives into subtypes relevant for ontology learning, we proceed in two steps: First, we assess the validity of the scheme in a human annotation task. In a second step, the resulting annotations are exploited in order to train a classification model for separating adjectives into property-denoting and relational lexical types in a weakly supervised manner.[1]

5.1 Corpus Annotation and Analysis

As a starting point for distinguishing adjective classes relevant for ontology learning, we adhere to the three-way classification that has been proposed for Catalan adjectives by Boleda (2006). According to the class labels (**b**asic, **e**vent-related and **o**bject-related), we name this classification scheme *BEO classification*. In the following, we give a brief overview of the properties exhibited by the BEO classes, paying special attention to their relevance for ontology learning.

5.1.1 Classification Scheme

Basic Adjectives. Basic adjectives denote values of an attribute exhibited by an entity. In case of *scalar* attributes (Levinson, 1983; Hatzivassiloglou and McKeown, 1993; de Melo and Bansal, 2013), adjectives either denote points or intervals on the scale, as in

[1] Major parts of the content of this chapter have been previously published as Hartung and Frank (2010a) and Hartung and Frank (2014).

(25) and (26), respectively. If the values of an attribute cannot be ordered on a scale (as for SHAPE, for instance), an adjective denotes an element in the set of possible values of the attribute, as in (27).

(25) blue car ↔ COLOR(car)=blue

(26) young girl ↔ AGE(girl)=young

(27) oval table ↔ SHAPE(table)=oval

Event-related Adjectives. These adjectives modify an associated event the referent of the noun takes part in, as illustrated by the following paraphrases (cf. Lapata, 2001):

(28) eloquent person ↔ person that *speaks* eloquently

(29) comfortable chair ↔ chair that is comfortable to *sit on*

(30) interesting article ↔ article that is interesting to *read*

Object-related Adjectives. This class comprises adjectives that are morphologically derived from a noun, denoted as $A_{/N}$ and N_b, respectively, as in (31)–(33). In these cases, N_b refers to an entity that acts as a semantic dependent of the head noun N.

(31) economic$_{[A_{/N}]}$ crisis$_{[N]}$ ↔ crisis of the *economy*$_{[N_b]}$

(32) political$_{[A_{/N}]}$ debate$_{[N]}$ ↔ debate on *politics*$_{[N_b]}$

(33) philosophical$_{[A_{/N}]}$ question$_{[N]}$ ↔ question about *philosophy*$_{[N_b]}$

Note that the paraphrases given in these examples move object-related adjectives into the proximity of filling a semantic role (Fillmore, 1968) offered by the semantics of the head noun. In (31), *economic* relates to the EXPERIENCER role of *crisis*; (32) and (33) exemplify THEME roles.

BEO classes in Formal Semantics. Note that, from a formal semantics perspective, basic adjectives are mostly *intersective* in the sense that the meaning of an adjective-noun phrase entails both the adjective and the noun meaning individually (Amoia and Gardent, 2007):

$$[AN] \models N$$
$$[AN] \models A$$

On the other hand, many event-related and object-related adjectives belong to the category of *subsective* modifiers. The underlying inferential pattern is characterized by the fact that the phrase meaning entails only the noun, but not the adjective meaning (Amoia and Gardent, 2007):

$$[AN] \models N$$
$$[AN] \not\models A$$

BEO classes in Ontology Learning. As seen above, the BEO classes distinguish properties (basic and event-related adjectives) from relational meanings (object-related adjectives). This distinction can be utilized in ontology learning for the acquisition of property-based concept descriptions and semantic relations between concepts, respectively.

5.1.2 Annotation Process

Methodology. To validate the BEO classification scheme, we ran an annotation experiment with three human annotators. We compiled a list of 200 high-frequency English adjectives from the British National Corpus[2] and for each of them randomly extracted five example sentences from the written section of the BNC. The annotators labeled each item as `BASIC`, `EVENT`, `OBJECT` or `IMPOSSIBLE`. The latter was supposed to be used in case the annotators were unable to provide a label due to erroneous examples[3], insufficient context, or instances belonging to alternative classes of adjectives not considered here.

Ambiguities between BEO Classes. The most notable ambiguity among BEO classes holds between basic and event-related adjectives. Consider the following competing analyses for *fast horse*:

(34) a. fast horse $\leftrightarrow$ SPEED(horse)=fast
b. fast horse $\leftrightarrow$ horse that *runs* fast

We argue that this ambiguity sheds light on the difference between *independent* and *founded* properties[4] of an object (cf. Guarino, 1992). For disambiguation, we propose the inference patterns[5] in (35).

[2] We used version 3 of the BNC XML Edition, available from: `http://www.natcorp.ox.ac.uk/`

[3] Part-of-speech tagging was the primary source of errors here.

[4] In its original statement, the notion of *foundation* is defined as follows: "For a concept α to be founded on another concept β, any instance χ of α has to be necessarily associated to an instance ϕ of β which is not related to χ by a part-of relation" (Guarino, 1992). We extend this notion from concepts to properties, arguing that event-based adjectives denote founded properties that are necessarily associated with an implicit event.

[5] Note that these patterns are mutually exclusive: (35a) applies to examples such as *comfortable chair* and *interesting article* in (29) and (30), where ENT fills the PATIENT role of EVENT. In contrast, *eloquent person* in (28) can be identified as event-based by (35b) only, as ENT acts as the AGENT of EVENT here (cf. Lapata, 2001). We expect that disambiguating basic and event-related readings should work best if (35a) is constrained such that EVENT may **not** be instantiated by perception verbs such as *look*, *feel*, *taste* etc.

	Annotator 1	Annotator 2	Annotator 3
Annotator 1	—	0.762	0.235
Annotator 2	0.762	—	0.285
Annotator 3	0.235	0.285	—

Table 5.1: Agreement figures in terms of Fleiss' κ

	BASIC	EVENT	OBJECT	IMPOSS
κ	0.368	0.061	0.700	0.452

Table 5.2: Category-wise κ-values for all annotators

(35) a. ENT(ity) can be attested to be ADJ(ective) by EVENT.
b. If ENT was not able to EVENT, it would not be an ADJ ENT.

Applied to (34), these patterns indicate that, in the case of a horse, being fast should be formalized as a property that is founded on the horse's inherent ability to run (or, at least, to move). If this ability was absent, it would no longer be possible to qualify the horse as being fast (cf. (35b)). Hence, we prefer an event reading for *fast horse*.

5.1.3 Agreement Figures

Table 5.1 displays agreement figures for our annotation experiment in terms of Fleiss' Kappa[6] (Fleiss, 1971). Total agreement between all three annotators amounts to $\kappa = 0.404$. Note that we observe substantial agreement of $\kappa = 0.762$ between two of the annotators, which suggests that the upper bound is higher than the observed overall agreement. Table 5.2 displays the overall agreement figures broken down into the four class labels. These results underline our intuition that the distinction between the classes BASIC and EVENT is very difficult even for human subjects.

This is corroborated by a thorough analysis of the cases of annotator disagreements in Table 5.3 on the facing page. This table overviews all cases where one annotator disagrees with the other two. The rightmost column indicates the total number of 2:1-disagreements for each class. The missing mass is due to the IMPOSSIBLE class. As can be seen, the situation where two annotators vote for BASIC, while one prefers the EVENT class, accounts for most of the disagreements among the annotators (172 cases in total). The following instances, taken from the set of disagreement cases, exemplify the problems encountered by the annotators when being confronted with the BASIC vs. EVENT distinction:

[6] κ measures the agreement among annotators in classification tasks. Its values reflect the degree of agreement *above chance*: $\kappa = 1$ indicates perfect agreement, whereas $\kappa = 0$ indicates an agreement that is merely due to chance (Fleiss, 1971).

		1 voter			
		BASIC	EVENT	OBJECT	Total
2 voters	BASIC	–	172	16	283
	EVENT	18	–	1	21
	OBJECT	54	10	–	66

Table 5.3: Distribution of Disagreement Cases over Classes

	BASIC+EVENT	OBJECT	IMPOSS
κ	0.696	0.701	-0.003

Table 5.4: Category-wise κ-values, binary classification scheme

(36) *Any changes should only be introduced after **proper** research and costing, and after an initial experiment.*

(37) *Matthew thought his mother sounded very young, her voice **bright** with some emotion he could not quite define.*

Resorting to (35), we argue for an event-based reading of *proper* in (36) (e.g., "research that has been properly conducted"), while *bright* in (37) should be given a basic interpretation.

As becomes evident from the quantitative analysis in Table 5.3 and these examples, the ambiguity between basic and event-related adjectives is the primary source of disagreement in our annotation experiment.

5.1.4 Re-Analysis: Binary Classification Scheme

This observation led us to re-analyze our data using a binary classification that collapses basic and event-related adjectives into one class. This re-analysis is merely a shift in granularity, as both basic and event-related adjectives denote properties, whereas object-related adjectives denote relations. Re-analyzing the data in this way improves overall agreement to $\kappa = 0.69$. See Table 5.4 for detailed agreement figures.

The remaining disagreements between annotators have been manually adjudicated. After adjudication, the data set contains 689 adjective *tokens* that are unambiguously annotated, given the respective context, as denoting a property, while 138 tokens are labeled as relational. In total, 190 (out of 200) lexical adjective *types* are covered. Again, the missing mass is due to items marked as IMPOSSIBLE by at least one annotator.

Type	after adjudication			before adjudication	
	#ATTR	#REL	#ambig.	#ATTR	#REL
black	2	2	0	2	2
male	4	1	0	4	1
personal	2	2	1	2	3
political	2	2	1	1	4
white	3	1	0	3	1
detailed	5	0	0	4	1
mental	0	5	0	2	3
military	0	5	0	1	4
nuclear	0	5	0	1	4
professional	0	5	0	3	2
regional	0	5	0	1	4
technical	0	4	0	1	3

Table 5.5: Overview of volatile adjectives in the data set

5.1.5 Class Volatility

In order to judge the possibility of a *type-based* automatic adjective classification, we need to quantify the degree of class volatility as observed in the annotated corpus, i.e., the proportion of lexical types that are assigned alternating class labels at the token level.

We identified 12 adjectives that are volatile in the sense that they can undergo a type shift between basic and event-related vs. object-related adjectives[7] on the token level. Thus, the proportion of volatile types in the data set amounts to 6.3%[8].

In a further adjudication step, the number of volatile types could be reduced to 5 by evaluating fine-grained interpretation differences. Table 5.5 displays the full list of adjectives considered before and after adjudication, including their frequency distribution over the two classes. The subset of adjectives established as "true volatiles" after adjudication is given in boldface. In the following, we discuss some typical cases of shifts between property-denoting and relational interpretations of adjectives.

Shifts from ATTR to REL

(38) a. *Certain stations in **black** rural areas or town locations were expected to be used exclusively by Africans.*

b. *The suburban commuter station was emphatically a **male** preserve at certain times of day.*

[7] Henceforth, we will refer to these binary classes as ATTR(ibutive) and REL(ational).

[8] In a selective investigation on more representative data, class volatility turns out to be only slightly higher (cf. Section 5.2.4).

Both *black* in (38a) and *male* in (38b) have to be assigned a relational interpretation even though the basic meaning of these adjectives is property-denoting. This shift can be analyzed as a metonymic process where the adjective is re-interpreted as referring to an entity to which the respective property applies (concretely: *black people*). This entity, in turn, acts as an argument in a relation with the head noun. Thus, *black rural areas* in (38a) and *male preserve* in (38b) can be paraphrased as *rural areas inhabited by black people* and *a preserve occupied by male people*, respectively.

Shifts from REL to ATTR

Clear Contextual Shifts. In the following example, we observe a shift from a relational to a property-based adjective reading:

(39) *But then aren't you taking a **political** stance, rather than an aesthetic one?*

(40) *Their reasons for study are various and include simple **personal** interest and skill acquisition in connection with present or possible future employment.*

We argue that a *political stance*, as in (39), does not denote a particular *stance on politics* (which would be the obvious relational interpretation), but a property: a stance that is *politically motivated* or *held for political reasons*. The given context crucially elicits the class-delineating function of the adjective, in that different subtypes of stances are contrasted.

The same holds for (40): Again, *personal* denotes a property that delineates a particular subtype of *interest*. This yields a semantic interpretation that is closer to a reflexive (*someone's own interest*) than to a relational reading (*someone's interest as a person/related to a person*).

Ambiguities. The following examples are considered ambiguous between a reading that has been shifted from relational to attributive and their original relational reading:

(41) *By offering a range of study modes and routes, including part-time associate status, individuals are encouraged to use the course for a variety of **personal** purposes.*

(42) *Owing to unexplained **political** pressures, General Choi then left the country.*

Both a reflexive and a subjective interpretation (see discussion above) are possible for *personal purposes* in (41). Analogously, there are two possible readings for *political pressures* in (42): Either the adjective is metonymically coerced to a noun reading (*people involved in politics*; see discussion of (38a) above) in order to fill the AGENT role of the noun, or the pressures are conceived of as being exerted for *political reasons*.

Comparing the examples in (39) and (40) to those in (41) and (42) sheds light on the possible influence of the head noun on the interpretation of the adjective. We presume that prototypical shifts as in (39) and (40) are licensed by a particular class of nouns we may call *psychological nouns*. Besides *interest* and *stance*, also *attitude*, *assessment*

and *confidence*, among others, might be representatives of this class, thus licensing the same shift in the context of adjectives such as *personal* or *political*. A more thorough investigation of this hypothesis, however, is beyond the scope of this work.

5.2 Automatic Type-based Classification of Adjectives

In this section, we report the results of a machine learning experiment addressing the feasibility of an automatic corpus-based classification of adjectives on the type level. We restrict the task to the distinction between property-denoting and relational adjectives in the first place as we are not aware of any overt features that are (i) sufficiently discriminative to capture the fine-grained distinction between basic and event-based adjectives in borderline cases such as 34 and (ii) frequently observable in corpora. Our particular focus in this experiment is on determining a feature set that yields robust performance on the binary classification task.

5.2.1 Features for Classification

Our classification approach is based on the observation that property-denoting and relational adjectives systematically differ with regard to their behaviour in certain grammatical constructions. These differences can be captured in terms of lexico-syntactic patterns (Amoia and Gardent, 2008; Beesley, 1982; Raskin and Nirenburg, 1998; Boleda, 2006). We cluster these patterns into groups (see Table 5.6[9]):

- Group `I`: features encoding **comparability**
- Group `II`: features encoding **gradability**
- Group `III`: features encoding **predicative use**
- Groups `IV` and `V`: features encoding the use in **particular constructions**
- Group `VI`: feature encoding **morphological derivation** from noun

The features from groups `I`–`V` encode grammatical properties that can be found with property-denoting adjectives only, while relational adjectives do not license them. As a positive feature for relational adjectives, we consider morphological derivation from nouns (group `VI`), e.g. *criminal – crime*, *economic – economy*). This information was extracted from the CELEX2 database (Baayen et al., 1996).

[9]The pattern descriptions used in the table make use of part-of-speech tags according to the Penn Treebank nomenclature Marcus et al. (1993).

Group	Feature	Pattern	Example
I	as	`as JJ as`	***as cheap as*** *possible*
	comparative-1	`JJR NN`	*halogen produces a* ***brighter light***
	comparative-2	`RBR JJ than`	***more famous than*** *your enemies*
	superlative-1	`JJS NN`	*this is the* ***broadest question***
	superlative-2	`the RBS JJ NN`	***the most beautiful buildings*** *in Europe*
II	extremely	`an extremely JJ NN`	***an extremely nice marriage***
	incredibly	`an incredibly JJ NN`	***an incredibly low downturn***
	really	`a really JJ NN`	***a really simple solution***
	reasonably	`a reasonably JJ NN`	***a reasonably clear impression***
	remarkably	`a remarkably JJ NN`	***a remarkably short amount*** *of time*
	very	`DT very JJ`	*gets onto* ***a very dangerous*** *territory*
III	predicative-use	`NN (WP|WDT)? is|was|are|were RB? JJ`	*my digital* ***camera is nice***
	static-dynamic-1	`NN is|was|are|were being JJ`	*the current* ***unit was being successful***
	static-dynamic-2	`be RB? JJ .`	***Be absolutely certain:***
IV	one-proform	`a/an RB? JJ one`	***a hard one***
V	see-catch-find	`see|catch|find DT NN JJ`	*90%* ***found the events relevant***
VI	morph	adjective is morphologically derived from noun	*culture* → *cultural*

Table 5.6: Set of features used for classification.

5.2.2 Heuristic Generation of Training Instances from Seeds

A major problem we encounter with the features presented above is their severe sparsity. Applied to our annotated corpus of 1000 sentences, the complete feature set yields only 10 hits.

Given the results of our corpus analysis in Section 5.1.5, however, we can raise the classification task to the type level, under the proviso that class volatility is limited to only a small number of adjective types and particular contextual occurrences. Under this assumption, we use our annotated data set as seed material for heuristically labelling adjective tokens in a large unannotated corpus. In this process, the unanimous class labels gathered from the manually annotated corpus are projected to the unannotated data. This means that potential class changes on the token level are completely disregarded.

5.2.3 Data Set Construction

Using the heuristic annotation projection technique described above, we created two data sets which provide the training and evaluation data for our classification experiments.

Data Set 1. The first data set we created is based on the manually annotated corpus described above. We identified all adjective types in the corpus that exhibit perfect agreement across all annotators and are not found to be volatile. This yields 164 property-denoting and 18 relational types, which we use as seeds for heuristic token-level annotation. For each lexical adjective type, we acquired a corpus of 5000 sentences from a subsection of the ukWaC corpus (Baroni et al., 2009) to which the labels from the annotated corpus were projected as described in section 5.2.2. We refer to this data set as DS1.

Data Set 2. In order to assess the soundness of our features on a larger and possibly more representative sample and to evaluate whether our method of heuristic annotation projection can be generalized to different data sets, we also compiled a gold standard of property-denoting and relational adjectives from WordNet 3.0.

Like any other part-of-speech category, adjectives in WordNet are organized in *synsets*, i.e., sets of (nearly) synonymous types. Every synset reflects fine-grained meaning differences in terms of *word senses*. All lexical knowledge in WordNet is encoded by semantic relations between word senses. The information of interest for our task is captured by the relations *attribute* and *pertainymy* (Miller, 1998): Presence of an *attribute* relation between an adjective and a noun sense indicates that the noun denotes a property and the adjective specifies a possible value of this property. A *pertainymy* relation[10]

[10]Note that the pertainymy relation in WordNet is uni-directional as it contains only links from adjectives

	DS1	DS2
Data Source	manual annotation	WordNet 3.0
Num. ATTR Types	164	246
Num. REL Types	18	140
Num. Training Tokens per Type	5000	5000
Labeling Procedure	heuristic annotation projection	WordNet relations `attribute` and `pertainymy`
Evaluation Mode	10-fold cross validation	train/test split (80%/20%)

Table 5.7: Characteristics of data sets used in adjective classification experiments

linking an adjective and a noun sense indicates a relational adjective meaning. If neither an *attribute* nor a *pertainymy* relation is specified for a given adjective, nothing can be inferred regarding the binary classification considered here.

For the construction of our gold standard, we collected all adjectives from WordNet that are unambiguously property-denoting or relational, meaning that *all* of their senses are marked with either the *attribute* or the *pertainymy* relation. This yields 3727 property-denoting and 3655 relational types (i.e., roughly one third of the overall 21486 adjective types in WordNet). We only considered adjectives with more than 2000 occurrences in the same subsection of the ukWaC corpus used for the construction of DS1. The final data set comprises 246 property-denoting and 140 relational adjective types. Again, we extracted up to 5000 sentences from ukWaC for each of these adjectives, and assigned them the class labels ATTR and REL, respectively. The resulting data set is referred to as DS2. The characteristics of both data sets are summarized in Table 5.7.

5.2.4 Experimental Evaluation

Classification algorithms. For evaluating our classification approach, we use two machine learning-based classification frameworks: *Decision Trees* and *Logistic Boosting*. Both of them provide an expressive output that can be used to gain insights into the feature space in order to determine the impact of different features or dependencies among them.

Decision trees represent the space of training instances in a tree structure. Internal nodes in the tree correspond to a subset of instances that are covered by splitting the possible values of one feature into several branches. Features to be split and the particular value to be used are selected according to the *information gain* provided by each

to their morphological base nouns, but not from derived nouns to base adjectives. For instance, *cultural* and *culture* or *dental* and *tooth* are linked by pertainymy, while no such link exists between *short* and *shortness*.

split. An unseen instance is classified by following the corresponding path in the tree from the root to a leaf node which provides the classification. Therefore, a decision tree can also be considered as a cascade of if-then classification rules. (Mitchell, 1997; Witten and Frank, 2005)

Logistic boosting is a meta-learning approach combining several complementary base classifiers into one ensemble classification model. Base classifiers are usually simple – in our case, they are one-level decision trees. Boosting implements an iterative forward selection procedure that starts with an empty ensemble. In each iteration, the base classifier maximizing the predictive performance of the ensemble as a whole is added. Throughout this process, it is guaranteed that the next model is particularly suitable for those instances that have not been correctly classified so far. (Witten and Frank, 2005)

In our experiments, we use the ADTree and LogitBoost implementations provided by the Weka toolkit (Witten and Frank, 2005).

Evaluation metrics. As our classification is intended to be used in ontology learning tasks, we evaluate the performance of the classifiers in separating property-denoting vs. relational adjectives in terms of precision and recall. Depending on whether attribute or relation learning is in focus, it is primarily important to achieve high performance for the respective target category of adjectives rather than good overall accuracy for both classes. However, in case this classification might be of interest for tasks different from ontology learning as well, we also report accuracy scores.

Feature combinations. We report the classification performance on both data sets, based on different feature combinations: In `all-feat`, all features are used individually, while in `all-grp` we collapsed them into groups (see Table 5.6). As a morphological lexicon might not be available in all domains and languages, we also experimented with a feature combination `no-morph` that incorporates all the collapsed features from `all-grp` except for the morphological derivation feature from group `VI`.

Baselines. We compare all these feature combinations against (i) a *majority* baseline that assumes that all adjective types are classified as belonging to the class that accounts for the majority of types in the data and (ii) a rule-based *morph-only* baseline that relies on the *morph* feature only: If an adjective is derived from a noun, it is classified as relational, otherwise as property-denoting. The performance of this decision rule allows to assess the added value that results from a classification approach capitalizing on multiple corpus-derived features in comparison to a simple rule-based approach that merely relies on an existing lexical resource.

Statistical significance. All results reported in the following are statistically significant ($p < 0.05$) relative to the baselines, according to McNemar's test (McNemar, 1947).

		ATTR			REL			
	Learner	P	R	F_1	P	R	F_1	Acc.
all-feat	ADTree	0.95	0.98	0.96	0.71	0.56	0.63	0.93
	Boosted	**0.96**	**0.99**	**0.97**	0.79	0.61	0.69	**0.95**
all-grp	ADTree	0.95	0.98	0.96	0.71	0.56	0.63	0.93
	Boosted	**0.96**	**0.99**	**0.97**	**0.85**	0.61	**0.71**	**0.95**
no-morph	ADTree	**0.96**	0.96	0.96	0.67	**0.67**	**0.67**	0.93
	Boosted	0.95	0.96	0.95	0.56	0.50	0.53	0.91
morph-only	ADTree	**0.96**	0.78	0.86	0.25	**0.67**	0.36	0.77
	Boosted	**0.96**	0.78	0.86	0.25	**0.67**	0.36	0.77
majority		*0.90*	*1.00*	*0.95*	*0.00*	*0.00*	*0.00*	*0.90*

Table 5.8: Class-based precision and recall scores on DS1 (cross-validation)

Cross Validation Results on Annotated Corpus (DS1)

We ran a first experiment on the heuristically annotated data set, using 10-fold cross validation. As the data in DS1 is highly skewed towards the property-denoting class, we also created a balanced data set by random oversampling (Batista et al., 2004).

Precision and recall figures for both classes of adjectives as achieved by the ADTree and the Boosted Learner, respectively, are summarized in Table 5.8. We observe very high precision values for the ATTR class, while precision for REL adjectives is lower. The decision tree performs surprisingly well on the unbalanced set with the `no-morph` feature combination. Interestingly, this holds for both classes, even though the morphological feature is the only positive feature we provided for the REL class. This suggests that morphological derivation as provided by CELEX2 does not perfectly discriminate the two classes.

In Table 5.9 on the next page, we show that even higher precision values, well above the baseline, can be obtained for both classes when an equal number of training instances is provided by random oversampling (Batista et al., 2004). This indicates that a corpus-based classification approach can be applied equally well for attribute and relation learning. Moreover, as revealed by the performance of the *morph-only* baseline in Table 5.9 on the following page, corpus-based learning is clearly superior to a simple lexicon lookup procedure that relies on morphological derivation as the only source of information.

Comparing the decision tree and the boosted learner, we observe slight improvements for the ATTR class, but – more importantly – a considerable increase on the REL class when the `all-grp` combination is used with boosting. Apparently, this classifier benefits from collapsing individual features into groups, thus merging the values of sparse features. For this classifier, at least, the morphological feature provides valuable

		ATTR			REL			
	Learner	P	R	F_1	P	R	F_1	Acc.
all-feat	ADTree	1.00	**0.93**	**0.97**	**0.94**	**1.00**	**0.97**	**0.97**
	Boosted	1.00	0.91	0.95	0.92	1.00	0.96	0.95
all-grp	ADTree	1.00	**0.93**	**0.97**	**0.94**	**1.00**	**0.97**	**0.97**
	Boosted	1.00	0.91	0.95	0.92	1.00	0.96	0.95
no-morph	ADTree	1.00	0.92	0.95	0.93	0.99	0.96	0.95
	Boosted	1.00	0.92	0.96	0.92	1.00	0.96	0.96
morph-only	ADTree	0.96	0.78	0.86	0.25	0.67	0.36	0.77
	Boosted	0.73	0.78	0.75	0.76	0.71	0.73	0.74
Baseline		*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*	*0.50*

Table 5.9: Class-based precision and recall scores on DS1 (cross-validation, random oversampling)

information[11], while the decision tree performs surprisingly well on the unbalanced set when this feature is omitted. Interestingly, this affects both classes, even though morphological derivation is the only positive feature we provided for the REL class. However, for the small set of relational adjectives in DS1, the morphological information is not sufficiently precise, as can be seen from the performance of the *morph-only* baseline in Table 5.8 on the previous page.

In sum, our results indicate that automatically distinguishing property-denoting and relational adjectives at the type level is possible with high accuracy, even on the basis of small training sets.

Results on WordNet Data (DS2)

With 246 property-denoting vs. 140 relational adjective types, the class distribution on DS2 is less skewed in comparison to DS1. Furthermore, DS2 offers sufficient training data for both classes. DS2 was therefore separated into training (80%) and test data (20%). The test set contains 49 property-denoting and 28 relational adjectives.

On DS2, the boosted classifier yields the best results. Detailed figures are displayed in Table 5.10. While all feature combinations outperform both baselines, the `all-grp` combination achieves the best results for both classes in terms of F-score and accuracy. Considering all features without collapsing them into groups yields lower performance in general, except for recall on the ATTR class. Again, *morph-only* constitutes a very strong baseline. Completely omitting the derivation feature leads to a slight decrease in

[11] Note, however, that the boosted learner benefits from morphological information only in combination with other features, as can be seen from the equal performance of both classifiers in the *morph-only* configuration.

	ATTR			REL			
	P	R	F_1	P	R	F_1	Acc.
all-feat	0.85	0.82	0.83	0.70	0.75	0.72	0.79
all-grp	**0.91**	0.80	**0.85**	**0.71**	**0.86**	**0.77**	**0.82**
no-morph	0.87	0.80	0.83	0.69	0.79	0.73	0.79
morph-only	0.80	**0.84**	0.82	0.69	0.64	0.66	0.77
majority	*0.64*	*1.00*	*0.53*	*0.00*	*0.00*	*0.00*	*0.64*

Table 5.10: Class-based precision and recall scores for the Boosted Learner on DS2

performance, while the best results are obtained by combining derivation information with the corpus-based features.

Comparing the performance on DS1 and DS2, we find, above all, that the REL class benefits from the less skewed class distribution in terms of recall. The results on DS2 underline that property-denoting adjectives can be identified with high precision and decent recall. With regard to relational adjectives, we also observe highly satisfactory recall scores, while precision is lower, but still acceptable.

Feature Analysis

In order to determine the features that are particularly valuable for classifying adjectives denoting properties and relations, we retrace the selection procedure of the boosted learner in the `all-grp` configuration on DS2. The results are shown in Table 5.11 on the following page. Rows in the table correspond to iterations of the learner; each row displays the feature group that has been selected in the respective iteration, as it guarantees a maximally informative partition when dividing the instance space at a particular split value (Witten and Frank, 2005). Thus, the order of features being selected in the boosting procedure indicates their relative impact on the performance of the ensemble learner. Positive feature weights (as given in the rightmost column of the table) are in favour of the ATTR class, negative weights in favour of REL.

Being selected in the first iteration, the predicative features (cf. Table 5.6) turn out to be most effective for the classification task as they imply a strong preference for the property-denoting class (feature weight: 1.63) if being frequently observed with an adjective. When being used as the only feature group in the Boosted Learner on DS2, these features yield a classification performance of P=0.87, R=0.65, F_1=0.74 for the ATTR class. We conclude that lexico-syntactic patterns of predicative use are cheap, but reliable indicators to detect property-denoting adjectives in corpora.

Iteration	Base Classifier	Split Value	Weight
1	Group III	≤ 2.5	-0.25
		> 2.5	1.63
2	Group I	≤ 13.5	-0.28
		> 13.5	1.44
3	morph	≤ 0.5	0.67
		> 0.5	-0.79
4	Group II	≤ 2.5	-0.29
		> 2.5	0.90
5	one-proform	≤ 6.5	-0.03
		> 6.5	1.53
6	see-catch-find	≤ 0.5	-0.10
		> 0.5	0.89
7	Group I	≤ 4.5	-0.15
		> 4.5	0.45
8	Group III	≤ 17.5	-0.04
		> 17.5	1.22
9	one-proform	≤ 0.5	0.23
		> 0.5	-0.19
10	one-proform	≤ 7.5	-0.02
		> 7.5	1.76

Table 5.11: Overview of most informative features and their weights for classifying property-denoting and relational adjectives (10 iterations of the boosted learner on DS2); positive weights are in favour of the property class (ATTR), negative weights in favour of the REL class.

Type	**ATTR Tokens**	**REL Tokens**	**IMPOSS Tokens**
beautiful (ATTR)	50	0	0
black (ATTR)	35	7	8
bright (ATTR)	45	1	4
heavy (ATTR)	42	0	8
new (ATTR)	50	0	0
civil (REL)	0	49	1
commercial (REL)	5	44	1
cultural (REL)	2	48	0
environmental (REL)	0	48	2
financial (REL)	0	46	4

Table 5.12: Volatility of prototypical class members

5.2.5 Discussion

As discussed in Section 5.1.5, a type-based classification approach runs the risk of being affected by class shifts on the token level. This is not reflected by the evaluation carried out on the heuristically acquired corpus. In order to investigate the strength of this effect, we selected five adjective types of each class and inspected a random sample of 50 tokens for each type. As example cases, we chose types that were automatically classified with high confidence scores, since, at this point, we were particularly interested in the class change potential of prototypical class members.

The results of this investigation are shown in Table 5.12. The columns labeled with ATTR and REL display counts of tokens that matched one of our target categories, whereas the rightmost column subsumes all tokens that could not be assigned to the ATTR or REL class. The majority of these cases is due to contexts where the adjective is part of a multi-word expression that does not elicit either a property or a relation, e.g. *black hole* or *heavy metal band*.

The average class volatility on the token level amounts to 8.6%. These figures can be considered as rough estimates for the average error that is introduced by raising our classification task to the type level irrespective of potential word sense ambiguities. Still, our findings suggest that class volatility is not an issue that affects entire classes on a large scale, but seems to be limited to individual contexts. This result is corroborated by examining WordNet: Analyzing the distribution of property-denoting and relational readings over the different word senses of adjectives in entire WordNet we found that 13.9% of all types exhibit volatile word senses that cannot be uniformly assigned a property-denoting or a relational reading. Even though this proportion is higher than the one we observed in our corpus (cf. Section 5.1.5), it is still tractable.

5.3 Summary

In this chapter, we investigated the task of automatically separating adjective types with regard to their ontological type in an empirical, corpus-based classification approach. Such a classification is expected to be useful as a filter in attribute selection in order to confine the range of adjectives considered to property-denoting ones.

In a corpus study based on human annotations, we find that only a coarse-grained classification into adjectives denoting properties and relations yields stable results in terms of annotator agreement. Similar to Boleda (2006), we do not find clear supporting evidence for a third class that highlights the fine-grained difference between independent and founded properties[12].

We show that by abstracting from this subtle difference, automatic classification of property-denoting and relational adjectives is feasible at high performance levels. To compensate for sparse and expensive training data on the token level, we generate additional training instances in a heuristic, weakly supervised manner. Our experiments show good and consistent results on two data sets, one of them manually annotated and another one acquired from WordNet. The pattern-based features we use for classification on the type level achieve high performance on the identification of property-denoting adjectives. Feature analysis reveals that patterns encoding predicative use are most effective in order to detect adjectives of this type.

An open issue concerns the feasibility of separating adjectives that are neither property-denoting nor relational (including intensional ones, for instance). Since adjectives of this kind are too sparse in our annotated data and they do not constitute a homogeneous class in WordNet, we could not investigate the problem here. Recent work in this direction (Boleda et al., 2013) underlines the difficulty of discriminating adjectives according to their inferential characteristics.

In summary, we consider the type-based adjective classification proposed here as an attractive method for supporting corpus-based ontology learning. Apart from developping a weakly supervised classification model for the ATTR/REL distinction, our experiments provide useful guidelines for the attribute selection task to be tackled in the remainder of this thesis by having identified the most reliable features for this task, viz. lexico-syntactic patterns encoding predicative use. Given that these patterns can easily be extracted from corpora, they are highly suitable for corpus-based detection of property-denoting adjectives in an unsupervised manner, without the need for an upstream classification process that requires costly annotations and/or the availability of lexical resources. In the following chapter, we demonstrate how these corpus-based indicators of property-denoting adjectives can be incorporated into distributional attribute models.

[12] This distinction might be substantiated in psycholinguistic rather than purely corpus-based settings, given that adjective-noun phrases such as *difficult mountain* have been found to cause human readers higher processing costs in eye-tracking studies than, e.g., *difficult exercise* (Frisson et al., 2011).

6 Attribute Selection from Adjective-Noun Phrases: Models and Parameters

In this chapter, we introduce our approach to attribute selection from adjective-noun phrases capitalizing on structured distributional attribute models. These models are based on (i) attribute-based vector representations for adjectives and nouns in a single distributional space, (ii) vector composition functions in order to construct composed phrase vector representations from individual word vectors and (iii) attribute selection functions in order to select those attribute(s) that are most prominent in the compositional semantics of an adjective-noun phrase from its composed vector representation. These components are introduced in Section 6.1. A pattern-based and a topic-based instantiation of these attribute models are described in Sections 6.2 and 6.3, respectively.[1]

6.1 Foundations of Structured Distributional Models for Attribute Selection

Definition. Following the definitions of general distributional models in Section 2.3, we define *structured distributional attribute models* as a special case of structured distributional models: We assume sets of target words W and attribute nouns A. This enables us to define the structured vector space V_{attr} being spanned by the set of orthonormal basis vectors $\{\vec{e}_a | a \in A\}$. Hence, *structured attribute vectors* representing the meaning of a target word $w \in W$ in V_{attr} are defined as follows:

$$\vec{w}_{attr} = \sum_{a \in A} \omega(w, a) \cdot \vec{e}_a \tag{6.1}$$

6.1.1 Attribute-based Distributional Representations of Adjective and Noun meaning

Contrary to prior work, we model attribute selection as involving *triples*

$$r = \langle noun, attribute, adjective \rangle$$

[1] Parts of the content of this chapter have been previously published in Hartung and Frank (2010b) and Hartung and Frank (2011b).

of nouns, attributes and adjectives. We propose to decompose r into *tuples*

$$r' = \langle noun, attribute \rangle \qquad \text{and} \qquad r'' = \langle attribute, adjective \rangle.$$

Previous learning approaches focussed on r' (Cimiano, 2006) or r'' (Almuhareb, 2006) in isolation only.

(43) a. $\text{blue}_{value}\ \text{car}_{concept}$

b. $\underbrace{\langle concept, \text{ATTR}, value \rangle}_{r} = \underbrace{\langle concept, \text{ATTR} \rangle}_{r'} \circ \underbrace{\langle \text{ATTR}, value \rangle}_{r''}$

c. $\text{ATTR}(concept) = value$

Our approach to attribute selection is illustrated in (43): Starting from adjective-noun phrases as in (43a), consisting of an adjective denoting a property value (e.g., *blue*) and a noun denoting a concept (e.g., *car*), our goal is to induce a triple r, as given in (43b). The triple r explictly relates the concept and the value to an attribute that is evoked in the semantics of the adjective-noun phrase without being made explicit on the phrase level. In the interest of better coverage in a corpus-based distributional model, r is decomposed into tuples r' and r'', as can also be seen in (43b). We assume that the distributional representation of r can be re-constructed from $r' \circ r''$, i.e., by composing the individual representations of r' and r'', using $\circ$ as an appropriate composition function[2]. Translating r into a logical form as given in (43c), the acquired triple can be used to populate a knowledge base or an ontology.

In our corpus-based approach to attribute selection, we model the semantics of adjectives and nouns in an attribute-based distributional model tailored to r' and r''. Thus, adjectives and nouns are represented in semantic vectors defined over pre-defined attributes as dimensions of meaning. Vector components are populated along linguistic patterns capturing meaningful co-occurrences of nouns and attributes (as for r') or adjectives and attributes (as for r''), as will be described in more detail for different instantiations of attribute selection models in Sections 6.2 and 6.3. A fragment of the distributional backbone of a generic attribute model is shown in Fig. 6.1, where the upper part displays examples of attribute-based vector representations for the adjective *enormous* and the noun *ball*, with dimensions of meaning being set to a range of ten attributes[3].

6.1.2 Vector Composition Functions

In order to reconstruct a distributional representation of the triple r from individual vector representations for r' and r'', vector composition is used as a hinge for their

[2] At this point, $\circ$ is deliberately left underspecified; concrete instantiations of composition functions will be discussed below.

[3] For illustration purposes, the vector components in Fig. 6.1 are set to raw corpus-based co-occurrence counts.

	COLOR	DIRECTION	DURATION	SHAPE	SIZE	SMELL	SPEED	TASTE	TEMPERATURE	WEIGHT
$\vec{enormous}$	1	1	0	1	45	0	4	0	0	21
$\vec{ball}$	14	38	2	20	26	0	40	0	0	20
$\vec{enormous} \odot \vec{ball}$	14	38	0	20	1170	0	160	0	0	420
$\vec{enormous} \oplus \vec{ball}$	15	39	2	21	71	0	44	0	0	41

Figure 6.1: Attribute-based vector representations of the adjective *enormous*, the noun *ball* and their compositions into a phrase vector representing *enormous ball*

combination. This serves two purposes: First, the fine granularity of linguistic patterns that capture the triple r comes at the cost of their sparsity when being applied to corpus data. Hence, this "reduce-and-reconstruct" approach can be seen as a strategy to recover from sparsity issues. Second, we argue that this approach is also linguistically sound, assuming that attribute selection is a compositional process rooted in formal semantics. From this point of view, vector composition can be seen as the distributional correlate of a formal semantic process where the adjective selects one or more of the roles provided by the deep semantic structure of the noun (Pustejovsky, 1995).

In our models, following Mitchell and Lapata (2010), we make use of two vector composition functions: *vector multiplication* (denoted as $\odot$ henceforth) and *vector addition* ($\oplus$) as defined in Equations (2.10) and (2.11) on page 31. We expect vector multiplication to perform best in attribute selection as it comes closest to the linguistic function of *intersective* adjectives (Amoia and Gardent, 2007), i.e., to promote dimensions that are prominent both for the adjective and the noun. In case of sparsely populated vector components, it may be reasonable to rely on vector addition, though.[4] In the lower part of Fig. 6.1, both composition functions are illustrated.

Note that both $\odot$ and $\oplus$ belong to the class of *vector mixture models* which rely on two fundamental assumptions (Baroni et al., 2014):

1. Individual word vectors and complex phrase vectors (as the result of the composition process) live in the same semantic space.

2. Syntactic structure in the constituents taking part in the composition does not matter (since vector mixtures are commutative operations).

These assumptions are certainly crucial to many problems in compositional distributional semantics. However, it is a widely held view that particular linguistic phenom-

[4] Mitchell and Lapata (2010) offer a range of other composition functions which are merely variations of vector multiplication and vector addition as defined above. As none of these alternatives seems to capture our intuitions on attribute selection as an intersective compositional process equally well, we decided to restrict ourselves to vector multiplication and vector addition.

ena beyond the word level are most adequately modeled in distributional semantics by individual composition functions reflecting the functional behaviour of these phenomena (cf. Section 3.5). Therefore, we argue that both assumptions are less critical to attribute selection. In fact, the first assumption is deliberately included in the design of our model, as we consider attributes as a layer of meaning that is intersectively shared between adjectives and nouns (cf. Section 4.3). Moreover, capitalizing on attributes as dimensions of meaning, we make adjectives and nouns interoperable in a compositional distributional model, which poses a challenge to other variants of vector mixture models (cf. Thater et al., 2010).

With respect to the second assumption, we anticipate that, in some cases, the syntactic head-modifier relationship in adjective-noun phrases might be more adequately captured by assigning different weights to the adjective and noun representations (cf. Mitchell and Lapata, 2010; Baroni and Zamparelli, 2010) rather than composing them in a fully symmetric manner. As it is an open question which of the two constituents of an adjective-noun phrase should be awarded a higher weight in their composition (cf. Pustejovsky (1995) for an adjective-centered and Asher (2011) for a noun-centered account)[5], we will remain within in the realm of symmetric composition functions in this thesis.[6]

6.1.3 Attribute Selection Functions

In order to select attributes from attribute-based vector representations, we propose four attribute selection functions:

- Most Prominent Component Selection (MPC)
- Threshold Selection (TSel)
- Entropy Selection (ESel)
- Median Selection (MSel)

These selection functions operate in a fully unsupervised manner and rely on free parameters to the smallest possible extent, as will be discussed in detail below. Their main purpose is to separate semantically meaningful dimensions in composed attribute-based vector representations from "noise", i.e., other dimensions of meaning which denote attributes that are not particularly prominent in the compositional semantics of a given adjective-noun phrase.

[5] This aspect and its impact on attribute selection will be investigated in Chapter 8.

[6] Further criticisms against vector mixture models put forward by Baroni et al. (2014) – issues in extracting distributional representations for grammatical and function words, incapability of representing differences in semantic structures, inability to account for recursion in modification contexts – do not apply to attribute selection.

Attribute selection from word and phrase vectors. Each of these selection functions can also be applied to word vector representations. Exploring attribute meanings in individual adjectives and nouns out of context may be interesting in order to assess (i) the ambiguity of adjectives with regard to the attributes they select and (ii) the disambiguation capacity of adjective and noun vectors when being considered jointly. Moreover, these attribute selection functions can be used to (iii) compare word and phrase vector representations in order to determine attributes that are prominent in a noun in isolation, without being selected on the phrase level. These effects can be observed in the phrase *enormous ball* (cf. Fig. 6.1), for example, which offers several dimensions of meaning. The adjective *enormous* may select a set of possible attributes (SIZE or WEIGHT, among others), while the noun *ball* elicits several attributes in accordance with its different word senses[7]. As can be seen from Fig. 6.1, these ambiguities are nicely captured by the separate vector representations for the adjective and the noun (upper part); by composing these representations, the ambiguity is resolved (lower part) and some of the dimensions that are prominent in the word meaning of the noun (e.g., SPEED) are down-weighted in the phrase vector.

Requirements. Depending on whether attributes are to be selected from adjective, noun or phrase vectors, an attribute selection function faces different requirements. This can be seen from Fig. 6.1 as well. This example describes a typical configuration with one vector representing a property-denoting adjective that exhibits relatively strong peaks on one or more dimensions, whereas noun vectors generally show a tendency for broad and flat distributions over their dimensions. This suggests using a selection function that is (i) rather strict (by choosing few very prominent dimensions) in case of adjectives, (ii) less restrictive for nouns (by licensing the inclusion of more dimensions of lower relative prominence), and (iii) largely flexible in case of phrase vectors (by adapting to the compositional processes underlying adjective-noun phrases). In general, we are interested in finding a selection function that relies on as few free parameters as possible in order to avoid efforts for optimizing them by supervised training or to adapt them to different component weighting functions or dimensionalities.

Most Prominent Component Selection (MPC). An obvious method for attribute selection is to choose the most prominent component from any vector (i.e., the highest absolute value). All other components are rejected, irrespective of their relative importance. MPC obviously fails to capture polysemy of targets, which affects adjectives such as *hot*, in particular.

Threshold Selection (TSel). TSel recasts the approach of Almuhareb (2006) in selecting all dimensions as attributes whose components exceed a frequency threshold.

[7] WordNet senses for the noun *ball* include, among others: 1. *round object [...] in games*; 2. *solid projectile*, 3. *object with a spherical shape*, 4. *people [at a] dance*.

This avoids the drawback of MPC, but introduces a parameter that needs to be optimized. Also, it is difficult to apply absolute thresholds to composed vectors, as the range of their components may be subject to great variation, and it is unclear whether the method will scale with increased dimensionality of the semantic space.

Entropy Selection (ESel). In information theory, entropy measures the average uncertainty in a probability distribution (Charniak, 1996). We define the entropy $H(\vec{w_{attr}})$ of an attribute-based vector $\vec{w_{attr}}$ over its components as:

$$H(\vec{w_{attr}}) = -\sum_{a \in A} p(w,a) \log p(w,a), \text{where}$$
$$p(w,a) = \frac{\omega(w,a)}{\sum_{a' \in A} \omega(w,a')}.$$

We use $H(\vec{w_{attr}})$ to assess the impact of singular vector components on the overall entropy of the vector: We expect entropy to detect components that contribute noise, as opposed to those that contribute important information.

We define the algorithm used for entropy-based attribute selection in Figure 6.2 on the facing page. Essentially, the algorithm returns a list of informative dimensions, `selectedAttributes`, by iteratively suppressing combinations of vector components one by one. The possible combinations of components (stored in `ocList`, cf. line 8) are determined by (i) sorting the n vector components in descending order of their value and (ii) collecting from this sorted list all n subsequences consisting of the first element and its successors up to the length of n. Each of these subsequences in `ocList` constitutes a `dimCombination` to be suppressed (cf. lines 11f.) in order to test whether this leads to a gain in vector entropy (cf. line 15). Given that a gain in entropy is equivalent to a loss of information and vice versa, we assume that every combination of components that leads to an increase in entropy when being suppressed is actually responsible for a substantial amount of information. The algorithm includes a back-off to MPC for the special case that a vector contains a singular peak (i.e., $H(\vec{w_{attr}}) = 0$; cf. lines 4–6), so that, in principle, it should be applicable to vectors of any kind. In case of vectors with very flat distributions over their dimensions, entropy selection may result in an empty attribute set, if no combination of components is found to be sufficiently informative to be selected. In the example given in Fig. 6.1, this holds for the word vector $\vec{ball}$ and the composed vector $\vec{enormous} \oplus \vec{ball}$. For the adjective vector $\vec{enormous}$ in the same example, ESel results in the selection of SIZE and WEIGHT, while the attributes SIZE, WEIGHT and SPEED are selected from $\vec{enormous} \odot \vec{ball}$.

Algorithm 1 Entropy Selection Algorithm

```
1:  procedure ENTROPYSELECTION(vector)
2:      entropy ← computeEntropy(vector)
3:      selectedAttributes ← empty list
4:      if entropy == -0.0 then
5:          selectedAttributes ← most prominent component
6:          return selectedAttributes                          ▷ back-off to MPC
7:      else
8:          ocList ← computeOrderedCombinations(dimensions)
9:          for all dimCombination in ocList do
10:             vectorPrime ← vector
11:             for all component in dimCombination do
12:                 vectorPrime ← set component to 0           ▷ suppress all components in dimCombination
13:             end for
14:             entropyPrime ← computeEntropy(vectorPrime)
15:             if entropyPrime > entropy then                 ▷ dimCombination was informative
16:                 selectedAttributes ← dimCombination
17:                 return selectedAttributes
18:             end if
19:         end for
20:         return selectedAttributes
21:     end if
22: end procedure
```

Figure 6.2: Entropy Selection Algorithm

Median Selection (MSel). As a further method we rely on the median m that can be informally defined as the value that separates the upper from the lower half of a probability distribution[8] (Krengel, 2003). It is less restrictive than MPC and TSel and may overcome a possible drawback of ESel. Using this measure, we select all dimensions whose components exceed m. Thus, for the vector representing *ball* in Fig. 6.1, the attributes WEIGHT, DIRECTION, SHAPE, SPEED and SIZE are selected.

As one of the main characteristics of the attribute selection problem, the number of attributes to be predicted for each adjective-noun phrase is unknown in advance. In that respect, attribute selection resembles a *multi-label classification problem*, where each data point is assigned a set of labels of varying cardinality (Tsoumakas and Katakis, 2007). Contrary to common practice in this area, our goal is to design unsupervised distributional models for attribute selection. From the attribute selection functions introduced here, we expect ESel to come closest to an *unsupervised* multi-label selection function, while MPC and MSel represent strong baselines that do not require parameter tuning either.

6.2 Pattern-based Distributional Model

In this section, we define components that are specific to pattern-based attribute models. These include (i) the inventory of lexico-syntactic patterns used in order to acquire initial adjective and noun vectors from corpora and (ii) further parameters used for filtering of extractions or weighting of singular patterns.

6.2.1 Lexico-syntactic Patterns for Attribute Acquisition

We use the following lexico-syntactic patterns[9] for the acquisition of vectors capturing the tuple $r'' = \langle attribute, adjective \rangle$. Even though some of these patterns (`A1` and `A4`) actually match triples of nouns, attributes and adjectives, we only use them for the extraction of binary tuples (underlined), thus abstracting from the modified noun.

(A1) `ATTR of DT? NN is|was JJ`

(A2) `DT? RB? JJ ATTR`

(A3) `DT? JJ or JJ ATTR`

(A4) `DT? NN's ATTR is|was JJ`

(A5) `is|was|are|were JJ in|of ATTR`

[8] The same notion can be straightforwardly applied to semantic vectors, even if their component weights may not always yield a proper propbability distribution.

[9] Some of these patterns are taken from Almuhareb (2006) and Sowa (2000). The descriptions rely on the Penn Treebank Tagset (Marcus et al., 1993). Optional elements in a pattern are marked by `?`.

In order to acquire noun vectors capturing the tuple $r' = \langle noun, attribute \rangle$, we rely on the following patterns. Again, we only extract pairs, as indicated by the underlined elements.

(N1) NN with|without DT? RB? JJ? ATTR

(N2) DT ATTR of DT? RB? JJ? NN

(N3) DT NN's RB? JJ? ATTR

(N4) NN has|had a|an RB? JJ? ATTR

These patterns were partly inspired by Almuhareb (2006) and Sowa (2000). We also experimented with further post-modification patterns, such as:

(*N5) NN for/at/of DT? RB? JJ? ATTR

However, due to the fact that prepositional phrases exhibit numerous ambiguities[10], which is particularly severe for the preposition *of*, we decided to do without them (cf. Poesio and Almuhareb, 2005).

6.2.2 Model Parameters

Some of the adjectives extracted by patterns (A1)–(A5) are not property-denoting and thus represent noise for attribute learning. This affects in particular pattern (A2) which extracts, among others, privative adjectives[11] like *former*, relational ones such as *economic* or *geographic*, or quantifiers like *more*.

This problem may be addressed in different ways: As shown in Chapter 5, it is possible to train a supervised classifier that is capable of automatically separating property-denoting adjectives from relational ones at decent performance levels. In the interest of staying within an *unsupervised* corpus-based framework, however, we opt for alternative approaches based on (i) using lexico-syntactic patterns for adjective target filtering or (ii) eliminating error-prone extractions using intersective pattern filtering.

Target Filtering. During target filtering, adjective extractions are checked against a predicative pattern (P1) :

(P1) DT NN is|was JJ

In our experiments on adjective classification, predicative use of adjectives turned out as the most informative feature for identifying property-denoting adjectives, achieving a precision of 87% (cf. Section 5.2.4)[12]. Pattern-based target filtering of adjectives is implemented such that all extractions of patterns (A1)–(A5) that do not match (P1) are ignored throughout the population of adjective vectors.

[10] See the vast literature on *PP Attachment*, e.g. Hindle and Rooth (1993); Merlo and Ferrer (2006).

[11] This class is characterized by the inference pattern $[AN] \models \neg N$ (Amoia and Gardent, 2007).

[12] This classification performance has been achieved on the dichotomy of property-denoting vs. relational adjectives. We hypothesize that the filtering capacities of (P1) go beyond relational adjectives, i.e., the

$$
\begin{aligned}
Ex(\texttt{Q1}) &= \{\langle \textit{hot}, \textsc{temperature}\rangle, \langle \textit{hot}, \textsc{temperature}\rangle\} \\
Ex(\texttt{Q2}) &= \{\langle \textit{hot}, \textsc{temperature}\rangle, \langle \textit{hot}, \textsc{speed}\rangle\} \\
Ex(\texttt{Q3}) &= \{\langle \textit{hot}, \textsc{taste}\rangle\}
\end{aligned}
$$

$$
\begin{aligned}
pf(\texttt{Q1}, \texttt{Q2}) &= \{\langle \textit{hot}, \textsc{temperature}\rangle, \langle \textit{hot}, \textsc{temperature}\rangle\} \\
pf(\texttt{Q2}, \texttt{Q1}) &= \{\langle \textit{hot}, \textsc{temperature}\rangle\} \\
pf(\texttt{Q1}, \texttt{Q3}) &= \emptyset \\
pf(\{\texttt{Q1}, \texttt{Q2}, \texttt{Q3}\}, \texttt{Q1}) &= \{\langle \textit{hot}, \textsc{temperature}\rangle, \langle \textit{hot}, \textsc{temperature}\rangle\}
\end{aligned}
$$

Figure 6.3: Pattern Filtering Example

Pattern Filtering. Inspired by Pantel and Pennacchiotti (2008), we aim at reducing the impact of noise due to low-confidence patterns. In our models, we experiment with a validation approach that checks the extractions of a particular pattern Q1 by filtering them against the extractions of another pattern Q2, such that only those extractions from Q1 are retained which are also extracted by Q2, whereas all others are discarded.

An example can be seen in Fig. 6.3, where $Ex(\texttt{Q})$ denotes the multiset of pairs extracted by some pattern Q, and $pf(\texttt{Q1}, \texttt{Q2})$ denotes the pattern filtering function for validating the extractions of pattern Q1 against the ones obtained from Q2. As shown in the last example in Fig. 6.3, *pf* can also be used to validate the extractions of several patterns at a time.

6.3 Distributional Attribute Models based on Weakly Supervised Topic Models

Distributional attribute models based on weakly supervised topic models differ from pattern-based models in two respects: First, they make use of dependency paths rather than lexico-syntactic patterns for acquiring word vectors from corpus data. Second, they generalize over observed co-occurrences of adjectives and attributes or nouns and attributes, respectively, by mapping them to abstract *topics* obtained from probabilistic topic models such as Latent Dirichlet Allocation (LDA; Blei et al., 2003). In our work, we adapt LDA in a weakly supervised manner such that the resulting topics are highly attribute-specific and can be injected into a structured distributional attribute model.

pattern may also be beneficial for eliminating privative adjectives and some quantifiers.

1: For each topic:
2: Draw a distribution over words $\vec{\beta}_k \sim Dir_K(\eta)$.
3: For each document:
4: Draw a vector of topic proportions $\vec{\theta}_d \sim Dir_V(\vec{\alpha})$.
5: For each word:
6: Draw a topic assignment $Z_{d,n} \sim Mult(\vec{\theta}_d)$, $Z_{d,n} \in \{1, \ldots, K\}$.
7: Draw a word $W_{d,n} \sim Mult(\vec{\beta_{Z_{d,n}}})$, $W_{d,n} \in \{1, \ldots, V\}$.

Figure 6.4: Generative process underlying LDA (Blei and Lafferty, 2009)

We proceed by giving a brief introduction to probabilistic topic models in Section 6.3.1, focussing on unsupervised (LDA) and a variant of supervised models (Labeled LDA). The details of inducing attribute-specific topics by weakly supervised variants of LDA and how these topics are embedded into a structured distributional attribute model are explained in Section 6.3.2.

6.3.1 Background: Probabilistic Topic Models

Latent Dirichlet Allocation. LDA is a generative probabilistic model for document collections. Each document is represented as a mixture of latent topics, where each topic is a probability distribution over words. These topics can be used as dense features for, e.g., document clustering (Blei et al., 2003; Steyvers and Griffiths, 2007). Depending on the number of topics, which has to be specified in advance, the dimensionality of the document representation can be considerably reduced in comparison to simple bag-of-words models.

The generative process underlying LDA is given in Fig. 6.4, following Blei and Lafferty (2009). Here, K denotes a pre-defined number of topics, V the size of the vocabulary. The vector α of size K is used as a Dirichlet prior on the document-specific topic proportions. The scalar η functions as a symmetric Dirichlet prior on the word-topic distributions.

Considering the documents $\vec{w_{1:D}}$ in the corpus D as observed variables, the posterior distribution of the latent variables given the observed (or pre-specified) ones, as stated in (6.2), can be determined by approximation techniques such as mean field variational inference or Gibbs sampling, among others (Blei and Lafferty, 2009).

$$p(\vec{\theta_{1:D}}, z_{1:D,1:N}, \vec{\beta_{1:K}} | w_{1:D,1:N}, \alpha, \eta) = \frac{p(\vec{\theta_{1:D}}, \vec{z_{1:D}}, \vec{\beta_{1:K}} | \vec{w_{1:D}}, \alpha, \eta)}{\int_{\vec{\beta_{1:K}}} \int_{\vec{\theta_{1:D}}} \sum_{\vec{z}} p(\vec{\theta_{1:D}}, \vec{z_{1:D}}, \vec{\beta_{1:K}} | \vec{w_{1:D}}, \alpha, \eta)} \tag{6.2}$$

As a result of the approximation, estimates of overall *word-topic probabilities* $\widehat{\beta_{k,v}}$, *topic proportions* per document $\widehat{\theta_{d,k}}$ and *word-topic assignments* $\widehat{z_{d,n,k}}$ in each document can be

1: For each topic $k \in \{1, \ldots, K\}$:
2: Generate $\beta_k = (\beta_{k,1}, \ldots, \beta_{k,V})^T \sim Dir(\cdot \mid \eta)$.
3: For each document d:
4: For each topic $k \in \{1, \ldots, K\}$.
5: Generate $\Lambda_k^{(d)} \in \{0, 1\} \sim Bernoulli(\cdot \mid \Phi_k)$.
6: Generate $\alpha^{(d)} = L^{(d)} \times \alpha$.
7: Generate $\theta^{(d)} = (\theta_{l_1}, \ldots, \theta_{l_{M_d}})^T \sim Dir(\cdot \mid \alpha^{(d)})$.
8: For each i in $\{1, \ldots, N_d\}$:
9: Generate $z_i \in \{\lambda_1^{(d)}, \ldots, \lambda_{M_d}^{(d)}\} \sim Mult(\cdot \mid \theta^{(d)})$.
10: Generate $w_i \in \{1, \ldots, V\} \sim Mult(\cdot \mid \beta_{z_i})$.

Figure 6.5: Generative process underlying L-LDA (Ramage et al., 2009)

used in order to decompose a document collection. Compared to exploring a corpus by merely inspecting the bag-of-words profiles of singular documents, these quantities provide a better abstraction over its contents (Blei and Lafferty, 2009).

Labeled LDA. L-LDA (Ramage et al., 2009) extends standard LDA by including supervision for specific target categories. The differences are as follows: (i) The generative process includes a second observed variable, i.e., each document is explicitly labeled with a target category. A document may be labeled with an arbitrary number of categories; unlabeled documents are also possible. However, L-LDA permits only binary assignments of categories to documents; probabilistic weights over categories are not intended. (ii) Contrary to LDA, where the number of topics has to be specified as an external parameter in advance, L-LDA sets this parameter to the number of unique target categories. Moreover, the model is constrained such that documents may be assigned only those topics that correspond to their observable category label(s).

More specifically, L-LDA extends the generative process of LDA by constraining the topic distributions over documents $\theta^{(d)}$ to only those topics that correspond to the document's set of labels $\Lambda^{(d)}$. This is done by projecting the parameter vector of the Dirichlet topic prior α to a lower-dimensional vector $\alpha^{(d)}$ whose topic dimensions correspond to the document labels (Ramage et al., 2009).

This extension is integrated in steps 5 and 6 of the generative process given in Fig. 6.5: First, in step 5, the document's labels $\Lambda^{(d)}$ are generated for each topic k, using a Bernoulli coin toss with a labeling prior Φ_k. The resulting vector of document labels $\lambda^{(d)} = \{k \mid \Lambda_k^{(d)} = 1\}$ is used to define a document-specific label projection matrix $L^{(d)}_{|\lambda^{(d)}| \times K}$, such that $L_{ij}^{(d)} = 1$ if $\lambda_i^{(d)} = j$, and 0 otherwise. This matrix is used in step 6 to project the Dirichlet topic prior α to a lower-dimensional vector $\alpha^{(d)}$, whose topic

dimensions correspond to the document labels. In step 7, a distribution of topics for the corresponding document is generated from this reduced parameter space (Ramage et al., 2009).

Irrespective of these adaptations in the generative process, posterior approximation for L-LDA is carried out analogously to LDA, resulting in word-topic probabilities, topic proportions and topic assignments as well.

6.3.2 Integrating Latent Topics into Distributional Attribute Models

Representing attribute meaning in pseudo-documents. Applying LDA to problems in lexical semantics, where the primary goal is not document modeling but the induction of semantic knowledge from high-dimensional co-occurrence data, requires certain adaptations of the framework in order to tailor the estimated topics to reflect *lexical semantic relations* rather than *document structure*. As shown in previous work on modeling selectional restrictions of verbs by inducing topic distributions that characterize mixtures of topics observed in verb argument positions (Ritter et al., 2010; Ó Séaghdha, 2010), this can be achieved by (i) collecting *pseudo-documents*, i.e., bags of words that co-occur in syntactic argument positions, and (ii) applying LDA to these pseudo-documents.

We apply a similar idea to the attribute selection problem: We compile *attribute-specific* pseudo-documents that characterize attributes by adjectives, nouns and verbs that co-occur with the attribute nouns in carefully selected dependency relations. The topic distributions obtained from fitting an LDA model to the collection of these pseudo-documents can then be injected into attribute-based vector representations for adjectives and nouns.

The list of dependency paths that are used for populating attribute-specific pseudo-documents is shown in Table 6.1 on the next page. The notation of paths follows the scheme $\langle$attribute$\rangle$: $\langle$path$\rangle$: $\langle$target$\rangle$, where = in a path description denotes concatenation of edges, and target words are constrained by their generalized part-of-speech category (`J` for adjective, `N` for noun, `V` for verb)[13]. Individual edges within a path are specified in terms of the dependency labels provided by the Malt parser (Nivre et al., 2007). Edge labels point from the syntactic dependent to the head by default, inverse edges pointing to the dependent are marked by `1` (as in `SBJ1`, for instance).

This method of compiling pseudo-documents is in line with our strategy of reducing the triple r into tuples r' and r'' (cf. Section 6.1). Accordingly, LDA is only exposed to binary tuples obtained from attributes and adjectives or nouns, respectively. As an additional source of distributional information, the dependency paths also collect verbs in particular syntactic environments of attribute nouns. These verbs are used for populating the pseudo-documents and, hence, for inducing attribute-specific topics from them, but they do not occur as target words in the resulting distributional attribute

[13] Generalized categories match all their fully specified sub-categories in the Penn Treebank nomenclature (Marcus et al., 1993). For instance, `J` matches `JJ`, `JJS`, `JJR`, etc.

Dependency Path	Example
N:NMOD1:J	*...[warm]* ***temperature****...*
N:SBJ=OBJ1=PRD1:J	*...the* ***color*** *seems to be [black].*
N:SBJ=PRD1:J	*...* ***texture*** *was [flat]...*
N:SBJ=VC1=PRD1:J	*...the* ***taste*** *has been [fantastic]...*
N:SBJ=VC1=OBJ1:J	*...* ***price*** *was considered [expensive].*
N:SBJ=VC1=VC1=PRD1:J	*...* ***price*** *might have been too [expensive]...*
N:SBJ=VC1=VC1=OBJ1:J	*...the* ***comfort*** *has been considered [excellent]...*
N:SBJ=VC1=VC1=VC1=OBJ1:J	*...the* ***comfort*** *might have been considered [excellent]...*
N:PMOD=NMOD=SBJ=PRD1:J	*...the level of the* ***smell*** *was [horrible].*
N:NMOD1:N	*...[car] 's* ***speed****...*
N:COORD1:N	*...in terms of* ***size*** *and [popularity]...*
N:SBJ=PRD1:N	*...* ***loyalty*** *is the only solid [foundation]...*
N:SBJ=VC1=PRD1:N	*...* ***distance*** *has been the only [impediment]...*
N:SBJ=VC1=OBJ1:N	*...* ***morality*** *is considered the [basis]...*
N:NMOD1=PMOD1:N	*...* ***speed*** *of the [car]...*
N:PMOD=NMOD=SBJ=PRD1:N	*...foundation of* ***success*** *is the [willingness] to...*
N:SBJ:V	***Popularity*** *[requires]...*
N:OBJ:V	*...[requires]* ***fairness****.*
N:SBJ=VC1:V	*The* ***price*** *will [increase]...*
N:SBJ=VC1=VC1:V	*More* ***power*** *would have [helped]...*
N:SBJ=VC1=VC1=VC1:V	*The* ***price*** *may have been [reduced]...*
N:SBJ=VC1=OBJ1:V	*...its* ***color*** *may seem to [change]...*
N:PMOD=ADV:V	*...[approved] of the* ***beauty****...*
N:SBJ=VC1:V	*...its* ***complexity*** *was [criticized]...*
N:SBJ=VC1=VC1:V	*...* ***potential*** *has been [exploited]...*
N:SBJ=OBJ1:V	*His* ***friendliness*** *used to [calm] people down...*

Table 6.1: Dependency paths used to generate attribute-specific pseudo-documents; attribute nouns given in boldface, context words used to populate pseudo-documents by matching the given dependency path in brackets.

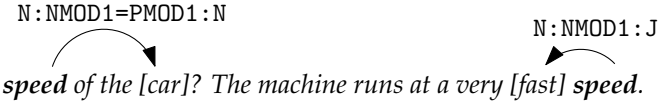


Figure 6.6: Example for populating an attribute-specific pseudo-document by matching dependency paths

model. Figure 6.6 gives an example (using the same nomenclature as in Table 6.1) of populating a pseudo-document for the attribute noun SPEED that contains the noun *car* and the adjective *fast*. With respect to possible word sense ambiguities in attribute nouns, we expect these dependency paths to be sufficiently precise in order to prevent substantial amounts of noise. Assuming that every occurrence of a candidate attribute noun in one of these paths actually denotes an attribute concept[14], we do not apply any additional methods dedicated to word sense disambiguation.

Inducing attribute-specific topics. As introduced in Section 6.3.1, LDA is an unsupervised process that estimates topic distributions θ_d over documents d and topic-word distributions ϕ_k with topics represented as latent variables. Estimating these parameters on a document collection yields document-specific *topic proportions* $p(k|d)$ and *word-topic distributions* $p(w|k)$ that can be used to compute a smooth distribution $p(w|d)$ as in (6.3), where k denotes a latent topic, w a word and d a document in the corpus.

$$p(w|d) = \sum_{k \in K} p(w|k)\, p(k|d) \tag{6.3}$$

In order to link the LDA-inferred topics to attribute meaning and integrate them into a distributional attribute model, we propose *Controled LDA* (C-LDA) as an extension to standard LDA that is capable of implicitly taking supervised attribute information into account. C-LDA will be compared against L-LDA which achieves the same goal by including an additional observable variable into the generative process.

Attribute-specific topics from Controled LDA. The generative story behind C-LDA is equivalent to standard LDA. However, the collection of pseudo-documents used as

[14] This assumption is similar to the *distant supervision* hypothesis commonly applied to relation extraction problems in the absence of labeled data (Mintz et al., 2009). Their extraction approach is supervised by a knowledge base that contains large amounts of pairs instantiating a particular relation type of interest. The distant supervision hypothesis states that "any sentence that contains a pair of entities that participate in a known [...] relation [as contained in the knowledge base] is likely to express that relation in some way" (Mintz et al., 2009). In our setting, the target categories (i.e., attributes) are also previously known from a knowledge resource (i.e., WordNet). In line with Mintz et al. (2009), we consider every occurrence of an attribute noun in one of the syntactic patterns defined in Table 6.1 as an instantiation of an ontological attribute relation. Contrary to their approach, however, our model does not take any previous knowledge about possible instantiations of these relations into account (in terms of pre-defined pairs of attributes and adjectives, for instance).

input to C-LDA is structured in a controled way such that each document conveys semantic information that specifically characterizes the individual categories of interest (attributes, in our case). Thus, the pseudo-documents constructed in this way can be regarded as distributional fingerprints of the meaning of the corresponding attribute.

Presenting LDA with these attribute-specific pseudo-documents can be seen as imposing weak supervision on the estimation process in two respects: The estimated topic proportions $p(k|d)$ will be highly attribute-specific, and similarly so for the word-topic distributions $p(w|k)$. We expect that this makes the model more expressive for the attribute selection task. Moreover, since C-LDA collects pseudo-documents focused on individual target attributes, we are able to link external categories to the generative process by heuristically labeling pseudo-documents with their respective attribute as target category. Thus, we approximate $p(w|a)$, the probability of a word given an attribute, by $p(w|d)$ as obtained from LDA:

$$p(w|a) \approx p(w|d) = \sum_{k \in K} p(w|k)\, p(k|d) \tag{6.4}$$

Finally, setting the component weighting function of the distributional attribute model to this quantity effectively couples probabilistic and distributional modeling for the purpose of attribute selection:

$$\vec{w}_{\text{C-LDA}} = \sum_{a \in A} p(w|a) \cdot \vec{e}_a \tag{6.5}$$

Attribute-specific topics from Labeled LDA. In our instantiation of L-LDA, we collect pseudo-documents for attributes exactly as for C-LDA. Documents are labeled with exactly one category, the attribute noun. This implies that each document is assigned exactly one topic, which renders L-LDA equivalent to a Naive Bayes model (Ramage et al., 2009).

Note that, even though the relationship between documents and topics is fixed, the one between topics and words is not. Any word occurring in more than one document will be assigned a non-zero probability for each corresponding topic. Consequently, under the assumptions of L-LDA, the component weighting function of the distributional attribute model is directly set to the estimate of $p(w|k)$:

$$\vec{w}_{\text{L-LDA}} = \sum_{\substack{k \in L(a), \\ a \in A}} p(w|k) \cdot \vec{e}_a, \tag{6.6}$$

where $L_{A \times K}$ denotes a label projection matrix that is an integral part of L-LDA (cf. line 6 in Fig. 6.5 on page 86) in order to map attribute labels to topics.

Comparison between C-LDA and L-LDA. With regard to attribute modeling, C-LDA and L-LDA build an interesting pair of opposites: L-LDA assumes that attributes are semantically primitive in the sense that they cannot be decomposed into smaller topical

units, whereas words may be associated with several attributes at the same time. C-LDA, at the other end of the spectrum, licenses semantic variability on both the attribute and the word level. Particularly, a word might be associated with some of the topics underlying an attribute, but not with all of them, and an attribute can be characterized by multiple topics. Thus, C-LDA and L-LDA focus on different aspects of corpus-based modeling of lexical meaning, i.e., *smoothing* and *disambiguation*.

Smoothing is generally understood as a strategy to overcome sparsity issues that are frequently encountered in probabilistic models (Chen and Goodman, 1999). In latent variable models such as LDA, this positive smoothing effect is achieved by marginalization over the latent variables (cf. Prescher et al., 2000). In case of C-LDA, for instance, it is unlikely to observe a dependency path linking the adjective *mature* to the attribute MATURITY. Such a relation is more likely for *young*, for example. If *young* co-occurs with *mature* in a different pseudo-document (AGE might be a candidate), this results in a situation where (i) *young* and *mature* share one or more latent topics and (ii) the topic proportions for the attributes MATURITY and AGE will become similar to the extent of common words in their pseudo-documents. Consequently, the final attribute model is assigns a (small) positive probability to the relation between *mature* and MATURITY without observing it in the training data.

On the other hand, the pseudo-documents collected for our extensions of LDA contain a substantial amount of words that are widely spread over several documents as they co-occur with many attribute nouns (e.g., *high*, *great* or *extreme*). Due to the sparse Dirichlet prior on the word-topic distributions, L-LDA effectively enforces a concentration of such adjectives and nouns to fewer attributes in their vector representations. Given that the attribute selection functions implemented in our attribute models reward clearly peaked distributions in word vectors rather than flat, uniform attribute profiles, L-LDA may specifically support the disambiguation capacities of a topic-enriched attribute selection model.

6.4 Summary

In this chapter, we have laid the foundations for learning attribute knowledge from adjective-noun phrases using a corpus-based approach. We have defined structured distributional attribute models for constructing attribute-based distributional representations of adjectives and nouns that can be composed into phrase vectors from which the attribute(s) that are most prominent in the compositional semantics of an adjective-noun phrase can be inferred by means of unsupervised attribute selection functions.

For the acquisition of word-level adjective and noun representations along attributes as dimensions of meaning, two variants of structured distributional models have been introduced: (i) a pattern-based model capitalizing on a small set of lexico-syntactic extraction patterns specifically tailored to capturing attribute meaning in adjectives and nouns, and (ii) a dependency-based model embedding weakly supervised, attribute-

specific topics from Latent Dirichlet Allocation. Incorporating the principle of compositionality and coupling distributional information with the smoothing capacities that can be expected from probabilistic latent variable models, our models are designed so as to balance the conflicting goals of specificity and sparsity in distributional semantic modeling.

In the following chapter, the attribute models introduced here will be subjected to an empirical evaluation on the task of attribute selection from adjective-noun phrases.

7 Attribute Selection: Experimental Evaluation

In this chapter[1], the attribute-based distributional models as introduced in the previous chapter are subjected to an experimental evaluation on the attribute selection task. We systematically compare the attribute selection performance of these models in two scenarios which we denote as *core attribute selection* and *large-scale attribute selection*. They differ in the inventories of attributes being considered: The core attribute selection task is restricted to ten attributes as proposed by Almuhareb (2006), whereas the inventory underyling the large-scale task encompasses more than 260 attributes as represented in WordNet. In the latter setting, our goals are to assess (i) the feasibility of building a large-scale attribute model that is applicable to various tasks in which different attribute inventories may be of interest, and (ii) the capacities of the models when maxing out the dimensionality of the attribute space.

We have constructed two data sets to be used as gold standards in these experiments. We first describe the construction of these data sets and their characteristics in Section 7.1. We present experiments for evaluating the pattern-based and the topic-based attribute selection models in Sections 7.2 and 7.3. Section 7.4 summarizes the results.

7.1 Construction of Labeled Data Sets

Due to the absence of lexical or ontological resources that provide reliable semantic links between attributes, nouns and adjectives, we have created two data sets for evaluating attribute selection in the *core attributes* and the *large-scale* setting. In this section, we describe the procedure of creating these gold standards and their most import characteristics. Both data sets are available to the research community.[2]

7.1.1 Core Attributes Gold Standard

The core attributes data consists of three gold standards for evaluating attribute selection on the core inventory of ten attributes (cf. Appendix A.1) from semantic vectors representing adjectives, nouns or adjective-noun phrases. We first describe the

[1] Parts of this chapter have been previously published in Hartung and Frank (2010b) and Hartung and Frank (2011b).

[2] `http://www.cl.uni-heidelberg.de/~hartung/data/`

approaches for sampling the data points in each of these gold standards, before we present the annotation procedure.

Data Sampling

Adjectives. An appropriate gold standard for attribute selection from adjective vectors can be compiled from WordNet (Fellbaum, 1998). In the interest of comparability, we replicate the procedure of Almuhareb (2006): We collect all adjectives that are linked to at least one of the core attributes by WordNet's `attribute` relation (including *similar*-links). This amounts to 1063 adjectives in total. The resulting data set will is referred to as **CoreAttributes-Adj**.

Nouns. The test set of nouns has been manually annotated with each of the ten core attributes they may elicit. This data set has been created as follows: We started from a representative set of nouns compiled by Almuhareb (2006), comprising 402 nouns that are balanced with regard to their semantic class (according to the WordNet supersenses), ambiguity and frequency. Running the extraction patterns (`N1`)–(`N4`) as introduced in Section 6.2 on the ukWaC corpus (Baroni et al., 2009) yields semantic vectors for 216 of these nouns. From this subset, we randomly sampled 100 nouns which were then manually annotated, resulting in the **CoreAttributes-Nouns** data set.

Adjective-Noun Phrases. For constructing a gold standard for attribute selection from adjective-noun phrases, we started from the same subset of 216 nouns described above. In order to select a set of property-denoting adjectives that are appropriate modifiers of these nouns, we applied the predicative extraction pattern (`P1`) on page 83 to ukWaC. This yielded 2085 adjective types which were further reduced to 386 by frequency filtering ($n \geq 5$). The phrases in the adjective-noun test set were sampled from all pairs in the cartesian product of the 386 adjectives and 216 nouns that occurred at least 5 times in a subsection of ukWaC. We controled for the number of ambiguous adjectives in the data by sampling in two-step procedure: First, we sampled four nouns each for a manual selection of 15 adjectives of all ambiguity levels in WordNet. Ambiguity levels and adjectives selected here are displayed in Table 7.1. This leads to 60 adjective-noun pairs. Second, another 40 pairs were sampled fully automatically. The resulting data set is referred to as **CoreAttributes-Phrases**.

Annotation Procedure

Both the noun and phrase samples as described above were manually labeled by the same three human annotators. Throughout the annotation process, all items (nouns or phrases) were presented to the annotators together with all ten attributes. Their task was to *remove* all attributes from each item that were *not* part of the noun meaning or the compositional semantics of the phrase, respectively, either because the attribute

Ambiguity Level	Num. Adj. Types	Examples
1	761	*green*, ***high***, ***enormous***
2	99	*bitter*, ***red***, *windy*
3	31	***narrow***, *foul*, *massive*
4	16	***blue***, *crisp*, ***great***
5	3	***short***, *little*, ***wide***
6	2	***yellow***, ***flat***
7	3	***warm***, *white*, *broad*
8	4	***heavy***, ***light***, *deep*
9	1	*straight*
10	1	***cold***
11	1	***hot***

Table 7.1: Ambiguity level of adjectives contained in the CoreAttributes-Phrases data (in terms of the number of different attribute senses they are linked to in WordNet). The number of adjective types per ambiguity level and examples for each level are given in the second and third column, respectively. Adjectives manually selected for the ambiguity-controlled sampling step are displayed in boldface.

does not apply to the noun or because it is not selected by the adjective. The annotators were free to accept any number of attributes per item. In case of word sense ambiguities, they were instructed to consider all possible senses of a word and to retain every attribute that was acceptable for at least one sense. Additionally, the annotators were allowed to provide alternative labels if they decided that none of the given attributes was appropriate.

Overall agreement among the three annotators in terms of Fleiss' Kappa (Fleiss, 1971) amounts to $\kappa = 0.69$ for nouns and $\kappa = 0.67$ for phrases. Detailed agreement figures broken down to singular attributes are shown in Table 7.2[3]. As can be seen from this table, the attributes SIZE and DIRECTION caused the annotators most problems on the noun level. Both these attributes were less prone to disagreement in adjective-noun phrases, whereas DURATION is most problematic on the phrase level.

After adjudication by majority voting, CoreAttributes-Nouns contains 424 attributes for 100 nouns. In CoreAttributes-Phrases, 86 attributes were assigned to 76 adjective-noun phrases. 24 phrases could not be assigned any attribute, either because the adjective did not denote a property, as in *private investment*, or the most appropriate attribute for the phrase was not included in the inventory, as in *new house*, for instance.

[3]The attributes SMELL and TASTE were not assigned to any of the phrases to be annotated. Therefore, the agreement score in the respective cells in the table is given as `NaN`.

Attribute	κ_{Nouns}	$\kappa_{Phrases}$
COLOR	0.72	0.81
DIRECTION	0.35	1.00
DURATION	0.51	0.28
SHAPE	0.72	0.59
SIZE	0.20	0.59
SMELL	0.62	NaN
SPEED	0.68	0.76
TASTE	0.69	NaN
TEMPERATURE	0.67	0.80
WEIGHT	0.71	0.48
overall	0.69	0.67

Table 7.2: Inter-annotator agreement in attribute assignment to nouns and phrases in terms of Fleiss' κ

Inspection of Cases of Disagreement

Selective manual inspection reveals the following major *sources of disagreement* among the annotators in the CoreAttributes-Phrases data. We expect that these factors not only cause difficulties to the annotators, but also shed light on particular problems which an automatic attribute selection system has to cope with. Throughout this discussion, we refer to the three annotators as A1, A2 and A3.

Different Interpretations of Attribute Meaning. As one source of disagreement, we observe differences in the interpretation of certain attributes among the annotators. We restrict our analysis on the attributes SHAPE, SIZE and WEIGHT here, as they can be seen to form a coherent cluster of physical properties from an ontological perspective.

A1 and A3 agree with respect to a semantic dependence of these attributes: According to their annotations, WEIGHT is completely subsumed by SIZE and may be subsumed by SHAPE. From the perspective of A2, all three attributes are semantically independent. Much of this difference can be explained by A2 having a wider interpretation of the meaning of SIZE. A2 consistently accepts the SIZE attribute for phrases that involve abstract nouns in combination with degree modifiers, such as *low income*, *high interest* or *wide acceptance*. Albeit plausible in its own right, this interpretation is not in accordance with the physical meaning of SIZE that was intended in the annotation experiment.

Word Sense Ambiguities. The impact of word sense ambiguities can be seen from the example *short corner*, for instance. This example has been annotated with DIRECTION by A3, while the other two annotators did not accept any of the predefined labels. A3's decision to label *short corner* with DIRECTION was based on a less prominent word sense

of *corner* that is related to the sports domain and has been overlooked by the other annotators[4]: In soccer, for instance, *short corner* denotes a corner kick that is not served straight into the box in front of the goal mouth, but passed to another player somewhere near the corner of the field. This interpretation clearly involves a DIRECTION aspect.

Interpretation Differences on Phrase Level. In the example *red deer*, the annotators disagree with respect to the attribute SHAPE. This is due to an ambiguity of *red deer* with regard to a compositional and a non-compositional reading: In the former case, the phrase denotes a deer that is colored red, while in the latter case the phrase refers to *red deer* as the denominator of a species (whose particular silhouette is arguably a protoypical feature that may be subsumed under the attribute SHAPE). As the phrases to be annotated have been sampled regardless of this difference and presented to the annotators without any contextual clues that might be informative for disambiguation, ambiguities of this kind are a potential source of disagreement among our annotators – and certainly are problematic for automatic attribute selection systems as well.

Semantic Associates. A semantic relation between a property-denoting adjective and an attribute may be expressed in different ways: An adjective can either denote one particular value or a set of possible values of the attribute (as discussed in Section 5.1.1) or the adjective and the attribute may be loosely linked by means of *semantic association*[5]. In our attribute selection data, the latter relation becomes manifest in the phrases *beautiful day*, *blue day*, *cloudy day* and the attribute TEMPERATURE.

We argue that in all these cases, the underlying semantic relation between the adjective and the attribute is merely associative as aspects of TEMPERATURE are certainly called to mind by these phrases, while they do not sufficiently determine points or intervals on a temperature scale. Arguably, a *beautiful day* is usually associated with a comfortable range of temperature, which may still vary from summer to winter, for instance.

Indirect Predications. The phrase *cloudy day* as discussed above is also interesting with regard to its relation to the attribute COLOR. A1 is the only annotator assigning this label to this phrase. Her notion of COLOR covers several typical COLOR-denoting adjectives (among them *red*, *blue* and *yellow*). It is certainly fair to say that the adjective *cloudy* does not completely fit with these adjectives when being analyzed out of context.

We argue, however, that the attribution of COLOR to the compositional phrase *cloudy day*, can be justified in consideration of an *indirect predication*. By this term we refer to

[4]This is not surprising given that the WordNet database (Fellbaum, 1998) does not even list this sense of *corner*.

[5]Semantic associates are defined as "those words spontaneously called to mind by a stimulus word", assuming that "these evoked words reflect highly salient linguistic and conceptual features of the stimulus word" (Schulte im Walde et al., 2008).

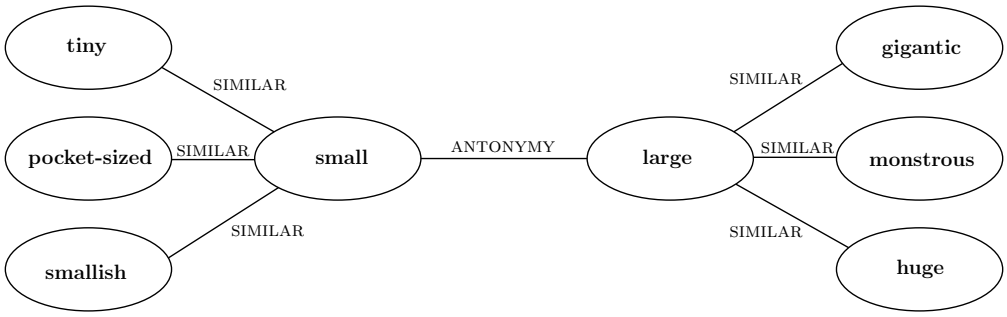


Figure 7.1: Dumbbell structure for representation of adjective meaning in WordNet (taken from Sheinman et al., 2012).

cases where the attribute relation does not hold between the adjective A and the noun N directly, but between A and an entity N' that is semantically related to N instead. Being a special case of *subsective* modification, indirect predications generalize over the categories of event-related and object-related adjectives as discussed in Section 5.1.1. In the case of *cloudy day*, a valid COLOR relation may be established by means of an indirect predication of *cloudy* over $N' = sky$.

7.1.2 Large-scale Gold Standard

The large-scale gold standard has been induced from WordNet. Even though attribute information in WordNet is explicitly encoded only for (some) adjectives and generally missing for nouns, we propose a strategy for acquiring valuable attribute information for adjective-noun phrases based on structural properties of the WordNet resource together with the glosses provided for each word sense. As will be discussed below, this strategy involves a manual validation step which has been carried out in a collaborative annotation setup at Heidelberg and Princeton. Therefore, we refer to the resulting large-scale gold standard as HeiPLAS (**Hei**delberg **P**rinceton **L**arge-scale **A**ttribute **S**election) data set.

Representation of Adjective Meaning in WordNet

Adjectives in WordNet are organized in a so-called *dumbbell structure* (Miller, 1998; Sheinman et al., 2012), as depicted in Fig. 7.1. In this structure, adjective meaning strictly unfolds along a small number of pairs of *anchors* which are explicitly linked by an *antonymy* relation. All other property-denoting adjectives are linked to an anchor by means of a *similar* relation. This leads to a very low degree of interconnectivity in the adjective network. In particular, only anchor adjectives are explicitly linked to an attribute concept. For all other adjectives, their related attributes can only be determined by following a *similar* link to their most closely related anchor. This causes problems due to the very heterogeneous nature of the *similar* links which conflate var-

Anchor	Attribute	Similar
pure	PURITY	sheer
corrupt	CORRUPTNESS	putrid
unoriginal	ORIGINALITY	stale
significant	SIGNIFICANCE	fundamental
fresh	FRESHNESS	hot
hot	TEMPERATURE	sweltering

Table 7.3: Examples of (partial) inconsistencies between anchors and similar adjectives with respect to attribute meaning

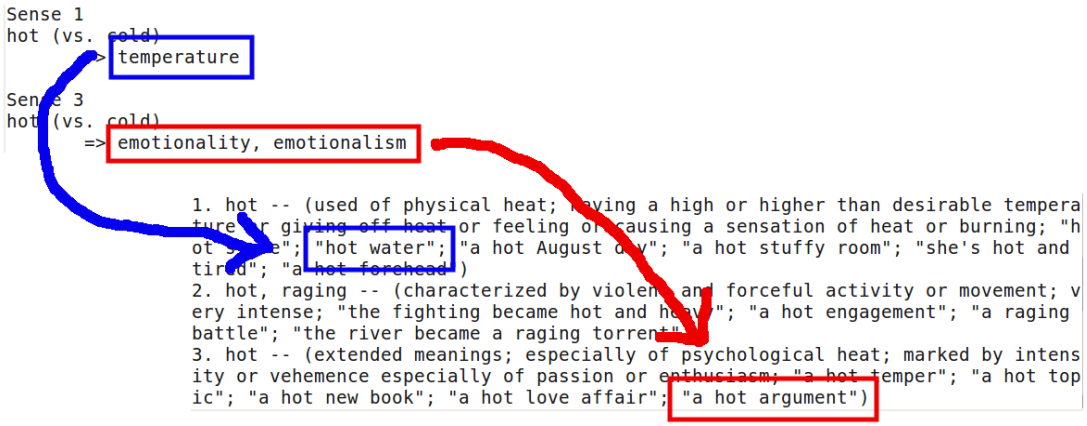


Figure 7.2: Automatic construction of labeled adjective-noun phrases from WordNet

ious degrees of semantic similarity, different classes of selectional preferences, or even different semantic scales (Sheinman et al., 2012, 2013). Table 7.3 shows some examples of anchors and similar adjectives which are not fully consistent with respect to their attribute meaning.

Therefore, our strategy to induce an attribute selection gold standard from WordNet involves two steps: First, all *similar* links in the network are expanded. This yields a maximally comprehensive collection of adjective-attribute pairs which are extended to triples of attributes, adjectives and nouns by mining example phrases in the glosses of these adjectives. Second, triples that are obtained from *similar* links are subjected to a filtering procedure based on an external ontological resource (i.e., the SUMO ontology) and human expert annotations. Details of these two steps are given in the following.

Acquisition of Labeled Adjective-Noun Phrases from WordNet

The acquisition process starts by looking up all adjective senses in WordNet that are linked to an attribute synset, either directly (in case of anchor adjectives) or indirectly

via *similar* links. For each of the resulting pairs, the gloss of the adjective is retrieved and scanned for example phrases containing adjective-noun phrases following the part-of-speech sequences given in (44) and (45)[6]:

(44) `NN* VB* JJ`

(45) `JJ NN !NN`

Given that these examples have been created by the WordNet editors in order to substantiate the meaning of the adjective in its respective attribute sense, we assume that this attribute is also manifest in the compositional semantics of the example phrase. Therefore, we use the attribute link originally provided for the adjective as a label for the complete adjective-noun phrase. This yields an intermediate result of 3755 adjective-noun phrases labeled with 285 unique attributes.

In the example given in Fig. 7.2 on the preceding page, attribute links provided for the pairs ⟨*hot*, TEMPERATURE⟩ and ⟨*hot*, EMOTIONALITY⟩ are propagated to the phrases *hot stove*, *hot water*, *hot forehead* and *hot temper*, *hot topic*, *hot argument*, respectively. Note that, for the sake of clarity, only one of these phrases per attribute is highlighted in the figure and that only direct attribute links are covered in this example. The same procedure is applied analogously for *similar* links as well; due to their semantic heterogeneity (as discussed above), example phrases acquired via *similar* links might introduce semantic drifts to different attribute meanings. Therefore, these cases are subjected to a two-step filtering procedure as described in the following.

Filtering Procedure

Step 1: SUMO Validation. As a first step, the attribute labels of the example phrases obtained via *similar* links are automatically validated against the SUMO ontology (Niles and Pease, 2001; Pease et al., 2002).

WordNet and SUMO offer different perspectives on attributes: While the focus in WordNet is on attributes as abstract linguistic concepts and their relation to singular properties they subsume, attributes in SUMO are defined from a knowledge representation point of view, with an emphasis on their class-constitutive and class-separating function. With regard to the notion of attributes in SUMO, Niles and Pease (2001) state that "the class of `Attributes` includes all qualities, properties, etc., that are not reified as `Objects`. For example, rather than dividing the class of `Animals` under `Objects` into `FemaleAnimals` and `MaleAnimals`, we make `Female` and `Male` instances of `BiologicalAttribute`, which is a subclass of `Attribute`". As a result, the granularity of attributes in SUMO and WordNet differ substantially. For instance, SUMO incorporates a class `SubjectiveAssessmentAttribute` for concepts involving "a criterion which varies from subject to subject and even with respect to the same subject over time" (Niles and Pease,

[6] Notation in these patterns follows the Penn Treebank Tagset. We use the wildcard symbol ∗ to match exactly one arbitrary character. Categories preceded by `!` are negated; i.e., pattern (45) does **not** match phrases including a noun compound as they are not covered by our system.

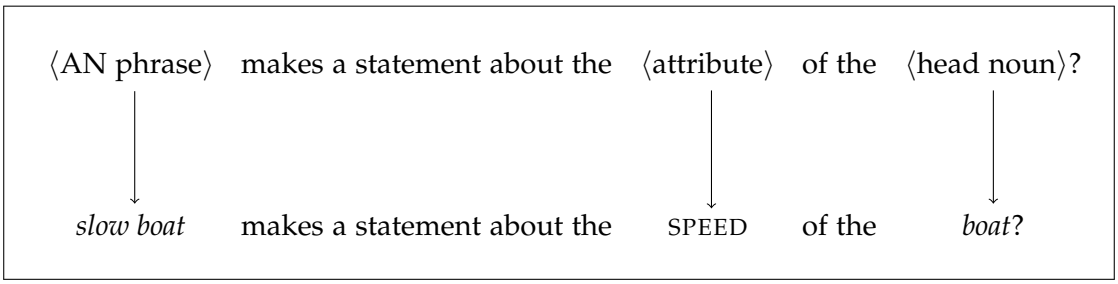


Figure 7.3: Classification test for manually rating example phrases

2003). This subclass conflates a substantial proportion of the attributes in WordNet. On the other hand, SUMO bears one clear advantage over WordNet in that attributes and their corresponding properties are directly linked, which makes SUMO a valuable complementary resource for consistency checks on the heuristically gathered attribute labels.

Our procedure is as follows: For each adjective in a labeled example phrase obtained via *similar* links, we determine its anchor from WordNet, its SUMO attribute and its anchor's SUMO attribute. The SUMO attributes are obtained from the WordNet-SUMO mapping provided by Niles and Pease (2003). The SUMO attributes of the adjective and its anchor are compared with each other: Cases of discordance in these attributes suggest a semantic drift in the underlying *similar* link in WordNet and lead to elimination of the phrase from the gold standard. After this check, 2621 phrases and 274 attributes remain.

Step 2: Manual validation. In a second step, we aim at eliminating remaining difficulties gone unnoticed by this coarse-grained method: e.g., semantic drifts that could not be identified by consulting SUMO[7] or phenomena already encountered in the core attributes gold standard such as indirect predications, semantic associates, cases of non-compositionality, etc. These subtle issues can only be assessed by human annotators. Therefore, we set up an annotation task with 11 participating annotators (all undergraduate students from Princeton University, native speakers of English and with prior experience in lexicographic work). They were asked to decide, for each of the remaining example phrases after SUMO validation, whether its heuristically assigned attribute label correctly reflects an implicit attribute meaning or not. The original annotation guidelines are given in Appendix B.

The task was run in an online environment, enabling annotators to carry out their

[7] Consider the case where SUMO assigns a `SubjectiveAssessmentAttribute` to both an adjective and its anchor. In this situation, the phrase is accepted, while possible drifts *within* the heterogeneous class of `SubjectiveAssessmentAttribute` may go unnoticed.

	Development Set	Test Set
Num. Attributes	254	222
Num. Unique Adj.-Noun Phrases	919	737
Num. Unique Adj. Types	919	737
Num. Unique Noun Types	612	540

Table 7.4: Properties of HeiPLAS gold standard

annotations in an intuitive way, by dragging phrase items over the screen and dropping them into one of two boxes corresponding to the categories *attribute* and *trash* (cf. Fig. B.1 in Appendix B). Annotators were instructed to base their rating on the test in Fig. 7.3, similar to Woods' linguistic test (Guarino, 1992).

The data were split among the annotators on a per-attribute basis such that three annotations were collected for each phrase on average. Attributes were grouped into frequency ranges (high, medium, low frequency according to their occurrence in the ukWaC corpus) and a maximally balanced number of attributes from each range was presented to each annotator. In order to make the annotators familiar with the intended meaning of an attribute and its subsumed properties, they were always presented a definition from WordNet alongside several adjectives and adjective-noun phrases for explication[8] before they were allowed to start annotating the candidate phrases of the respective attribute. Inter-annotator agreement in terms of Fleiss' Kappa (Fleiss, 1971) on all data points for which at least three judgments could be collected amounts to $\kappa = 0.28$. This must be considered a fairly low agreement which underlines the difficulty of the task even for human expert annotators. To increase the reliability and consistency of the annotated data, we decided to retain only those phrases with unanimous agreement between all annotators. The resulting gold standard was randomly split into a development and test section, as given in Table 7.4.[9] Note that, as an artifact of the acquisition procedure and unlike the core attributes gold standard, this data set does not contain any adjective-noun phrases that are explicitly marked as ambiguous by being assigned more than one attribute label.

7.1.3 Summary of Data Sets

The data sets just discussed are summarized in Table 7.5 on page 104, together with their most important characterists and references to the experiments they are used in. The table describes, for each data set, statistics about the number of data points (DPs) contained and their type (adjectives, nouns or phrases), the total number of attributes

[8] Explicative adjectives consisted of anchor adjectives, for which reliable `attribute` links are available. Explicative phrases were extracted from WordNet glosses describing these anchor adjectives.

[9] Differences in numbers of attributes between development and test section are due to 32 attributes for which only one adjective-noun phrase could be retained. In these cases, the respective attribute was decided to become part of the development section.

assigned to these data points and the number of different attributes that assigned to each data point on overage (columns 2–4). If this ratio equals 1 (as in the HeiPLAS data), every data point is annotated with exactly one attribute; otherwise, a subset of data points is assigned more or less than one correct attribute. The fifth column provides an insight into the distribution of data points over attributes in terms of the most frequent and (one of) the most infrequent attribute(s) in the data, respectively. Column 6 briefly summarizes the compilation process for each data set; column 7 points to the descriptions of the experiments these data sets are used in.

Name	Data Points	Num. Attrs.	Avg. Num. Attrs. per Data Point	Distribution of DPs over Attrs.	Data Source	Used in Experiments
CoreAttributes-Adj	1063 Adjectives	10	1.02	COLOR: 381 ⋮ WEIGHT: 16	reconstructed from WordNet following Almuhareb (2006) (cf. Section 7.1.1)	Attribute Selection from Adjective Vectors (Section 7.2.1)
CoreAttributes-Nouns	100 Nouns	10	4.24	SIZE: 93 ⋮ DIRECTION: 8	manual annotation of partially controled sample (sem. class, frequency, ambiguity) (cf. Section 7.1.1)	Attribute Selection from Noun Vectors (Section 7.2.2)
CoreAttributes-Phrases	100 Phrases	8	0.86	SIZE: 32 ⋮ DIRECTION: 1	manual annotation of partially controled sample (frequency, ambiguity) (cf. Section 7.1.1)	Attribute Selection from Phrase Vectors (Section 7.2.3); C-LDA Attr. Selection from Phrase Vectors (Section 7.3.1)
HeiPLAS-Dev	919 Phrases	254	1.00	SIZE: 20 ⋮ WILDNESS: 1	acquired from WN glosses, filtered by SUMO, manually validated (cf. Section 7.1.2)	Large-scale Attr. Selection from Phrase Vectors (Section 7.3.3); Distributional Enrichment (Section 9.4)
HeiPLAS-Test	737 Phrases	222	1.00	SIZE: 19 ⋮ WILDNESS: 1		

Table 7.5: Overview of data sets used for the experiments reported in this thesis. Columns 2–4 show relevant statistics about the data points (DPs) contained in each data set and the attributes assigned to them; column 5 indicates the overall range of the distribution of data points over attributes; columns 6 and 7 summarize the compilation process for each data set and refer to the experiments they are used in.

Pattern Label	Num. Hits (Web)	Num. Hits (ukWaC)
(A1)	2249	815
(A2)	36282	72737
(A3)	3370	1436
(A4)	–	7672
(A5)	–	3768
(N1)	–	682
(N2)	–	5073
(N3)	–	953
(N4)	–	56

Table 7.6: Number of pattern hits on the Web (Almuhareb, 2006) and on ukWaC

7.2 Evaluation of the Pattern-based Attribute Model

We evaluate the attribute selection performance of the pattern-based attribute model in three experiments on the core attributes data: In Experiment 1 and Experiment 2, we evaluate the individual qualitiy of attribute-based word vector representations capturing adjective and noun meaning, respectively. Experiment 3 investigates the selection of hidden attributes from vector representations constructed by composition of adjective and noun vectors. All experiments are evaluated in terms of Precision, Recall and F_1 score. Note that in these experiments it may be required to evaluate m predictions against n labels in the gold standard (due to attribute selection functions returning more or less than one attribute and gold standard phrases being assigned more or less than one attribute, respectively). In these cases, each of the m predictions must be correct in order to achieve $P = 1$ and each of the n labels must be predicted in order to achieve $R = 1$.

7.2.1 Experiment 1: Attribute Selection from Adjective Vectors

The first experiment evaluates the performance of attribute-based vector representations on attribute selection from adjectives. We compare this model against a re-implementation of Almuhareb (2006).

Experimental settings and gold standard. In order to reconstruct Almuhareb's approach, we ran patterns (A1)-(A3)[10] on the ukWaC corpus. Table 7.6 shows the number of hits when applied to the Web (Almuhareb, 2006) vs. ukWaC. (A1) and (A3) yield less extractions on ukWaC as compared to the Web.[11] We introduced two additional patterns, (A4) and (A5), that contribute about 10,000 additional hits. The extractions of all patterns are evaluated individually and in combination.

[10] All patterns referred to in this section are defined in Section 6.2.1 on page 82.
[11] The difference for A2 is an artifact of Almuhareb's extraction methodology.

Pattern(s)	Almuhareb (reconstr.) P	R	F_1	Thr	PattAM (TSel + Target Filter) P	R	F_1	Thr	FPatt	PattAM (ESel + Target Filter) P	R	F_1	FPatt
(A1)	0.183	0.005	0.009	5	0.300	0.004	0.007	5	(A3)	**0.519**	0.035	0.065	(A3)
(A2)	0.207	0.039	0.067	50	**0.300**	0.033	0.059	50	(A1)	0.240	0.049	0.081	(A3)
(A3)	0.382	0.020	0.039	5	**0.403**	0.014	0.028	5	(A1)	0.375	0.027	0.050	(A1)
(A4)					**0.301**	0.020	0.036	10	(A3)	0.272	0.020	0.038	(A1)
(A5)					0.295	0.008	0.016	24	(A3)	**0.315**	0.024	0.045	(A3)
(A1)–(A5)					**0.420**	0.024	0.046	183	(A1)	0.225	0.054	0.087	(A3)

(a) PattAM models based on TSel and ESel with target and pattern filtering being applied

Pattern(s)	PattAM (ESel) P	R	F_1
(A1)	0.231	**0.045**	**0.076**
(A2)	0.084	**0.136**	**0.104**
(A3)	0.192	**0.059**	**0.090**
(A4)	0.135	**0.055**	**0.078**
(A5)	0.105	**0.056**	**0.073**
(A1)–(A5)	0.076	**0.152**	**0.102**

(b) PattAM model based on ESel without target and pattern filtering

Table 7.7: Evaluation results for Experiment 1 (Attribute Selection from Adj. Vectors)

We adopted Almuhareb's manually chosen thresholds for attribute selection for his original patterns (A1)-(A3); for (A4), (A5) and a combination of all these patterns, we manually selected optimal thresholds. With regard to attribute selection functions, we compare TSel (as used by Almuhareb), ESel, MSel and MPC.

In this experiment, we use the CoreAttributes-Adj data set (cf. Section 7.1.1) as gold standard, which facilitates comparability to Almuhareb (2006).

Evaluation results. Results for Experiment 1 are displayed in Table 7.7. Rows indicate the individual or combinations of patterns; the performance of different models is compared across columns. Those columns labeled with PattAM refer to the results of our pattern-based attribute model. PattAM is instantiated using various attribute selection methods and combinations of target and pattern filtering settings, the best-performing ones of which are summarized in Table 7.7a. Regarding pattern filtering, we only report the best filter pattern for each configuration, denoted as FPatt[12] in the table.

The results for our re-implementation of Almuhareb's individual patterns are comparable to his original figures[13], except for (A3) which seems to suffer from quantitative differences of the underlying data. Combining all patterns leads to an improvement in precision over (our reconstruction of) Almuhareb's best individual pattern when TSel

[12] FPatt refers to the second argument of the pattern filtering function *pf* as defined in Section 6.2.2 and exemplified in Fig. 6.3 on page 84. The first argument of *pf* is instantiated with the pattern(s) given in the first column of the respective row in the table.

[13] Precision scores given by Almuhareb (2006) are as follows: P(A1)=0.176, P(A2)=0.218, P(A3)=0.504.

	MPC			ESel			MSel		
Pattern(s)	P	R	F	P	R	F	P	R	F
(N1)	0.22	0.06	0.10	**0.29**	0.04	0.07	0.22	**0.09**	**0.13**
(N2)	**0.29**	0.18	0.23	0.20	0.06	0.09	0.28	**0.39**	**0.33**
(N3)	**0.34**	0.05	0.09	0.20	0.02	0.04	0.25	**0.08**	**0.12**
(N4)	0.25	0.02	0.04	**0.29**	0.02	0.03	0.26	0.02	**0.05**
(N1)–(N4)	**0.29**	0.18	0.22	0.20	0.06	0.09	0.28	**0.43**	**0.34**

Table 7.8: Evaluation results for Experiment 2 (Attribute Selection from Noun Vectors)

is used with target and pattern filtering. MPC and MSel perform worse (not reported here). As for pattern filtering, (A1) and (A3) generally work best.

Both TSel and ESel benefit from the combination with pattern filtering with respect to precision, where the largest improvement (and the best overall result) is observable for ESel on pattern (A1) only. This is the pattern that performs worst in Almuhareb's original setting. From this, we conclude that both an entropy-based attribute selection function and pattern-based strategies for filtering error-prone extractions are valuable extensions to pattern-based attribute models when precision is in focus.

Similar to Almuhareb, recall is problematic. Even though ESel, when being used without any pattern or target filtering (cf. Table 7.7b), leads to slight improvements, the scores are far from satisfying (and at the expense of considerable loss in precision). In line with Almuhareb, we note that this is mainly due to a high number of extremely fine-grained adjectives in WordNet that are rare in corpora.[14]

7.2.2 Experiment 2: Attribute Selection from Noun Vectors

Experiment 2 evaluates the performance of attribute selection from attribute-based noun vectors, using the labeled nouns from the CoreAttributes-Nouns Gold Standard as ground truth.

Evaluation results. Results for Experiment 2 are given in Table 7.8. Performance is lower in comparison to Experiment 1, which suggests that the tuple r'' might not be fully captured by overt linguistic patterns. Note that, in the interest of acquiring distributional representations that reflect the often very broad attribute profile of nouns as much as possible, we do not apply any pattern filtering in this experiment.

Against this background, MPC is relatively precise, but poor in terms of recall. ESel, being designed to select more than one prominent dimension where appropriate, counterintuitively fails to increase recall, suffering from the fact that many noun vectors show a rather flat distribution without any strong peak. MSel turns out to be most suitable for this task: Its precision is comparable to MPC – with (N3) as an outlier –, while

[14] For instance: *bluish-lilac*, *chartreuse* or *pink-lavender* as values of the attribute COLOR.

	MPC			ESel			MSel		
	P	R	F	P	R	F	P	R	F
$\odot$	0.60	0.58	**0.59**	**0.63**	0.46	0.54	0.27	0.72	0.39
$\oplus$	0.47	0.56	0.51	0.42	0.51	0.46	0.18	**0.91**	0.30
BL-Adj	0.44	0.60	0.50	0.51	0.63	0.57	0.23	0.83	0.36
BL-N	0.27	0.35	0.31	0.37	0.29	0.32	0.17	0.73	0.27
BL-P	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Table 7.9: Evaluation results for Experiment 3 (Attribute Selection from Adjective-Noun Phrase Vectors)

recall is about twice as high. Overall, these results indicate that attribute selection for adjectives and nouns, though similar, should be viewed as distinct tasks that require different attribute selection methods.

7.2.3 Experiment 3: Attribute Selection from Phrase Vectors

In this experiment, we compose noun and adjective vectors in order to yield an attribute-based phrase representation, on which attribute selection is performed.

Experimental settings. The quality of the attributes being selected is assessed against the manually labeled adjective-noun phrases from the CoreAttributes-Phrases Gold Standard. Individual vectors for the adjectives and nouns from the test pairs were constructed using all patterns (A1)–(A5) and (N1)–(N4).

For attribute selection, we tested MPC, ESel and MSel. The results are compared against three baselines: BL-P implements a purely pattern-based method, i.e., running the patterns that extract the triple r – i.e., (A1), (A4), (N1), (N3) and (N4), with JJ and NN instantiated accordingly – on the pairs from the test set. BL-N and BL-Adj are back-offs for vector composition, taking the respective noun or adjective vector, as investigated in Experiments 1 and 2, as surrogates for a composed vector.

Evaluation results. Results are given in Table 7.9. Attribute selection based on the composition of adjective and noun vectors yields a considerably higher precision and recall figures than observed in Experiments 1 and 2[15].

Comparing the results of Experiment 3 against the baselines reveals two important aspects of our work. First, the complete failure of BL-P[16] underlines the attractiveness of our method to build structured vector representations from patterns of reduced complexity. Second, vector composition is suitable for selecting hidden attributes from

[15]Note, however, that these figures are not exactly comparable, due to the differences in the underlying data sets (cf. Section 7.1.1).

[16]The patterns used yield no hits for the test pairs at all.

adjective-noun phrases that are jointly encoded by adjective and noun vectors: Both composition methods we tested outperform BL-N.

However, the choice of the composition method matters: $\odot$ performs best with a maximum precision of 0.63. This confirms our expectation that vector multiplication is a good approximation for attribute selection in adjective-noun semantics. Being outperformed by BL-Adj in most categories, $\oplus$ is less suited for this task[17].

All selection methods outperform BL-Adj in precision. Comparing MPC and ESel, ESel achieves better precision when combined with the $\odot$-operator, while doing worse for recall. The robust performance of MPC is not surprising as the test set contains only ten adjective-noun pairs that are still ambiguous with regard to the attributes they elicit. The stronger performance of the entropy-based method with the $\odot$-operator is mainly due to its accuracy on detecting false positives, in that it is able to return "empty" selections. In terms of precision, MSel did worse in general, while recall is decent. This underlines that vector composition generally promotes meaningful components, but MSel is too inaccurate to select them.

Given that attribute selection from individual word vectors has been shown to be a difficult task in Experiments 1 and 2, we consider these very promising results for our approach to attribute selection from structured vector representations. The results also corroborate that previous approaches to attribute learning from adjectives in isolation fall short of the precision that can be achieved in a compositional approach.

7.2.4 Discussion

In this section, we evaluated a pattern-based attribute model as a framework for inferring hidden attributes from the compositional semantics of adjective-noun phrases.

By reconstructing Almuhareb (2006), we showed that attribute-based vector representations of adjective meaning consistently outperform simple pattern-based learning, up to 13 points in precision. A combination of target filtering and pattern weighting turned out to be effective here, by restricting the extractions of lexico-syntactic patterns to particularly reliable ones and filtering adjectives that are not property-denoting.

Our distributional attribute model offers a natural account for resolving sense ambiguity of adjectives and nouns by means of vector composition. Thus, the composition of pattern-based adjective and noun vectors robustly reflects aspects of compositionality in the tradition of formal semantics. Moreover, in a comparison against a purely pattern-based baseline in Experiment 3, we showed that composition of vectors representing complementary meaning aspects is beneficial to overcome sparsity effects in

[17] In the $\text{ESel}_{\oplus}$ setting, a considerable boost in precision can be gained from transforming vector components of the composed phrase vector to the log_{10} scale: $P = 0.71$, $R = 0.35$, $F = 0.47$. This result is an artifact of the smaller range of component values enforced by the logarithm function: Running ESel on vector representations with a broader, less peaked attribute profile produces more results in which no attributes are predicted. Thus, the increase in precision is traded against a substantial drop in recall.

acquiring complex semantic relations.

However, our compositional approach clearly meets its limits if the patterns capturing adjective and noun meaning in isolation are too sparse to acquire sufficiently populated vector components from corpora. Our experimental results from Experiment 2 and the performance of the noun-based baseline from Experiment 3 suggest that this issue is most severe for noun representations. In the following section, we investigate whether an alternative attribute model can alleviate this problem by relying on a dependency-based extraction strategy and incorporating attribute-specific topic modeling.

7.3 Evaluation of Topic-based Attribute Models

In this section, we evaluate the performance of the attribute models based on C-LDA and L-LDA. Our experiments are conducted in two contrastive settings: First, we explore the impact of embedding topic models into attribute-based distributional models by comparing them against a pattern-based attribute model on the CoreAttributes data. Second, we assess the prospects of using distributional attribute models spanning 286 attributes as dimensions for large-scale attribute selection.

Evaluation Settings. The gold standards used in these experiments are the CoreAttributes-Phrases and the HeiPLAS-Dev data sets as introduced above (cf. Table 7.5 on page 104). We report precision, recall and F_1-score. Where appropriate, we test differences in the performance of various model configurations for statistical significance in a randomized permutation test (Yeh, 2000), using the `sigf` tool (Padó, 2006).

Baselines. We compare our models against two baselines, PattAM and DepAM. PattAM refers to the best pattern-based attribute models from the previous experiments in Section 7.2 (cf. Table 7.9 on page 108). We consider ESel and MPC in combination with $\odot$ and $\oplus$, without any pattern oder target filtering being applied. DepAM is similar to PattAM; however, it relies on dependency paths that connect the target elements and attributes in local contexts. The paths are identical to the ones used for constructing pseudo-documents in C-LDA and L-LDA (cf. Table 6.1 on page 88). As in PattAM, the vector components are set to raw frequencies over extracted paths.

LDA Implementations. Our models were implemented using MALLET (McCallum, 2002) for C-LDA and the Stanford Topic Modeling Toolbox[18] for L-LDA. In both cases, 1000 iterations of Gibbs sampling were run, relying on default values for all hyperparameters.

[18] Available from `http://nlp.stanford.edu/software/tmt/`.

Training Data. The pseudo-documents are collected from dependency paths obtained from the parsed pukWaC corpus (Baroni et al., 2009).

7.3.1 Experiment 4: Topic-based Attribute Selection on Core Attributes

In this experiment, we evaluate the performance of C-LDA and L-LDA on the attribute selection task over 10 core attributes. Apart from a comparison to the pattern-based and dependency-based PattAM and DepAM attribute models, we are particularly interested in the relative performance of the LDA models.

Table 7.10 summarizes the results for attribute selection over 10 attributes against the labeled adjective-noun pairs in the CoreAttributes-Phrases set, using ESel and MPC as selection functions on vectors composed by multiplication (Table 7.10a) and addition (Table 7.10b). The results reported for C-LDA correspond to the best performing model (with number of topics empirically set to 42, as this setting yields the best and most constant results over both composition operators).

C-LDA shows highest F_1 scores and recall over all settings, and highest precision with vector addition.[19] We obtain the best overall results in this experiment with vector addition[20] (ESel: P: 0.55, R: 0.66, F: 0.61; MPC: P: 0.59, R: 0.71, F: 0.64). The difference between C-LDA and L-LDA is small but significant for vector multiplication; for vector addition, it is not significant.

Compared to the LDA models, the pattern-based and dependency-based attribute models are competitive[21], but tend to perform lower. This effect is statistically significant for ESel with vector multiplication: Each of the LDA models statistically significantly outperforms one of DepAM and PattAM. With ESel and vector addition, both LDA models outperform both DepAM and PattAM statistically significantly. The $\text{LDA}_{ESel,\oplus}$ models outperform the $\text{PattAM}_{ESel,\oplus}$ model by a high margin in F_1 score: +0.15 for C-LDA; +0.09 for L-LDA. Compared to the stronger multiplicative settings $\text{PattAM}_{ESel,\odot}$ and $\text{PattAM}_{MPC,\odot}$ this still represents a plus of +0.07 (p=0.072) and +0.02 ($p \gg 0.1$) in F_1 score for C-LDA, respectively.

We further observe a clear improvement of the LDA models over PattAM and DepAM in terms of recall (+0.20, $\text{C-LDA}_{ESel,\oplus}$ vs. $\text{PattAM}_{ESel,\odot}$), at the expense of some loss in precision (-0.08, $\text{C-LDA}_{ESel,\oplus}$ vs. $\text{PattAM}_{ESel,\odot}$). This clearly confirms a stronger generalization power of attribute models with embedded topic models compared to purely distributional models.

[19] In Table 7.10, statistical significance of the differences between the models is marked by the superscripts L, D and P, denoting a significant difference over **L**-LDA, **D**epAM and **P**attAM, respectively. All differences reported are significant at $p < 0.05$, except for the difference between C-LDA and L-LDA in Table 7.10a ($p < 0.1$).

[20] In line with Mitchell and Lapata (2010), who also achieved better correlation scores with human judgements from additive rather than multiplicative models in a similarity prediction task on adjective-noun phrases, using a distributional model with LDA-induced topics as dimensions of meaning.

[21] In contrast, recall from Experiment 3 that a purely pattern-based baseline entirely fails on the data set investigated here (cf. the BL-P setting in Table 7.9 on page 108).

	ESel, $\odot$			**MPC**, $\odot$		
	P	R	F	P	R	F
C-LDA	0.58	**0.65**	**0.61**^{L,P}	0.57	**0.64**	**0.60**
L-LDA	**0.68**	0.54	0.60^{D}	0.55	0.61	0.58^{D}
DepAM	0.48	0.58	0.53^{P}	0.57	0.60	0.58
PattAM	0.63	0.46	0.54	**0.60**	0.58	0.59

(a) Vector composition by $\odot$

	ESel, $\oplus$			**MPC**, $\oplus$		
	P	R	F	P	R	F
C-LDA	**0.55**	**0.66**	**0.61**^{D,P}	**0.59**	**0.71**	**0.64**
L-LDA	0.53	0.57	0.55^{D,P}	0.50	0.45	0.47^{D,P}
DepAM	0.38	0.65	0.48^{P}	0.57	0.60	0.58
PattAM	0.42	0.51	0.46	0.47	0.56	0.51

(b) Vector composition by $\oplus$

Table 7.10: Performance of topic-based attribute models (C-LDA and L-LDA) in Experiment 4 (Attribute Selection from CoreAttributes-Phrases)

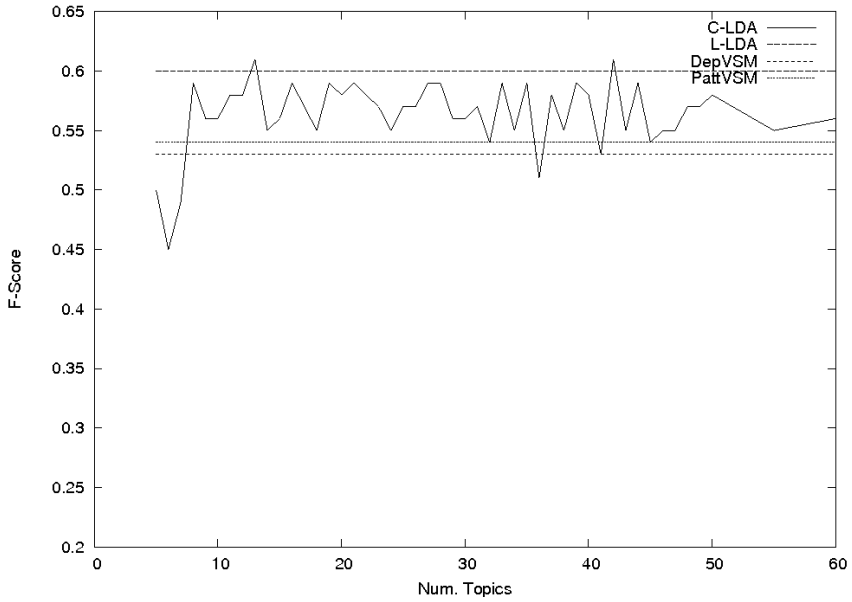


Figure 7.4: Performance of C-LDA$_{ESel,\odot}$ in Experiment 4 for different topic numbers, compared against all other models

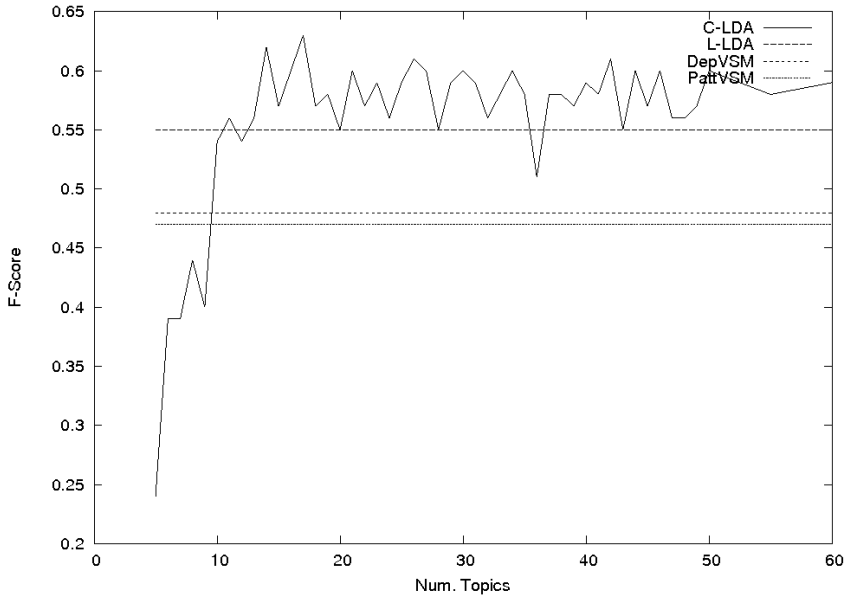


Figure 7.5: Performance of C-LDA$_{ESel,\oplus}$ in Experiment 4 for different topic numbers, compared against all other models

With regard to selection functions, we observe that MPC tends to perform better in DepAM and PattAM, while ESel is more suitable for the LDA models.

Figures 7.4 and 7.5 display the overall performance curve ranging over different topic numbers for C-LDA$_{ESel,\oplus}$ and C-LDA$_{ESel,\odot}$ – compared to the remaining models that are not dependent on topic size. For topic numbers smaller than the attribute set size, C-LDA underperforms, for obvious reasons. Increasing ranges of topic numbers to 60 does not show a linear effect on performance. Parameter settings with performance drops below the baselines are rare, which holds particularly for vector addition at topic ranges larger than 10. With vector addition, C-LDA outperforms L-LDA in almost all configurations, yet at an overall lower performance level of L-LDA (0.55 with addition vs. 0.6 with multiplication). Note that in the multiplicative setting, C-LDA reaches the performance of L-LDA only in its best configurations, while with vector addition it obtains high performance that exceeds L-LDA's top F_1 score of 0.6 for topic ranges between 10 and 20.

Based on these observations, vector addition seems to offer the more robust setting for C-LDA, the model that is less strict with regard to topic-attribute correspondences. Vector multiplication, on the other hand, is more suitable for L-LDA and its stricter association of topics with class labels.

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	**0.39**	**0.31**	**0.35**	**0.37**	**0.27**	**0.32**
L-LDA	0.30	0.18	0.23	0.20	0.18	0.19
DepAM	0.20	0.10	0.13	0.37	0.26	0.30
PattAM	0.00	0.00	0.00	0.00	0.00	0.00

(a) Vector composition by $\odot$

	ESel			MPC		
	P	R	F	P	R	F
C-LDA	**0.43**	**0.33**	**0.38**	**0.44**	**0.28**	**0.34**
L-LDA	0.34	0.16	0.22	0.37	0.18	0.24
DepAM	0.16	0.17	0.17	0.36	0.21	0.27
PattAM	0.13	0.04	0.06	0.17	0.25	0.20

(b) Vector composition by $\oplus$

Table 7.11: Smoothing power of attribute models on sparse vectors

7.3.2 Smoothing Power

Our hypothesis was that LDA models should be better suited for dealing with sparse data, compared to purely distributional pattern-based or dependency-based approaches. While this is broadly confirmed in the above results, we conduct a special evaluation focused on those pairs in the core attributes test set that suffer from sparse data. We selected all adjective and noun vectors that did not yield any positive component values in the PattAM model. The 22 adjective-noun pairs in the evaluation set affected by these "zero vectors" were evaluated using the remaining models.

The results in Tables 7.11a and 7.11b yield a very clear picture: C-LDA obtains highest precision, recall and F_1 score across all settings, followed by L-LDA and DepVSM$_{ESel}$, while their ranks are reversed when using MPC. Again, MPC works better for the purely distributional models (DepAM and PattAM), ESel for the LDA models. Vector addition performs best for C-LDA with F_1 scores of 0.38 and 0.34 – outperforming the pattern-based results on sparse vectors by orders of magnitude.

The results also show that the LDA models clearly benefit from the more general and flexible method of acquiring distributional information on attribute nouns from dependency paths rather than from lexico-syntactic patterns. On top of that, C-LDA and L-LDA$_{ESel}$ contribute additional capacities in order to alleviate sparsity in vector representations: Given that C-LDA and L-LDA estimate attribute-specific topic distributions in the structured pseudo-documents under different assumptions regarding the correspondence of attributes and topics, the impact of C-LDA is due to directly smoothing insufficiently populated vector components, while the focus of L-LDA is on sharpening the attribute profiles of highly ambiguous target words (cf. Section 6.3.2). Hence, this analysis suggests that smoothing is more important than disambiguation for attribute selection from a confined set of core attributes.

7.3.3 Experiment 5: Large-scale Attribute Selection

This experiment is designed to max out the space of attributes to be modeled, to assess the capacity of both LDA models and the DepAM baseline model in the attribute se-

	$\odot$			$\oplus$		
	P	R	F_1	P	R	F_1
C-LDA	0.08	**0.05**	**0.06**^{L,D}	0.03	0.03	0.03
L-LDA	0.15	0.02	0.03	0.05	0.02	0.03
DepAM	**0.16**	0.02	0.03	0.09	0.03	0.05

Table 7.12: Performance figures of C-LDA$_{ESel}$, L-LDA$_{ESel}$ and DepAM$_{ESel}$ in Experiment 5 on 254 attributes

lection task on a large attribute space.[22] Moreover, we are interested in exploring the prospects of training a model on a general, largely domain- and task-independent inventory of attributes. In contrast to the previous experiment with its confined semantic space of 10 target attributes, this represents a huge undertaking.

Overall performance. Table 7.12 displays the performance of all models on large-scale attribute selection on the HeiPLAS development set[23] which covers a range of 254 attributes. We compare vector addition and multiplication. For C-LDA, the number of topics was empirically set to 400.

Overall performance on the large-scale task in terms of F_1 score is very low for all three models and both composition methods. C-LDA performs significantly better[24] than L-LDA and DepAM in the multiplicative setting, yet at an unsatisfactory level. The relative superiority of C-LDA is due to recall (which underlines the strong smoothing capacities of the model once again), whereas DepAM and L-LDA yield better precision.

Examples. Evidently, raising the attribute selection task from 10 to 254 attributes poses a true challenge to our models, by the sheer size and diversity of the semantic space considered. Table 7.13 gives an insight into the nature of the data and the difficulty of the task, by listing correct and false preditions of C-LDA for a small sample of adjective-noun pairs. Possible explanations for false predictions are manifold, among them near misses (e.g., *weak president*, *short flight*, *rough bark*), or idiomatic expressions (e.g., *faint heart*, *fluid society*).

Performance of individual attributes. To gain a deeper insight into the modeling capacity of the LDA models for this large-scale selection task, Table 7.14 (column **all**)

[22] We did not apply PattAM to this large-scale experiment, as only poor performance can be expected in the first place, due to very few pattern hits for a large number of attributes.

[23] HeiPLAS-Test has been held out until the final evaluation that involves distributional enrichment of attribute models (reported in Section 9.4).

[24] Again, statistically significant differences are marked by superscripts (cf. footnote 19). All differences reported are significant at $\alpha < 0.05$, except for C-LDA$_{\oplus}$ vs. L-LDA$_{\oplus}$ ($\alpha < 0.1$).

	Prediction	Correct
thin layer	THICKNESS	THICKNESS
heavy load	WEIGHT	WEIGHT
shallow water	DEPTH	DEPTH
short holiday	DURATION	DURATION
short hair	LENGTH	LENGTH
weak president	POSITION	POWER
fluid society	REPUTE	CHANGEABLENESS
short flight	DISTANCE	DURATION
rough bark	TEXTURE	EVENNESS
faint heart	CONSTANCY	COWARDICE

Table 7.13: Sample of correct and false predictions of C-LDA$_{ESel,\odot}$ in large-scale attribute selection

presents a partial evaluation of attributes that could be assigned to adjective-noun pairs at an individual performance of $F_1 > 0$ by C-LDA$_{ESel,\odot}$ when being trained on the entire range of 254 attributes. Despite the disappointing overall performance of the LDA models on this large attribute space, it is remarkable that C-LDA is able to induce distinctive topic distributions for 23 attributes which yield an average F_1 score of 0.44. In comparison, L-LDA$_{ESel,\odot}$ yields 11 attributes with an average F_1 score of 0.34.

7.3.4 Re-Training on Confined Subsets of Attributes

In an attempt to improve the attribute selection performance of topic-based attribute models, we re-train them on various subsets of the previously considered 254 attributes. These subsets (denoted as *property attributes*, *measurable attributes* and *selected attributes*) are designed such that they confine the large-scale inventory in a meaningul way in order to reduce the complexity of the task. At the same time, the resulting subsets still exceed the cardinality of the 10 core attributes introduced by Almuhareb (2006) in order not to lose sight of the intended large-scale coverage of the model.

Property attributes. Although the 254 attributes used in the large-scale experiment are rather diverse, including concepts such as HEIGHT, KINDNESS or INDIVIDUALITY, we observe a high proportion of core attributes that are successfully modeled (7 out of 10, cf. column **all** in Table 7.14)[25]. Given that they are categorized into the *property* class in WordNet[26], we presume that the varying performance across attributes might be influenced by their ontological subtype. This hypothesis is validated by re-training our attribute models on the 73 attributes pertaining to the *property* subtype in WordNet[27].

[25] Their averaged performance amounts to P=0.50, R=0.38, F_1=0.43.
[26] WordNet separates attributes into *properties*, *qualities* and *states*, among several others.
[27] Refer to Appendix A.2 for a comprehensive list of these *property* attributes.

	all			**property**			**measurable**		
	P	R	F_1	P	R	F_1	P	R	F_1
QUANTITY	0.80	0.22	0.35	0.33	0.22	0.27	1.00	0.22	0.36
WIDTH	1.00	0.80	0.89	1.00	0.80	0.89	0.75	0.60	0.67
CONSISTENCY	0.75	0.40	0.52	0.50	0.20	0.29	0.25	0.20	0.22
AGE	0.22	0.33	0.27	0.20	0.40	0.27	0.50	0.40	0.44
POSITION	0.08	0.11	0.09	0.18	0.22	0.20	0.33	0.11	0.17
LIGHT	0.17	0.33	0.22	0.07	0.33	0.12	0.13	0.33	0.18
COLOR	0.27	0.33	0.30	0.23	0.43	0.30	0.33	0.43	0.38
COMPLEXION	0.06	0.33	0.11	0.11	0.33	0.17	0.10	0.33	0.15
TEMPERATURE	0.60	0.67	0.63	0.54	0.78	0.64	0.70	0.78	0.74
SIZE	1.00	0.15	0.26	0.50	0.30	0.38	1.00	0.25	0.40
SPEED	0.60	0.33	0.43	0.83	0.56	0.67	1.00	0.44	0.62
TEXTURE	0.15	0.40	0.22	0.13	0.20	0.15	0.22	0.40	0.29
WEIGHT	0.67	0.67	0.67	0.50	0.67	0.57	0.50	0.67	0.57
DISTANCE	0.25	0.29	0.27	0.25	0.29	0.27	0.40	0.29	0.33
DEPTH	0.33	0.25	0.29	1.00	0.50	0.67	0.50	0.50	0.50
DURATION	0.75	0.43	0.55	0.60	0.43	0.50	0.83	0.71	0.77
COMPLEXITY	1.00	0.17	0.29				1.00	0.17	0.29
VOLUME				0.25	0.25	0.25	0.33	0.25	0.29
QUALITY				0.18	0.11	0.14	0.33	0.06	0.10
STRENGTH				0.25	0.17	0.20	1.00	0.17	0.29
SEX				1.00	0.33	0.50	0.50	0.33	0.40
LENGTH				0.33	0.33	0.33	0.14	0.33	0.20
PITCH				0.33	0.50	0.40	0.08	0.50	0.13
CRISIS	0.25	0.33	0.29						
REALITY	0.25	0.14	0.18						
IMPORTANCE	1.00	0.17	0.29						
NORMALITY	0.25	0.20	0.22						
ABSORBENCY	1.00	1.00	1.00						
REGULARITY	1.00	0.33	0.50						
DEGREE				0.33	0.07	0.12			
CONTINUITY				1.00	0.17	0.29			
MODERATION				0.25	0.08	0.12			
SHARPNESS				0.14	0.50	0.22			
STATURE				0.17	0.25	0.20			
POWER				0.20	0.20	0.20			
HEIGHT							0.33	0.40	0.36
THICKNESS							0.33	0.20	0.25
INTELLIGENCE							0.33	0.33	0.33
SIGNIFICANCE							0.25	0.20	0.22
average	0.54	0.36	0.44	0.41	0.34	0.37	0.49	0.36	0.41
avg. overall	0.08	0.05	0.06	0.23	0.19	0.21	0.26	0.21	0.23

Table 7.14: Attribute selection on 254 attributes (column **all**), 73 property attributes (column **property**) and 65 measurable attributes (column **measurable**); performance figures of C-LDA$_{ESel,\odot}$ for best attributes ($F_1 > 0$)

	$\odot$			$\oplus$		
	P	R	F_1	P	R	F_1
C-LDA	0.23	**0.19**	**0.21**^{L,D}	0.15	**0.13**	**0.14**^{D}
L-LDA	**0.26**	0.03	0.06	**0.17**	0.08	0.11
DepAM	0.23	0.04	0.07	0.16	0.06	0.09

Table 7.15: Performance figures of C-LDA$_{ESel}$, L-LDA$_{ESel}$ and DepAM$_{ESel}$ in Experiment 5 on 73 property attributes (after re-training)

The evaluation set was restricted accordingly, resulting in 303 pairs from HeiPLAS-Dev that are assigned a *property* attribute. The number of topics in the re-trained model was empirically set to 125.

The overall performance of the models in this experiment is shown in Table 7.15. With vector multiplication, the best-performing composition function across all models, In comparison to large-scale attribute selection on the entire attribute inventory (cf. Table 7.12), C-LDA shows a considerable benefit of $+0.16$ points in F_1 score, which amounts to an improvement by 320%. This constitutes a statistically significant advantage of C-LDA over both L-LDA and DepAM, for which smaller improvements of $+0.03$ and $+0.04$ points are observed. In this confined setting, the superiority of C-LDA over L-LDA and DepAM in recall (at the expense of lower precision relative to L-LDA, though) is even more accentuated compared to the large-scale results. With vector addition, the performance gains are lower in general. The advantage of C-LDA over L-LDA and DepAM diminishes to $+0.02$ and $+0.05$ points in F_1 score, respectively, which still yields a statistically significant difference between C-LDA and DepAM. Note that the affinity of C-LDA with vector addition and L-LDA with vector multiplication, which has been observed in Experiment 4 (cf. Table 7.10 on page 112), is inverted here.

While these overall results are still far from satisfactory, they clearly indicate that the C-LDA attribute model works effectively for at least a subset of attributes, outperforming both L-LDA and the DepAM baseline. Again, a more detailed analysis is given in Table 7.14 (column **property**), showing the performance of the best individual property attributes ($F_1 > 0$) in the restricted experiment. Average performance in this subset of attributes amounts to $F_1 = 0.38$. As expected, narrowing down the attribute inventory results in a higher number of property attributes with $F_1 > 0$. However, in comparison to the unrestricted setting, only some of the property attributes previously modeled successfully (cf. column **all**) benefit from model training on selective data (e.g., COLOR, SPEED or WEIGHT). Thus, apparently, some of the adjectives associated with non-property attributes in the full set provide some discriminative power that is helpful to distinguish property types.

In a qualitative analysis of the non-property attributes filtered out in this experiment, we find that SUMO (Pease et al., 2002) does not provide differentiating definitions for about 60% of these attributes, linking them to a single *subjective assessment attribute*

	$\odot$			$\oplus$		
	P	R	F_1	P	R	F_1
C-LDA	0.26	0.21	0.23	0.15	0.14	0.14
L-LDA	**0.28**	**0.22**	**0.25**	0.13	0.14	0.14
DepAM	0.26	0.19	0.22	**0.18**	**0.16**	**0.17**

Table 7.16: Performance figures of C-LDA$_{ESel}$, L-LDA$_{ESel}$ and DepAM$_{ESel}$ in Experiment 5 on 65 measurable attributes (after re-training)

instead. This suggests that in many cases the distinctions drawn by WordNet are too subtle even for humans to reproduce.

Measurable attributes. Based on the observation that a substantial proportion of attributes successfully modeled by C-LDA is physically measurable, we construct a semantically coherent subset of *measurable attributes* by manually selecting all attributes in the HeiPLAS-Dev data which satisfy at least one of the following criteria:

1. Are humans equipped with a *sensory organ* that enables them to detect different values of the attribute?
2. Does an objective *unit of measurement* exist in order to distinguish different values of the attribute?
3. Does a *technical device* (such as an artificial sensor) exist in order to distinguish different values of the attribute?
4. Does the attribute denote a *scientific concept* that can be operationalized by verifiable criteria or axiomatically grounded?

This method yields 65 measurable attributes; a comprehensive list is included in Appendix A.3. The results of re-training our models on this subset of attributes are shown in Table 7.16. In this experiment, the number of C-LDA topics has been empirically set to 130.

We find a further overall improvement of all models compared to the property subset (cf. Table 7.15). In contrast to the previous configurations, however, the best performance in this experiment is obtained by L-LDA with vector multiplication. With vector addition, the dependency-based model performs surprisingly well. The differences between the models are not statistically significant, though.

Selected attributes. As seen above, all attribute models benefit from being fitted to semantically confined subsets of attributes. Their overall performance still does not fully meet our expectations, though. Therefore, we re-train our attribute models on an

	$\odot$			$\oplus$		
	P	R	F_1	P	R	F_1
C-LDA	0.41	**0.39**	**0.40**	0.33	0.33	0.33
L-LDA	**0.51**	0.33	**0.40**	0.35	0.33	0.34
DepAM	0.46	0.29	0.35	0.32	0.31	0.32

Table 7.17: Performance figures of C-LDA$_{ESel}$, L-LDA$_{ESel}$ and DepAM$_{ESel}$ in Experiment 5 on 23 selected attributes (after re-training)

inventory of *selected* attributes. This subset comprises 23 attributes which have obtained an individual performance of $F_1 > 0$ in the initial large-scale setting (cf. Table 7.14, column **all**). The number of topics in the C-LDA model has been empirically set to 50. This experiment can be seen as an attempt to maximize the attribute selection performance of unsupervised distributional models while still keeping large-scale coverage in focus.

The results are summarized in Table 7.17. Both LDA models and DepAM benefit considerably from the more confined space of attributes. In this configuration, C-LDA and L-LDA are on a par. Their equal performance of F_1=0.40 reflects different proportions of precision and recall, however: C-LDA balances precision and recall almost harmonically (P=0.41, R=0.39), whereas L-LDA strongly prefers precision (P=0.51, R=0.33). These results are in line with C-LDA's advantage in recall as found in previous experiments. DepAM follows the pattern observed in L-LDA, at an overall lower performance, though. Also in line with findings in previous large-scale experiments, multiplicative vector composition is clearly superior to vector addition across all models.

7.3.5 Discussion

Overall findings. Taken together, the experiments conducted in this section in order to compare different types of attribute models with respect to their attribute selection performance on various inventories of attributes follow three trends:

1. Topic-based attribute models (C-LDA, L-LDA) are clearly superior to purely dependency-based models (DepAM). This finding is stable across all inventories of attributes investigated here (either large-scale or confined according to size or semantically motivated subtype).

2. Large-scale attribute selection is a difficult task which cannot be solved at satisfactory performance levels by the unsupervised attribute models investigated here (neither topic-based nor purely dependency-based ones). However, selection performance improves considerably across all models on more confined subinventories, up to F_1=0.40 for *selected* attributes, for instance. This tendency is corroborated by a separate evaluation of C-LDA and L-LDA on the core attributes

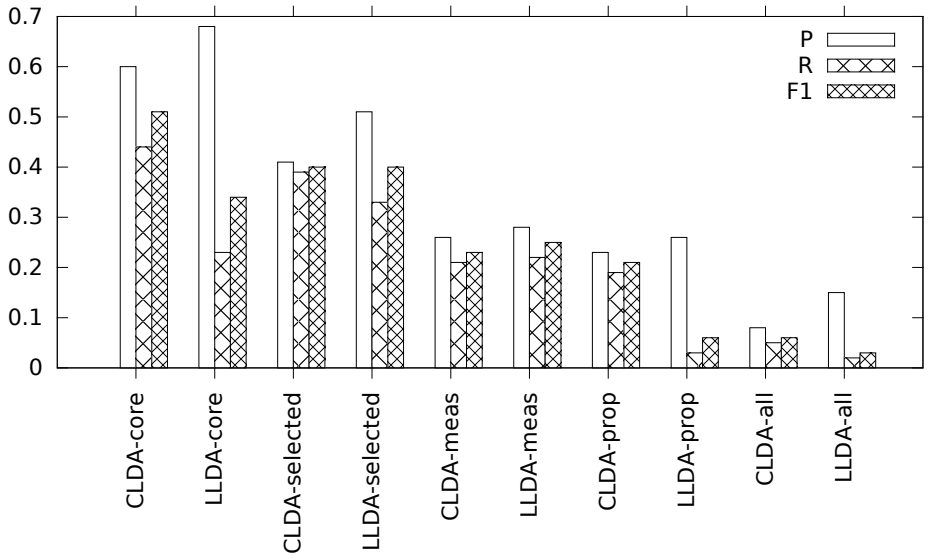


Figure 7.6: Comparison of C-LDA and L-LDA attribute selection performance (ESel, $\odot$) after re-training on different subsets of the HeiPLAS-Dev data

subset within HeiPLAS-Dev: In this evaluation, C-LDA$_\odot$ yields P=0.60, R=0.44, F_1=0.51; L-LDA$_\odot$ achieves P=0.68, R=0.23, F_1=0.34.[28]

3. For the sake of a comparative summary of all experiments conducted in this section, Fig. 7.6 contrasts the performance of C-LDA and L-LDA attribute models when (i) the entire space of all 254 attributes from HeiPLAS-Dev is maxed out, and (ii) when the models are re-trained on smaller subsets. We find that on small ranges of attributes, the difference between C-LDA and L-LDA largely boils down to a trade-off between precision and recall, whereby superior recall turns out as the distinctive feature of C-LDA. In larger spaces of attributes, C-LDA significantly outclasses L-LDA (e.g., by 11 points in F_1 score on the *property* subset), as L-LDA tends to offer better precision in general, but cannot keep up with the growing need for smoothing in these settings.

Smoothing and disambiguation capacities of topic-based attribute models. As discussed before, we argue that the quantitative differences between C-LDA and L-LDA are largely due to their different behavior with respect to smoothing and disambiguation. In order to shed light on these differences, we investigate a selected sample of

[28] Note, however, that the same models yield a performance of F_1=0.61 (C-LDA) and F_1=0.60 (L-LDA) in Experiment 4 on the CoreAttributes-Phrases data set (cf. Table 7.10a on page 112). These differences suggest that the observed difficulties in the large-scale task are not entirely due to shortcomings of our models, but may also reflect adverse conditions in the underlying data.

	DepAM	L-LDA	C-LDA
deep sleep	–	–	DEPTH
deep concentration	COLOR	–	DEPTH
wide margin	WIDTH, SIZE	WIDTH	WIDTH
right side	LIGHT	–	DIRECTION
cool breeze	–	–	TEMPERATURE
short life	DURATION, DISTANCE	DURATION	DURATION, DISTANCE

Table 7.18: Example predictions of DepAM, L-LDA and C-LDA in Experiment 5 (after re-training on *selected* attributes)

instructive example phrases from the HeiPLAS-Dev data for which C-LDA and L-LDA differ in their attribute selections (when being trained on the *selected* attributes subset). The result of this study is shown in Table 7.18, together with DepAM predictions for comparison.

The table clearly shows that L-LDA, compared to C-LDA and DepAM, yields the smallest number of predictions for the example cases under consideration. This confirms the general pattern that underlies L-LDA performance throughout all experiments in this section by consistently favoring high precision over low recall. L-LDA is particularly beneficial in cases of ambiguity which cannot be resolved by a purely dependency-based attribute model, e.g., *wide margin* or *short life*. In both these examples, L-LDA yields the correct disambiguation, whereas only in case of *wide margin*, C-LDA is capable of replicating this result. C-LDA, on the other hand, is clearly superior in cases where DepAM either yields (i) no prediction at all, because the relation between the correct attribute and the adjective and the noun in the phrase cannot be sufficiently substantiated based on overtly observable dependency paths alone, as in *deep sleep* or *cool breeze*, or (ii) a spurious prediction of DepAM must be overridden, as in *deep concentration* or *right side*. In all of these cases, C-LDA returns the correct prediction, whereas L-LDA abstains from selecting any attribute.

Comparison against related work. Recently, Tandon et al. (2014) have proposed the *WebChild* knowledge base which contains triples of attributes, nouns and adjectives that are automatically acquired from adjective-noun phrases (cf. Section 3.2). With respect to individual performance per attribute, the authors encounter a similar effect, i.e., despite being trained on the entire attribute inventory provided by WordNet, their system successfully acquires triples for 19 attributes[29] only.

For this subset, however, the system is highly accurate, achieving an F_1 score of 0.65 ($P = 0.93$, $R = 0.50$). In comparison, their re-implementation of C-LDA performs at F_1=0.27 (P=0.33, R=0.23). Even though these figures are not exactly comparable to

[29]These are displayed in Table A.4 in Appendix A.4. WebChild covers 7 of the 10 core attributes and 10 of the 73 property attributes.

the ones reported in Tables 7.12 and 7.14, as they have been determined on a previous version of the gold standard[30], they clearly point out a performance gap between WebChild and C-LDA. This can be explained by the fact that WebChild capitalizes on *semi-supervised* learning in a knowledge-based environment: While WebChild uses the `attribute` relations between anchor adjectives and attributes as provided by WordNet as seed material for label propagation, C-LDA is required to acquire this information in an unsupervised manner from raw corpus data. On the other hand, this comparison also confirms that attribute selection on the large scale is still challenging even for systems operating at a higher level of supervision.

7.4 Summary

In this chapter, we evaluated two purely distributional (pattern-based and dependency-based) attribute models and two attribute models which integrate attribute-specific topics induced by weakly supervised variants of LDA (C-LDA and L-LDA) on the attribute selection task. These experiments were carried out in two settings contrasting inventories of 10 core attributes and a large-scale set of more than 250 attributes.

On the CoreAttributes dataset, we outperform previous work on attribute selection from adjectives only (Almuhareb, 2006) by wide margins. Extending the task to a linguistically more adequate scenario in which attributes are selected from adjective-noun phrases, our models achieve robust performance, with F_1 scores above 0.60.

Throughout all experiments, the attribute models incorporating attribute-specific topics consistently outperform the purely distributional approaches, though the pattern-based model offers slight margins in precision in singular configurations. We show that the general advantage of C-LDA and L-LDA is due to their specific smoothing and disambiguation capabalities which help to overcome inherent sparsity and ambiguity issues of pattern-based and dependency-based models.

Comparing C-LDA and L-LDA, we find an overall preference for C-LDA which is most clearly visible in the large-scale experiments, whereas on smaller inventories, the differences are not always significant. With respect to an overall assessment of vector composition operators and attribute selection functions, our results remain inconclusive. In the interest of flexibility, we have a preference for entropy selection, which tends to mesh best with vector multiplication. In the large-scale setting, this combination is the only one that yields significant differences between the models, with a clear advantage of C-LDA over both L-LDA and the dependency-based model.

Raising the attribute selection task to the large scale poses a grand challenge to our models. In fact, all models in all configurations investigated obtain very low performance in this experiment, with C-LDA standing out as "the best of a bad bunch". Improved selection performance on more confined attribute subsets across all models lead

[30] Released with Hartung and Frank (2011b).

to the question as to what extent the observed difficulties in large-scale attribute selection are actually due to shortcomings of our models or reflect adverse conditions in the underlying data. Therefore, the next chapter will be devoted to a thorough multivariate linear regression analysis in order to (i) discover performance factors of the C-LDA attribute model that might explain its observed low selection quality and (ii) identify optimization potentials for large-scale attribute selection.

8 Explaining C-LDA Performance in Large-scale Attribute Selection

In this chapter, we thoroughly analyze the performance of C-LDA in large-scale attribute selection as reported in the previous chapter. Our goal is to gain insights into key properties of the model, along with its strenghts and systematic errors, in order to derive strategies for refinement and improvement.

Our line of analysis encompasses three steps. We first assess the validity of two possible explanations for the observed low performance of large-scale attribute selection: It might be due to (i) issues in the approach taken by C-LDA in order to cope with semantic compositionality in adjective-noun phrases (Section 8.2), or (ii) on the level of individual representations of adjective and noun meaning. In order to explore these hypotheses, we subject C-LDA performance on the levels of lexical and phrasal meaning to a linear regression analysis (Section 8.3) that investigates various explanatory variables in order to account for both (i) inherent semantic properties underlying the data and (ii) aspects of our implementation of adjective-noun compositionality in an unsupervised distributional framework. Finally, we describe how the conclusions to be drawn from this analysis can be utilized in order to devise a distributional optimization procedure that aims at improving overall C-LDA performance in large-scale attribute selection (Section 8.4).

Before turning to these aspects in the given order, we first define the explanatory variables guiding the subsequent analyses (Section 8.1).

8.1 Explanatory Variables

The explanatory variables to be investigated can be grouped into six categories: *semantic features*, *morphological features*, *corpus frequency features*, *ambiguity-related features*, features assessing the degree of *uncertainty* in a vector representation, and *vector quality features*. Unless explicitly stated otherwise, these variables have been explored based on the data in HeiPLAS-Dev after filtering *out-of-vocabulary* (OOV) words, i.e., leaving all instances from the gold standard aside for which no reliable C-LDA phrase vector could be constructed due to sparsity issues on the level of word representation. OOV filtering is performed at two levels: At level 1, we exclude all phrases whose phrase representations are composed from a sparse adjective and a sparse noun (i.e., neither the adjective nor the noun are found to co-occur with any attribute noun in the corpus

	Attribute	Adjective	Noun	Phrase
Semantic Features	AbstractnessAttr PropertyAttr MeasurableAttr	AbstractnessAdj AdjMorphAttr	AbstractnessNoun	
Morphological Features		AdjMorphAttr		
Frequency Features	AttrPseudoDocsFreq	AdjFreq AdjPseudoDocsFreq	NounFreq NounPseudoDocsFreq	PhraseFreq
Ambiguity Features	NumAttrSenses AttrSemcorEntropy PropAttrReadings			
Uncertainty Features		AdjEntropy	NounEntropy	PhraseEntropy
Vector Quality Features		AttrRankAdj DeltaAdj	AttrRankNoun DeltaNoun	AttrRankComp

Table 8.1: Explanatory variables investigated in order to explain C-LDA performance

data). At level 2, all phrases are excluded that are composed from either a sparse adjective or a sparse noun. Note that OOV filtering has only been applied in the post-hoc analyses carried out in this section, not in the evaluation of the models in Section 7.3.

8.1.1 Semantic Features

Abstractness. Following Andrews et al. (2009), we hypothesize that attributes are more likely to adhere to concrete rather than abstract words. Given that concrete words "refer to things, events, and properties that we can perceive directly with our senses", while abstract words "refer to ideas and concepts that are distant from immediate perception" (Turney et al., 2011), the position of a concept on a scale ranging from high abstractness to high concreteness might serve as an approximation for the difficulty of assigning an attribute to it.

We extract abstractness scores for adjectives (`AbstractnessAdj`), nouns (`AbstractnessNoun`) and attributes (`AbstractnessAttr`) from a data set presented by Turney et al. (2011). Their data contains 114,501 words (nouns, verbs, adjectives) that are semi-automatically assigned a score ranging from 0 (indicating high concreteness) to 1 (high abstractness).[1] Attributes differ in their abstractness themselves. Consider, for instance, SIZE vs. NICENESS. According to the definitions stated above, SIZE is considerable more concrete than NICENESS, which is also reflected by their abstractness scores (SIZE: 0.53711; NICENESS: 0.75087). If our assumptions are correct, an attribute selection system should face less problems in identifying phrases invoking SIZE rather than NICENESS. Consequently, we extract abstractness scores for attributes as well (`AbstractnessAttr`).

[1] The data set is available from Peter Turney on request.

Property Attributes. As detailed in Sections 7.3.1 and 7.3.3, the overall performance of C-LDA decreases considerably when the underlying attribute inventory is enlarged from 10 core attributes to more than 250 attributes in total. Based on the observation that all attributes in the core inventory share the common feature of being part of the *property* sub-hierarchy in WordNet, we were able to show that reducing the large-scale inventory to a subset of property attributes improves performance by approximately factor 3 (cf. Table 7.14 on page 117).

In order to relate this effect to the impact of other factors, we augmented all attributes in the development set with information about its *property* status in WordNet (`PropertyAttr`): If at least one sense of the attribute noun is a hyponym of *property* in WordNet, the attribute receives `TRUE` as its property label, otherwise `FALSE`. For a comprehensive list of the resulting 73 property attributes, please refer to Appendix A.2.

Measurability. Additional performance gains can be achieved by narrowing down the attribute space to *physically measurable* attributes (cf. Table 7.14 on page 117 again).

In order to determine how measurability interacts with other potential performance factors, we manually judged all attributes in the development set according to the criteria introduced on page 119. If at least one of these criteria is accepted for a particular attribute, it receives `TRUE` as its measurability label, otherwise `FALSE`. This method yields 65 measurable attributes; a comprehensive list is included in Appendix A.3.

8.1.2 Morphological Features

Morphological Relation between Adjectives and Attribute Nouns. If an attribute has an *orientation* with regard to the properties it denotes, one of these properties can be considered its default or *unmarked* value, while all other values on the scale are *marked* deviations from this default. The adjective denoting the unmarked value of such an attribute is usually *morphologically related* to the attribute name, unlike most of its marked counterparts (Miller, 1998). For example, the attribute DEPTH has *deep* as its unmarked, morphologically related default value, while its direct antonym, *shallow*, is regarded as the most prominent deviation from this default.

Recall from Section 6.3 that overtly observable dependency paths are the backbone for populating pseudo-documents in order to model attribute meaning in C-LDA (cf. Table 6.1 on page 88). Due to the morphological relationship between an unmarked adjective and the respective attribute noun, they rarely co-occur explicitly in corpora, i.e., unmarked adjectives tend to be underrepresented in pseudo-documents. The same holds true for antonyms if they are morphologically derived from unmarked adjectives by prefixation.

The left part of Table 8.2 on the following page demonstrates these sparsity effects by way of an example involving the attribute IMPORTANCE. The table displays, for each adjective, its morphological relation to either the attribute name or the unmarked

Adjective	Morphological Relation	Corpus Frequency
important	+	17
unimportant	+	0
crucial	–	186
great	–	999
central	–	140

Phrase	AdjMorphAttr
important rule	TRUE
unimportant feature	TRUE
crucial information	FALSE
great work	FALSE
central cause	FALSE

Table 8.2: Comparison of adjectives related to the attribute IMPORTANCE with regard to their morphological relation to attribute noun or unmarked value and their corpus frequency in context of attribute noun (left); AdjMorphAttr labels for example phrases related to the attribute IMPORTANCE (right)

value, and their corpus frequency in the context of the attribute IMPORTANCE[2]. The upper part of the table contains the adjective denoting the unmarked value, *important*, and the corresponding antonym derived from it by prefixation, *unimportant*. Both are morphologically related to the attribute noun. Relative to the morphologically unmarked marked values in the lower part of the table, the corpus frequencies of *important* and *unimportant* are much lower, and thus do not reflect their prominence for the attribute IMPORTANCE.

In order to analyze the impact of these morphological effects on C-LDA performance, we manually labeled all phrases in HeiPLAS-Dev with regard to the morphological status of the adjective involved: If the adjective is morphologically related to the attribute noun or to another adjective that denotes the unmarked value of the attribute, the phrase is assigned TRUE as its AdjMorphAttr label, otherwise FALSE. See Table 8.2 (right) for illustration. In the following, we will refer to adjectives with AdjMorphAttr=TRUE as *morphologically marked*, all others as *morphologically unmarked* ones.

8.1.3 Ambiguity Features

Like many other words, attribute nouns may also denote various meanings or *word senses*, depending on their respective context (Navigli, 2009). Therefore, if an attribute noun is ambiguous, the respective pseudo-document runs the risk of being populated with context words that are unrelated to this noun's attribute sense, which may introduce noise into the resulting attribute model. Thus, C-LDA may be affected from the problem of word sense conflation that is generally faced by distributional semantic models (Erk, 2012).

The variables introduced in this section are intended to analyze the impact of word sense ambiguities in attribute nouns on attribute selection performance. For the re-

[2]Note that this is equivalent to the frequency of the adjective in the pseudo document used to model the attribute.

mainder of this discussion, we divide the set of word senses pertaining to an attribute noun, $S(n)$, into two subsets

$$S(n) = S_{attr}(n) + S_{non\text{-}attr}(n),$$

where $S_{attr}(n)$ refers to *attribute senses* and $S_{non\text{-}attr}(n)$ to *non-attribute senses*. Relying on the sense inventory provided by WordNet, the former category applies to all senses that are subsumed by the *attribute* concept, the latter to all others. Consider the noun *volume* as an example: Among the six different word senses of *volume* listed in WordNet, we identify three attribute senses expressing attribute meanings defined as "the amount of three-dimensional space occupied by an object", "the property of something that is great in magnitude" or "the magnitude of sound", respectively.[3] The remaining non-attribute senses listed for *volume* have a clear sortal interpretation, referring to *books*, *publications* and *amounts* (in the context of fluids). As becomes evident from this example, it may not always be possible to determine one singular attribute sense for an attribute noun. We investigate the following features in order to assess the ambiguity potential of an attribute noun.

Number of Word Senses per Attribute Noun. For each of the attribute nouns in HeiPLAS-Dev, we investigate the number of its word senses according to WordNet:

$$\texttt{NumAttrSenses} := |S(n)|$$

Attribute Sense Entropy. A high number of different senses for a word is not necessarily problematic for predicting the correct word sense, as many of the senses listed in WordNet are very rarely instantiated in natural language use (Kilgarriff and Rosenzweig, 2000). In fact, many word sense disambiguation (WSD) systems face severe difficulties outperforming the so-called *most frequent sense* baseline (Manandhar et al., 2010), which underlines the importance of sense-specific frequency information.

For the purpose of the present study, this means that merely focussing on the number of word senses per attribute noun might over-estimate the impact of attribute ambiguity in cases where some of the non-attribute senses are very infrequent and thus very unlikely to occur as confounders in the training data presented to C-LDA. As a more robust indicator, we compute a frequency distribution over all word senses of each attribute noun from a sense-labeled corpus and determine the entropy of this distribution:

$$\texttt{AttrSemcorEntropy} := -\sum_{s \in S(n)} p(s) \log p(s), \text{ where}$$

$$p(s) = \frac{\text{freq}(s)}{\sum_{s_i \in S(n)} \text{freq}(s_i)}$$

[3] All quotes in this passage are due to definitions in WordNet 3.0 (Fellbaum, 1998).

The underlying intuition is that the entropy metric indicates, for each attribute noun, the risk of introducing noise into the C-LDA model that is due to sense ambiguities. Attribute nouns with only a few frequent word senses exhibit a lower entropy, while attribute nouns with several frequent senses receive a higher entropy.

For reasons of reliability, we do not perform automated WSD on our training data in order to determine `AttrSemcorEntropy`, but extract word sense frequencies from the manually curated Semcor resource[4], assuming that the frequency patterns underlying the use of attribute nouns will be roughly similar in the samples constituted by Semcor and our data.

Proportion of Attribute/Non-Attribute Senses. We also investigate the impact of confounding senses from different semantic classes, i.e. *attribute* vs. *non-attribute* senses – a distinction that is disregarded by `NumAttrSenses` and `AttrSemcorEntropy`. Relying on word sense frequencies obtained from Semcor, we determine the following factors, accounting for the proportion of attribute senses and non-attribute senses, respectively:

$$\texttt{PropAttrSenses} := \frac{\sum_{s_i \in S_{attr}(n)} \text{freq}(s_i)}{\sum_{s_j \in S(n)} \text{freq}(s_j)}$$

$$\texttt{PropOtherSenses} := 1 - \texttt{PropAttrSenses}$$

Adjective and Noun Ambiguity. In contrast to attribute nouns, we do not compute any ambiguity features for the adjectives and nouns used as distributional descriptors of attribute meaning, due to the fact that adjectives and nouns enter a pseudo-document only to the extent that they explicitly co-occur with the respective attribute noun. We consider this a sufficient disambiguation constraint in order to prevent adjectives and nouns from contributing substantial noise to attribute representations.

Apart from that, the disambiguation capacity provided by existing lexical resources such as WordNet in order to disambiguate adjectives and regular nouns with regard to their attribute senses is rather limited, given that WordNet does not contain any explicit links between noun senses and attributes, and the `similar`-links that need to be traversed in order to decide whether or not an adjective sense denotes a particular attribute meaning are too heterogeneous to be reliable[5] (Sheinman et al., 2013).

[4] SemCor covers a subset of the Brown corpus (Kucera and Francis, 1967) with content words (nouns, verbs, adjectives and adverbs) being manually annotated with part-of-speech, lemma, and word sense information (Miller et al., 1993). In total, SemCor annotations comprise more than 230,000 tokens from 352 texts. While the manual annotations have originally been carried out on the sense inventory from WordNet 1.6, an automatic mapping to WordNet 3.0 is provided by Rada Mihalcea under `http://lit.csci.unt.edu/~rada/downloads/semcor/semcor3.0.tar.gz`.

[5] See also the discussion in Section 7.1.2 of this thesis.

8.1.4 Frequency Features

As attested by Bullinaria and Levy (2007) in a *distance comparison* task[6], words with higher frequency of occurrence have a higher chance a priori to be accurately represented in distributional semantic models, because more training material can be acquired for them. We consider it highly plausible that this *frequency hypothesis* should also hold in the context of related semantic tasks, and in particular also for attribute selection. Consequently, we propose pseudo-document frequency and global frequency as measures to investigate frequency effects on the attribute selection performance of C-LDA.

Pseudo-Document Frequency. Following the frequency hypothesis outlined above, better attribute selection performance can be expected for phrases that invoke an attribute for which a larger amount of training material is available. We measure this quantity in terms of the number of tokens of adjectives, nouns and verbs that are contained in the pseudo document representing the respective attribute (`AttrPseudoDocsFrequency`).

Global Frequency. We measure *global* frequency of adjectives (`AdjFreq`), nouns (`NounFreq`) and adjective-noun phrases (`PhraseFreq`) in unrestricted contexts by querying the ukWaC corpus (Baroni et al., 2009) via the CQP engine (Christ et al., 1999). Queries were formulated such that lemmas and their part-of-speech categories were targeted, without imposing any constraints on the context of occurrence.

8.1.5 Uncertainty Features

Using C-LDA attribute models in a similarity prediction experiment (Hartung and Frank, 2011a), we found that (i) adjective vectors as generated by C-LDA exhibit lower entropy than noun vectors, and (ii) that lower entropy within a vector representation tends to correlate with better performance in similarity prediction. It seems plausible that C-LDA reveals similar patterns also in attribute selection. Therefore, we determine the entropy $H(\vec{w})$ of each vector $\vec{w}$ representing either an adjective (`AdjEntropy`), a noun (`NounEntropy`) or a phrase (`PhraseEntropy`) within the development set as fol-

[6] This task can be seen as a variation of a *pseudo disambiguation task* (Rooth et al., 1999): A distributional semantic model is used to predict, for a collection of 200 target words, the semantic distance of each target to a number of response words that are collected in a supervised way such that one of them is semantically related, while ten others are random confounders. Bullinaria and Levy (2007) evaluate the performance of their model in terms of the proportion of random response words for which the model predicts a larger distance compared to the actually related response word.

lows:

$$H(\vec{w}) = -\sum_{a \in A} p(w,a) \log p(w,a), \text{ where}$$

$$p(w,a) = \frac{\omega(w,a)}{\sum_{a' \in A} \omega(w,a')}$$

Here, $\omega(w,a)$ denotes the value of the vector component that relates the target word w to the attribute a. Normalizing this value to $p(w,a)$ guarantees that $H(\vec{w})$ is equivalent to the standard notion of entropy as originally defined for probability distributions (Shannon, 1948).

The rationale underlying vector entropy is that it distinguishes vectors with a smaller number of relatively pronounced, i.e., *informative*, components from others exhibiting a rather flat, near-uniform distribution. Thus, $H(\vec{w})$ quantifies the amount of *uncertainty* faced by C-LDA when being confronted with the problem of selecting one or more attributes from $\vec{w}$. Intuitively, lower vector entropy corresponds to more accentuated peaks in the distribution over vector components, which reduces the difficulty for an attribute selection system to decide for individual components whether they contribute information or noise. Note that this idea is also utilized by our entropy-based attribute selection method (cf. Section 6.1.3). Thus, evaluating the impact of vector entropy also serves as another benchmark for assessing the appropriateness of using ESel for attribute selection.

If a vector representation generated by C-LDA exhibits a flat distribution, however, there may be two reasons for this: Either the respective target word resists distributional modelling in an attribute space due to inherent semantic properties, or the particular approach taken by C-LDA modeling fails at promoting the most important vector components.

8.1.6 Vector Quality Features

Rank Features. As the most direct way to assess the quality of semantic vectors, we determine the rank of the correct attribute according to the gold standard within the ordered list of all vector components, assuming that in an ideal attribute-based meaning representation, the correct attribute should be ranked at first position. More generally, the lower the rank of the correct attribute in a vector representation (i.e., the closer it is to rank 1), the higher its quality.

We define a function $rank : W \times A \rightarrow \mathbb{N}$ that effectively re-arranges the components of the original vector $\vec{w}$ in decreasing order and assigns integer values to each of them, such that, for all $a_i, a_j \in A$:

$$rank(w,a_i) < rank(w,a_j) \text{ if } \omega(w,a_i) > \omega(w,a_j)$$
$$rank(w,a_i) = rank(w,a_j) \text{ if } \omega(w,a_i) = \omega(w,a_j)$$
$$rank(w,a_i) > rank(w,a_j) \text{ if } \omega(w,a_i) < \omega(w,a_j)$$

Based on this function, the factors AttrRankAdj and AttrRankNoun can be determined in terms of $rank(w, a_{corr})$, where w stands for an individual adjective or noun and a_{corr} denotes the correct attribute according to the gold standard. Note that the gold standard provides correct attributes for phrases only, from which they are propagated to the adjective and noun vectors representing the constituents of the phrase. Thus, the factors AttrRankAdj and AttrRankNoun always reflect the correct attribute(s) *in the phrase context*. Note that, even though the *rank*-function establishes a partial order on the attributes, we do not observe any *ties* (i.e., two or more attributes being assigned the same rank) in the data (except for out-of-vocabulary terms, where all vector components are 0 anyway). In case of several correct attributes provided for a particular vector representation, the *rank*-function picks out the lowest of the corresponding ranks. In order to determine attribute ranks for phrase vectors in an analogous way (AttrRankComp), the domain of the *rank*-function as given above is extended to cover phrases as well.

Compositionality Features. In order to investigate the effect of vector composition, we are interested in comparing the quality of phrase vectors to vectors representing their constituents. Following Boleda et al. (2012), we define two further factors based on the *rank* function described above, DeltaAdj and DeltaNoun. These factors are computed as the difference of the rank of the correct attribute in the individual vector and its rank in the composed vector:

$$\texttt{DeltaAdj} = \texttt{AttrRankAdj} - \texttt{AttrRankComp} \tag{8.1}$$

$$\texttt{DeltaNoun} = \texttt{AttrRankNoun} - \texttt{AttrRankNoun} \tag{8.2}$$

Intuitively, these factors take high (i.e, positive) values if vector composition yields an improvement of the correct attribute in the phrase vector beyond the individual noun vector, low (negative) values otherwise.

8.2 Compositionality in C-LDA

The design of the attribute models proposed in this thesis has been based on the assumption from formal semantics that attribute selection instantiates a compositional process in which the adjective selects particular aspects of meaning provided by the deep lexical semantics of the noun (Pustejovsky, 1995). In our attribute models, this assumption is reflected in the intersective approach to constructing vector representations of adjective-noun phrase meaning from individual adjective and noun vectors by way of multiplicative vector composition, followed by an entropy-based method to select the most informative attributes from the composed phrase vector.

Hypothesis. In this section, we seek to show that the distributional approach to attribute selection developed in this thesis follows formal semantic principles, using the

Phrase	Attribute	DeltaAdj	DeltaNoun
confusing signal	CLARITY	7	57
ineffective legislation	EFFECTIVENESS	8	20
wrong assumption	CORRECTNESS	5	32
perfect reproduction	PERFECTION	20	16
good knife	QUALITY	29	0
meager resource	SUFFICIENCY	28	2

Table 8.3: Examples of compositional gains due to phrase rank improvements over individual adjective or noun vector (as indicated by positive DeltaAdj or DeltaNoun values, respectively)

method of *proof by contradition*. The hypothesis to be falsified is that the low performance of the C-LDA attribute model in large-scale attribute selection is due to weaknesses in the model which prevent it from capturing traits of compositionality in the adjective-noun data it is evaluated on. We argue that in order to falsify this hypothesis, at least two criteria must be met by C-LDA:

1. An attribute model reflecting compositional principles should yield *compositional gains* in the sense that phrase vectors combine individual attribute profiles of adjectives and nouns in such a way that the rank of the correct attribute in the phrase vector is lower than in the word vectors.

2. The predictions of the model should be traceable to *compositional semantic processes*, i.e., linguistically meaningful patterns in the interaction of adjective and noun meaning. In order to assess this criterion, we investigate as to what extent observed attribute ranks in composed phrase vectors being high or low, respectively, can be reduced to meaningful patterns in the attribute ranks of the word vectors contributing to the phrase representation.

Compositional Gains. The first criterion is assessed in terms of the compositional features defined in Equations 8.1 and 8.2 on the previous page: We expect attribute selection from a composed adjective-noun vector to result in positive values of either DeltaAdj or DeltaNoun. In fact, we observe *compositional gains* in terms of a positive delta in either the adjective or the noun vector in more than 91% of the instances in the development set after OOV filtering at level 1 (as described in Section 8.1 on page 125). Hence, in line with what can be expected from a linguistically adequate distributional compositional model, the approach taken by C-LDA is generally capable of contextually improving the quality of individual attribute-based adjective or noun vector representations. A small sample of selected examples of compositional gains is shown in Table 8.3.

	RankComp≤ 10		RankComp> 10	
	RankNoun≤ 10	RankNoun> 10	RankNoun≤ 10	RankNoun> 10
RankAdj≤ 10	ADJ-N-COMP (36; 4.1%)	ADJ-n-COMP (62; 7.1%)	ADJ-N-comp (0)	ADJ-n-comp (32; 3.7%)
RankAdj> 10	adj-N-COMP (37; 4.3%)	adj-n-COMP (29; 3.3%)	adj-N-comp (47; 5.4%)	adj-n-comp (626; 72.0%)

Table 8.4: Overview of data subsets after dichotomization of `AttrRankAdj`, `AttrRankNoun` and `AttrRankComp` at rank 10; cells of the table contain subset identifier and number/proportion of items within the respective set

Compositional Processes. In a first step towards assessing the second criterion, we explore several segments of the large-scale development data set that are obtained from dichotomizing the variables `AttrRankAdj`, `AttrRankNoun` and `AttrRankComp` at rank 10, respectively. Thus, each of these variables is split into ranges indicating high (attribute rank ≤ 10) and low vector quality (attribute rank > 10).

Table 8.4 gives an overview of all subsets resulting from this segmentation process, together with their cardinalities. For instance, the upper left cell of the table displays that the data set contains 36 instances in total where all three rank variables have values lower than 10. The lower right cell, on the other hand, aggregates all 626 instances in the data with ranks above 10 for all three variables. Throughout the following discussion, we refer to these subsets by the following convention: If a variable has values lower than 10 (indicating that the corresponding vectors are of *HIGH* quality), its shorthand notation will be written in uppercase (*ADJ*, *N* or *COMP*, respectively), otherwise in lowercase letters.[7] Consequently, the subset in the upper left and lower right cells are abbreviated as *ADJ-N-COMP* and *adj-n-comp*, for instance.

The cardinalities of the subsets as given in Table 8.4 are insightful in various respects: First, instances with high-quality phrase vectors are clearly in the minority, as can be seen from comparing *-*-*COMP* (164 cases) and *-*-*comp* (705 cases). Apparently, the compositional gains outlined above are, in most cases, not strong enough in order to promote the correct attribute to ranks that are within the scope of entropy-based attribute selection. Second, within the *-*-*COMP* fraction, the majority of cases is due to high-quality adjective vectors (98 vs. 66 instances), which suggests that, in successful attribute selection, the contribution of adjective vectors to the composed meaning representation of an adjective-noun phrase is more prominent than the contribution of the noun. This is intuitively plausible given that nouns tend to offer a wider range of attributes in their semantics, from which the adjective selects the most appropriate one(s) in the given phrasal combination. From this perspective, the observations in the segments *ADJ-N-COMP* and *ADJ-n-COMP* can be attributed to the vital contribution

[7] Moreover, we will use $*$ as wildcard symbol denoting the following disjunctions: *ADJ* or *adj*, *N* or *n*, *COMP* or *comp*.

Phrase	Attribute	Rank Adj.	Rank Noun	Rank Phrase
breathtaking adventure	EXCITEMENT	65	2	3
low hill	HEIGHT	17	2	2
high ceiling	HEIGHT	89	1	3
broad shoulder	WIDTH	55	2	1
deep voice	PITCH	79	5	7

Table 8.5: Selected examples from *adj-N-COMP* subset

of the adjective vector, thus putting the compositional capacities of C-LDA in line with Pustejovsky (1995).

In case of *adj-N-COMP*, however, C-LDA seems to adapt to a claim put forward by Asher (2011): According to his argument, adjectival modification of nouns preserves the general semantic type of the noun. From this perspective on compositionality, adjectives are arguments of the noun conforming to the type presuppositions of the noun. Table 8.5 contains a selection of examples from the *adj-N-COMP* subset. These examples show a much stronger preference for the correct attribute in the lexical meaning of the noun, which can be readily interpreted in terms of Asher's notion of noun-triggered type presuppositions.

In case of suboptimal noun vectors, a high-quality adjective representation may even be sufficient in order to achieve an adequate phrase representation, as demonstrated by the *ADJ-n-COMP* segment. Likewise, if both the adjective and the noun are represented reliably, their composition will very unlikely result in a phrase vector that lags behind the quality of its individual constituents (cf. $|$*ADJ-N-comp*$| = 0$). Conversely, if neither the adjective nor the noun vector are of reliable quality, it is plausible to assume that the resulting phrase vector is not reliable either (cf. *adj-n-comp*).

Compositionality Puzzles. With regard to the hypothesis to be falsified – i.e, that linguistic principles of compositionality are insufficiently reflected in C-LDA –, we can state that compositional gains are indeed an integral part of the model's behaviour, which satisfies the first requirement for falsification. Concerning the second criterion, the subsets *ADJ-n-comp*, *adj-n-COMP* and *adj-N-comp*[8] still pose a puzzle for the otherwise consistent explanation of C-LDA predictions along the lines of compositional semantic processes. Given that these subsets account for only 12% of the data points in HeiPLAS-Dev, we conclude that C-LDA is in fact largely aware of compositionality.

Issues on the Level of Word Meaning. This leads to the conjecture that the poor performance of the model in large-scale attribute selection is primarily due to particular aspects on the level of word meaning which seem to obstruct the corpus-based induction of attribute-based representations of adjective and noun meanings in the first place.

[8]For illustration, all instances comprising these subsets are shown in Tables C.1-C.3 in Appendix C.

In fact, one major question that remains unanswered by the analysis conducted yet is: Why are more than 70% of the instances in the development data (cf. *adj-noun-comp* subset) affected by individual adjective and noun vectors that are highly error-prone in that the correct attributes are barely pronounced in these vectors?

In the next section, we explore a variety of semantic criteria in order to identify the most detrimental factors opposing effective attribute-based representations of word meaning in C-LDA and how they might be improved upon.

8.3 Linear Regression of C-LDA Performance at the Intersection of Word and Phrase Meaning

The previous analysis has revealed that the compositional aspects of the attribute selection task seem to be robustly captured by C-LDA in principle in the sense that compositionality gains are achieved by composing adjective and noun vectors to phrase-level representations. In practice, however, the traits of compositionality underlying C-LDA are not always sufficient in order to promote the correct attribute(s) in the phrase vector such that they are selected by the model. In order to subject this tension in C-LDA performance to a closer analysis, our discussion proceeds in three further steps:

1. On the level of phrase meaning, we assess the relative impact of various factors on *phrase vector quality*. To this end, we exhaustively examine all explanatory variables summarized in Table 8.1 on page 126 in a linear regression model of attribute ranks in phrase vectors. This analysis serves two purposes: From an explorative perspective, it allows for a more thorough investigation of possible predictors of phrase vector quality. Moreover, the learned regression model can afterwards be used as an objective function for optimizing C-LDA performance via complementary distributional information, as will be outlined in Section 8.4.

2. Zooming in on the word level, we take individual *word vector quality* into focus. This is achieved by means of a linear regression of attribute ranks in adjective and noun vectors on the same variables used in the first step.

3. Eventually, the perspectives taken in the previous steps are linked to each other. Our particular interest is to obtain a more comprehensive picture of the *compositional processes* at the intersection of the word and the phrase level in C-LDA.

Before delving into these three steps of analysis, we give a brief survey of the foundations of linear regression modelling in order to clarify the basic notions and the terminology to be used in the subsequent study.

8.3.1 Foundations of Linear Regression Modelling

This outline is entirely based on Cohen et al. (2003) and Fox (1997), unless explicitly stated otherwise.

Goals of Regression Modelling. Linear regression models are used to investigate the relationship between particular outcomes of a *dependent variable* (DV) Y as a function of one or a combination[9] of *independent variables* X (IVs; also denoted as *explanatory variables*, *regressors* or *predictors*):

$$Y = f(X) \tag{8.3}$$

Linear regression analysis postulates that f be a *linear* function in the sense that Y is determined by taking the weighted sum of the independent variable X and a constant a. Thus, the *model equation* has the following form:

$$Y = a + bX \tag{8.4}$$

Thinking of linear regression in geometrical terms, Equation (8.4) denotes the curve[10] relating Y to X. The weight or *coefficient* b determines the *slope* of this curve, while the constant a denotes its *intercept*, i.e. the point of intersection with the y-axis.

The primary goal of linear regression analysis is to assess the validity of a model that posits a hypothesis about the true relationship of a DV and the IVs in the form of Equation (8.4). Moreover, finding optimal estimates for the coefficients and the intercept contained in the model facilitates predicting the outcome of the DV for new data points. These goals are achieved by fitting the curve given by a function in the form of Equation (8.4) to a number of empirical observations.

Residual Error. In practice, however, a perfect fit of the model is rarely encountered. The *residual error* E is defined as the difference between the observed value Y and the predicted value $\hat{Y}$ for each observation:

$$E = Y - \hat{Y} = Y - (a + bX) \tag{8.5}$$

In the particular framework of *least-squares regression* that we will use for the subsequent analyses, the best regression curve incorporating optimal estimates for the coefficients and the intercept is obtained by minimizing S, the sum of the squared residuals over all N observations (Fox, 1997):

$$S(a, b) = \sum_{i=1}^{N} E_i^2 = \sum_{i=1}^{N} (Y_i - a + bX_i)^2 \tag{8.6}$$

[9] For the sake of simplicity, we limit this initial discussion to the case where the DV is determined by only one IV. Note, however, that linear regression models can be straightforwardly extended to an arbitrary number of IVs.

[10] If the model equation contains two IVs, its geometrical equivalent is a plane in three-dimensional space. Obviously, it is impossible to continue this geometrical analogy for larger numbers of IVs.

Interpreting Regression Results

Effect Size of Predictors. The most important aspect brought to light by a linear regression model concerns the strength of the relationship between the DV and the IVs. For each IV involved in the model equation, an individual coefficient indicating its *effect size* is computed. The effect size of each IV is characterized in terms of the magnitude of the coefficient and its sign, indicating a positive or a negative association to the DV.

In that sense, regression coefficients are similar to correlation coefficients such as Spearman's ρ or Pearson's r (Spearman, 1904; Pearson, 1896). Compared to these *single-factor* correlation models, linear regression models have the advantage that they are capable of studying multiple variables that *simultaneously* influence the DV, while separating their individual impact. It is important to consider this aspect in interpreting the coefficients estimated by a multi-factor regression model: Assuming a model

$$Y = a + b_1 X_1 + b_2 X_2,$$

the coefficient b_i pertaining to X_i indicates the average increase[11] of Y that is associated with a one-unit increase in X_i, if all X_j with $j \neq i$ (i.e., all other IVs in the model) are held constant. The intercept a has to be interpreted as the expected mean value of Y when $\sum_{i=1}^{N} X_i = 0$.

Interaction Terms. It is also possible to include *interaction terms* in a linear model. For instance, a model accounting for the effect on Y that is due to an interaction between a *predictor* X and a *moderator* Z can be designed as:

$$\hat{Y} = a + b_1 X + b_2 Z + b_3 XZ \tag{8.7}$$

The underlying assumption is that the relationship between Y and X is *moderated* by Z, i.e., the coefficient b_3 quantifying the relationship between Y and X is not assumed as constant, but varying with changes in Z – and analogously so, when predictor and moderator are interchanged (Cohen et al., 2003).

In interpreting interactions, we follow Cohen et al. (2003), who suggest to analyze the values predicted for Y by X at several meaningful values of Z. For this purpose, (8.7) is first refactored into the following *simple regression equation*:

$$\hat{Y} = (b_1 + b_3 Z)X + (a + b_2 Z) \tag{8.8}$$

Note that this equation actually describes a line, with its slope being determined by $b_1 + b_3 Z$ and $a + b_2 Z$ as its intercept, both depending on Z. The simple regression equation is very useful for analyzing interactions, as inserting meaningful values of Z (e.g., the mean, the maximum and the minimum) licenses to evaluate the change in $\hat{Y}$

[11] Consequently, a *decrease* in Y associated to X is indicated by a negative sign of b. Also note that, unlike single-factor correlation coefficients, regression coefficients are *not* confined to ranging from 0 to 1.

that is due to X, while Z is being controlled for in an instructive way. Moreover, a *simple regression line* can be constructed for each value of Z, in order to assess the interaction in graphical terms: Only in case of *non-parallel* regression lines, indicating a change of the regression of Y on X as a function of Z, we can readily accept the presence of an interaction between X and Z. Otherwise, the possibility of an interaction must be rejected as the regression of Y on X is constant for all values of Z. (Cohen et al., 2003)

Reliability of Linear Regression Models

In interpreting the results of a linear regression model, several aspects have to be taken care of, both on the level of the regression as a whole and the individual regression coefficients.

Overall Level. As an indicator of the "goodness of fit" between the empirical observations and the model predictions, we inspect the *squared multiple correlation*, R^2:

$$R^2 = \frac{\sum(\hat{Y}_i - \bar{Y})^2}{\sum(\hat{Y}_i - \bar{Y})^2 + \sum E_i^2} \tag{8.9}$$

R^2 indicates the proportion of the variation in Y that is accounted for by the combination of IVs in the model. Ranging between 0 and 1, higher values of R^2 indicate a better fit of the model. As noted by Cohen et al. (2003), R^2 is slightly biased in the sense that it tends to increase with more IVs in the model, even if these additional variables do not have any explanatory power. A more conservative measure that is sensitive for the number k of IVs in the model is R^2_{adj}, shorthand for *adjusted* R^2 as given in Equation (8.10). Here, N denotes the number of observations.

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{N - 1}{N - k - 1} \tag{8.10}$$

A second question of interest is whether the predictive power of the regression model is statistically significant. This question is commonly answered by an F-test evaluating the null hypothesis that all coefficients of the IVs in the model (except the intercept) are 0, which amounts to no predictive power at all. As usual in statistical significance testing, the p-value of this test gives the probability of the null hypothesis. If this probability is above 0.05, i.e., the null hypothesis can *not* be rejected, the p-values of the individual coefficients (see below) should be disregarded.

Note that the F-test is based on the assumption of normally distributed residual errors. If the residuals deviate markedly form normality, the test becomes invalid. In this case, the reliability of the regression model is questionable. In particular, the individual coefficients and their p-values should be interpreted with caution, as the model is likely to miss a major explanatory factor (Baayen, 2008).

The requirement of normally distributed residuals can also be motivated from the perspective that there should be no consistent over- or underestimation in the model.

Therefore, the *standard error* S_E being close to 0 is another desirable property of a linear regression model. This quantity is measured in terms of the average size of the residuals as defined below:

$$S_E = \sqrt{\frac{\sum E_i^2}{n - p - 1}} \tag{8.11}$$

In the denominator of (8.11), n stands for the total number of data points underlying the estimation, p for the number of IVs in the model. The denominator as a whole is often also referred to in terms of *degrees of freedom (df)* of the model.

Individual Coefficients. On the level of individual coefficients, the statistical significance of the estimates obtained from the model is also assessed in terms of p-values. The individual p-value of each coefficient indicates the probability of the null hypothesis that the coefficient has a value of 0 (i.e., no impact on the DV at all). If $p < 0.05$, we can be confident that the effect of the respective predictor on the DV as revealed by the model is *statistically significant*.

The quality of a regression model is also influenced by possible interactions among the predictors. In general, relationships between predictors may range from full *orthogonality* to strong *correlations*. In the former case, all preditors are completely *independent* of one another, i.e., each X_i explains a different part of the outcome of Y, which is very rare in practice. In the latter situation, also known as *multi-collinearity* (Belsley et al., 1980), "highly correlated independent variables are explaining the same part of the variation in the dependent variable, so their explanatory power and the significance of their coefficients is divided up between them." (Cohen et al., 2003)

If a severe degree of multi-collinearity is present in a model, individual regression coefficients are likely to change their magnitude and possibly even their sign when being considered in combination. Therefore, it is important to pay attention to this quantity when interpreting regression coefficients. A common measure of multi-collinearity is the *variance inflation factor (VIF)* that is computed for each independent variable X_i as given in (8.12):

$$\text{VIF}_i = \frac{1}{1 - R_i^2} \tag{8.12}$$

Here, R_i^2 denotes the squared multiple correlation (as introduced in Equation (8.9) above) of a regression model that incorporates X_i as the DV and all other $X_j (j \neq i)$ as IVs. Thus, a situation where the combination of all other IVs bears no explanatory power for X_i (i.e., they are fully uncorrelated) yields $\text{VIF}_i = 1$. The stronger the correlation between X_i and the other IVs, the larger VIF_i will increase. As a rule of thumb, Cohen et al. (2003) propose to take a VIF_i score above 10 as evidence for serious multicollinearity in the model that is to be attributed to X_i.

8.3.2 Phrase Level: Least Squares Regression of Phrase Vector Quality

The explanatory variables summarized in Table 8.1 on page 126 cover a wide range of properties of the adjective-noun phrases and their constituents involved in the attribute selection task. We now investigate these variables in a least squares regression model in order to identify the *main effects*, i.e., those factors with the strongest impact, on C-LDA performance.

This experiment was carried out on the HeiPLAS large-scale development set. After OOV filtering at level 2 (cf. Section 8.1 on page 125), 665 phrases (distributed over 254 attributes) are retained. Being annotated with values for each of the explanatory variables displayed in Table 8.1 on page 126, these phrases constitute the record of observations that enter the regression model. `AttrRankComp` is selected as the dependent variable.

Discussion of Full Model

We start out with a full regression model[12], taking into account all explanatory variables from Table 8.1. All variables indicating ranks or frequencies (including the DV) were transformed to a logarithmic scale in order to smooth their distributions and account for possible outliers.[13]

The full model yields an R^2 score of 0.926, which means that our selection of variables is definitely reasonable, as their combination explains almost 93% of the variance in the ranks of correct attributes in C-LDA phrase vectors. The regression as a whole is highly significant ($p < 2.2 \cdot 10^{-16}$) with a relatively small standard error ($S_E = 0.4488$). The main results with regard to the individual IVs are summarized in Table 8.6. The columns in this table show, for each IV in the model, its regression coefficient, VIF, p-value and a significance code[14].

As can be seen from this table, the variables `AttrRankAdj`, `AbstractnessAdj`, `AttrRankNoun`, `AbstractnessNoun` and `PhraseEntropy` have the strongest effect on the DV. All of them are statistically significant, except for `AbstractnessNoun`. Moreover, a small effect can be observed for `PropertyAttr` and `AdjEntropy`, at a rather low significance level, though. All variables in the model show very tolerable VIF scores clearly below 10, which suggests that multi-collinearity seems to be unproblematic in this model. In fact, most of the VIFs are close to the minimum of 1, with exceptions only in those cases where several variables from the same feature group are present in the model (e.g., several frequency or ambiguity features).

[12] All regression models reported in the following have been implemented in `R` (R Core Team, 2013), using the `lm` function. For model inspection and evaluation, the packages `rms` (Harrell, 2013) and `car` (Fox and Weisberg, 2011) were used.

[13] The same holds for all subsequent analyses as well.

[14] All significance codes follow the conventions used in `R`: $0 < p < 0.001$: '***'; $0.001 \leq p < 0.01$: '**'; $0.01 \leq p < 0.05$: '*'; $0.05 \leq p < 0.1$: '.'. Results above the 0.1 level are considered not significant. (R Core Team, 2013)

	Coefficient	VIF	p-value	Sig.
(Intercept)	-1.370900	d.n.a.	$6.80 \cdot 10^{-6}$	***
AttrRankAdj	0.715268	1.387924	$< 2 \cdot 10^{-16}$	***
AttrRankNoun	0.638044	1.215864	$< 2 \cdot 10^{-16}$	***
PropertyAttr=TRUE	-0.075221	1.578067	0.09747	·
MeasurableAttr=TRUE	-0.028699	1.690963	0.55679	
AbstractnessAttr	0.066882	1.138450	0.50401	
AdjEntropy	0.044287	1.412564	0.09323	·
NounEntropy	0.016622	1.296238	0.59310	
PhraseEntropy	-0.106237	1.661267	$8.52 \cdot 10^{-9}$	***
AttrSemcorEntropy	-0.015895	5.018194	0.75347	
PropAttrReadings	0.042109	1.808370	0.42265	
NumAttrSenses	0.001921	3.779184	0.93302	
AttrPseudoDocsFreq	-0.002986	2.517128	0.73807	
AdjFreq	-0.012117	6.780736	0.61573	
AdjPseudoDocsFreq	-0.004919	6.332131	0.82958	
AbstractnessAdj	-0.168825	1.155386	0.00285	**
AdjMorphAttr= TRUE	0.002593	1.258837	0.94791	
NounFreq	-0.001787	3.744689	0.92949	
NounPseudoDocsFreq	-0.004652	3.692467	0.82083	
AbstractnessNoun	0.114662	1.149755	0.22367	
PhraseFreq	0.009975	1.266668	0.14940	

Table 8.6: Results of full regression model, using all explanatory features for predicting `AttrRankComp`

	Coefficient	VIF	p-value	Sig.
(Intercept)	-1.271656	d.n.a.	$1.10 \cdot 10^{-11}$	***
AttrRankAdj	0.719686	1.223624	$< 2 \cdot 10^{-16}$	***
AttrRankNoun	0.640025	1.141171	$< 2 \cdot 10^{-16}$	***
PropertyAttr=TRUE	-0.089508	1.168155	0.02118	*
AdjEntropy	0.044310	1.305858	0.07877	·
PhraseEntropy	-0.104388	1.401142	$6.04 \cdot 10^{-10}$	***
PhraseFreq	0.009724	1.169529	0.14105	
AbstractnessAdj	-0.159653	1.069653	0.00317	**
AdjFreq	-0.017720	1.213130	0.08105	·
AbstractnessNoun	0.115377	1.046346	0.19645	

Table 8.7: Refined regression model after backward elimination (BE model)

Backward Elimination of IVs from Full Model

This initial model is subjected to a stepwise refinement procedure in order to eliminate (i) potential noise due to the relatively large number of insignificant predictors and (ii) redundancy due to several features from the same group. We follow an iterative strategy known as *backward elimination* (Miller, 2002): In each step, the variable with the largest p-value is removed from the model, as long as R^2_{adj} increases and S_E decreases. The model resulting from this procedure is summarized in Table 8.7 and will be referred to as *BE model* henceforth.

Results. First and foremost, backward elimination does not cause any substantial change in overall model behavior: The BE model is still highly significant as a whole ($p < 2.2 \cdot 10^{-16}$). $R^2 = 0.928$ and $S_E = 0.4459$ differ to an extent that is barely noticeable. This underlines that the variables removed during backward elimination are justified in their own right; some of them might even turn out as meaningful predictors whose impact is overridden by other variables in the model.

Next, we investigate important characteristics of the residuals. As discussed in Section 8.3.1 above, their normality and independence are requirements that must be largely met in order for a regression model to be fully reliable. As can been seen from the left side of Fig. 8.1 on the next page, however, the distribution of studentized residuals of the BE model actually resembles a *heavy-tailed* rather than a normal distribution. This means that, its mean close to 0 and the small deviations from this mean in the central part of the distribution notwithstanding, the BE model has a hard time predicting the correct outcome of the DV for the observations located at the tails of the distribution. Not only does this indicate a substantial proportion of outliers in the data (Fox, 1997); moreover, these outliers are extreme outliers compared to the data points in the center of the distribution. This analysis is supported by the quantile-quantile plot on the right side of Fig. 8.1 which compares the actual distribution of residuals against an idealized

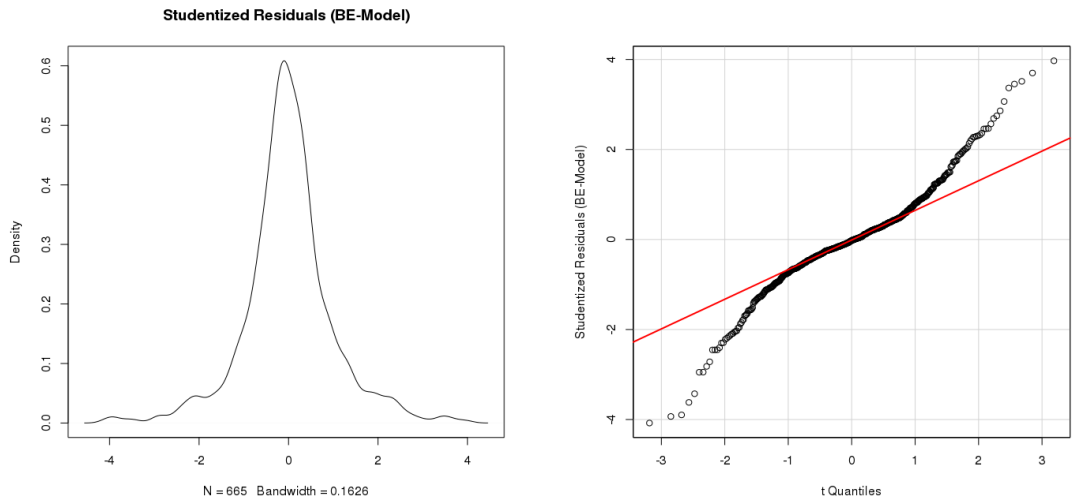


Figure 8.1: Studentized residuals in BE model, displayed as density plot (left) and quantile-quantile plot with superimposed line of normality (right)

line of normality superimposed on the plot. In summary, the BE model explains most of the observations very well, on the one hand, but also yields large errors for a minority of data points, on the other.

Outlier Removal. In order to explore to what extent the estimates of the coefficients and their significance levels are affected by these outliers[15], we follow a strategy of re-fitting the same model to a revised data set from which 20% of the data points accounting for the most extreme residuals have been purged beforehand (10% from the left and the right tail, respectively). The remaining 80% of the data fit the BE model much more accurately. As can be seen from the residuals plot in Fig. 8.2 (left), the heavy-tail issue is largely alleviated, apart from slight deviations from normality on the right tail of the distribution. In fact, almost all residuals emerging from the new fit follow a normal distribution within a confidence interval of 0.95 as indicated by the dashed curves surrounding the superimposed normality line in Fig. 8.2 (right).

Outlier removal leaves the regression coefficients and significance levels of the individual variables as displayed in Table 8.7 largely unchanged, which is why we do not explicitly report them here. Nevertheless, we consider this an important finding as it suggests that the results of the BE model are sufficiently reliable for our purposes.

[15]Recall from Equation (8.6) that, due to the least-squares minimization criterion, the models discussed here are in general very susceptible to large residuals.

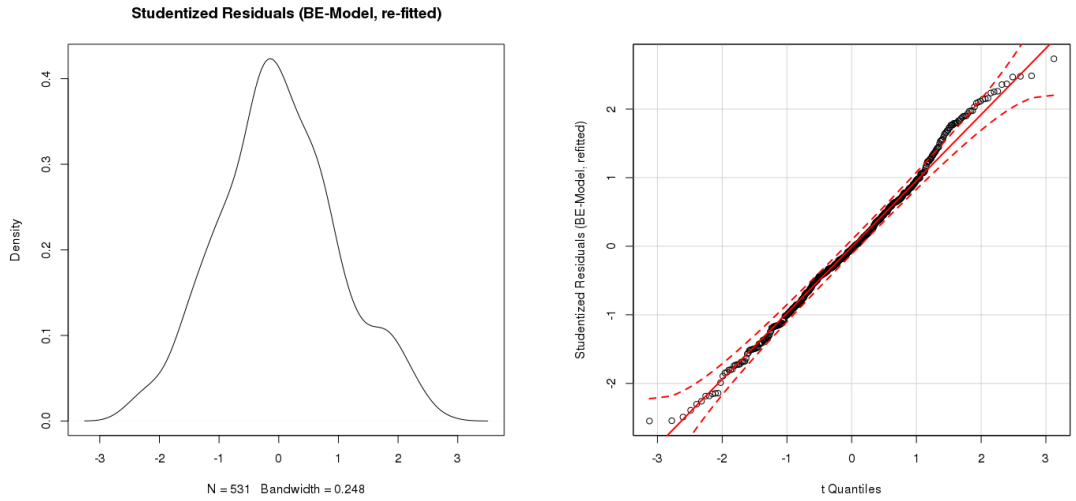


Figure 8.2: Studentized residuals in BE model after outlier removal, displayed as density plot (left) and quantile-quantile plot with superimposed line of normality surrounded by 0.95 confidence interval (right)

Interpretation of Backward Elimination Results

Main Effects. After outlier removal, the following IVs are identified as *simple main effects*: `AttrRankAdj`, `AttrRankNoun`, `PhraseEntropy`, `AbstractnessAdj` and `PropertyAttr`. Among them, `AttrRankAdj` and `AttrRankNoun` exhibit by far the largest effect sizes (0.719686 for adjective and 0.640025 for noun ranks, respectively)[16].

Apart from their effect sizes, further evidence for the importance of the rank variables comes from their relative contribution to the overall explanatory power of the BE model: A comparative regression model containing only `AttrRankAdj` and `AttrRankNoun` as IVs already achieves an R^2_{adj} score of 0.9211 at a very high level of significance ($p < 2.2 \cdot 10^{-16}$) and at the only expense of a slightly larger standard error (S_E=0.4632). Comparing these figures against the BE model (R^2_{adj}=0.9269; S_E=0.4459) points out that the composed rank as predicted by C-LDA is, above anything else, a function of the individual ranks.

Compositionality. These findings clearly corroborate the role of compositionality in C-LDA: On average, vector composition by pointwise multiplication improves upon

[16]These figures have to be interpreted such that, whenever an attribute rank in an adjective vector changes by one unit, the same attribute in a phrase vector with the corresponding adjective as one of its constituents can be expected to change by 0.72 units into the same direction on average, assuming that all other variables in the model are held constant. The same holds analogously for noun vectors, at a smaller effect size, though. Note that ranks and frequencies were transformed to log space, i.e., k units of change on the decimal scale equal e^k units on the logarithmic scale.

the quality of individual vector representations, yielding a composed vector with the correct attribute located at a lower rank (even though the degree of improvement is often not sufficient in order to be effective in attribute selection, as discussed in Section 8.2). In that process, the contribution of the adjective vector is more influential than the one of the noun. Additional support for this finding comes from an inspection of the means of `DeltaAdj` and `DeltaNoun`[17] as computed over all instances in the development set: On average, `DeltaNoun` is higher than `DeltaAdj` (14.41 vs. 11.66), which suggests that adjective vectors not only have a stronger impact on vector composition, but are also superior in terms of their individual quality.

Negative Suppressor Variables. Besides the paramount explanatory power of `AttrRankAdj` and `AttrRankNoun`, further significant contributions to phrase vector quality are due to `PhraseEntropy` and `AbstractnessAdj`. Note, however, that the negative sign of their coefficients seems controversial: If our intuitions as outlined in Section 8.1 were correct, we would expect lower entropy in a phrase vector to coincide with the correct attribute being located on a *lower* rank (and vice versa). Similarly, lower scores on the abstractness scale (denoting higher concreteness) should also result in *lower* ranks. Thus, positive coefficients for both `PhraseEntropy` and `AbstractnessAdj` would reflect our intuitions much better. Indeed, each of these variables shows a positive effect when being correlated with `AttrRankComp` in a bivariate setting, using Spearman's rank-correlation coefficient ρ: `PhraseEntropy` yields $\rho = 0.21$ ($p = 3.029 \cdot 10^{-8}$), `AbstractnessAdj` a slightly lower $\rho = 0.16$ ($p = 3.567 \cdot 10^{-5}$).

These discrepancies suggest that `AttrRankAdj` and `AttrRankNoun` might have a dominating influence not only on the DV, but on the other IVs contained in the BE model as well. In fact, `PhraseEntropy` and `AbstractnessAdj` show all properties of *negative suppressor variables*[18] of irrelevant variance in `AttrRankAdj` or `AttrRankNoun` (Darlington, 1968; Lutz, 1983).

Unveiling Negative Suppression Effects

In the interest of an unobstructed understanding of the effects associated with the potential suppressor variables, we compute another regression model, BE-NoR, by run-

[17] Recall from the definitions in Section 8.1 that `DeltaAdj` refers to the improvement of a noun representation that has been triggered by composition with an adjective vector; analogously, `DeltaNoun` refers to the improvement of an adjective vector due to the compositional contribution of a noun. Note that neither `DeltaAdj` nor `DeltaNoun` have been included in the full regression model (cf. Table 8.6), due to severe multi-collinearity.

[18] As summarized by Pandey and Elliott (2010), a negative suppressor is characterized by "removing irrelevant variance from a predictor (or set of predictors), increasing the predictor's regression weight, and increasing overall predictability of the regression equation". Furthermore, the negative suppressor exhibits a "positive zero-order correlation with other predictor variable(s) and with the outcome variable; however, when entered in multiple regressions, [...], contrary to what is expected, the regression weight of the negative suppressor has an opposite sign."

	Coefficient	VIF	p-value	Sig.
(Intercept)	3.13635	d.n.a.	$6.77 \cdot 10^{-11}$	***
PropertyAttr=TRUE	-0.20375	1.371234	0.121524	
MeasurableAttr=TRUE	-0.91891	1.488522	$2.16 \cdot 10^{-10}$	***
AttrPseudoDocsFreq	-0.14057	1.303204	$4.97 \cdot 10^{-12}$	***
AdjEntropy	0.12754	1.280087	0.102531	
AbstractnessAdj	-0.22184	1.067169	0.188554	
AbstractnessNoun	0.70757	1.048366	0.011621	*
PhraseEntropy	0.17651	1.325297	0.000516	***

Table 8.8: Regression model without dominating factors `AttrRankAdj` and `AttrRankNoun` (*BE-NoR* model)

ning an independent backward elimination on the BE model, initially leaving aside `AttrRankAdj` and `AttrRankNoun`.

A summary of the BE-NoR model is shown in Table 8.8. The regression as a whole is highly significant ($p < 2.2 \cdot 10^{-16}$). R^2 amounts to 0.2913, which is considerably lower compared to the BE model, but still indicates that the predictive power of a regression model explaining attribute ranks in C-LDA phrase vectors does not entirely depend on the presence of `AttrRankAdj` and `AttrRankNoun`.

Predictors in BE-NoR. Contrary to the BE model, BE-NoR does no longer contain the frequency features `PhraseFreq` and `AdjFreq`. It seems that global frequencies are somewhat helpful as correctives for predicting phrase ranks, but do not contribute any explanatory power individually. `PropertyAttr` and `AdjEntropy` are still present in BE-NoR, but no significant effect can be attested for them. Importantly, however, `AbstractnessNoun`, after being part of the BE model as an insignificant effect, now turns out as highly influential at a robust level of significance ($\beta = 0.70757; p = 0.011621$). Hence, completely in line with expectations from the literature (Andrews et al., 2009), our data shows that adjective-noun phrases containing a concrete noun fare considerably better in attribute selection than ones with abstract nouns do.

Additionally, BE-NoR involves two variables not included in the BE model: `MeasurableAttr`, actually being the strongest predictor in the model ($\beta = -0.91891; p = 2.16 \cdot 10^{-10}$), and `AttrPseudoDocsFreq` ($\beta = -0.14057; p = 4.97 \cdot 10^{-12}$). The estimated measurability coefficient indicates that, whenever the correct attribute is *not* measurable, its rank in the composed vector is about 2.5 positions higher[19], on average, holding all other variables constant. The observed effect of `AttrPseudoDocsFreq` is less strong, yet highly plausible, indicating that attributes with less training data are harder to rep-

[19] Note that `AttrRankComp` is transformed to log scale, while `MeasurableAttr` is not. Therefore, the exact increase in `AttrRankComp` that is associated with a 1-unit decrease of `MeasurableAttr` is $e^{0.91891} = 2.484323$.

	Coefficient	VIF	p-value	Sig.
(Intercept)	−0.98379	d.n.a.	0.53827	
MeasurableAttr=TRUE	−0.53978	2.241759	$8.68 \cdot 10^{-10}$	***
PropertyAttr=TRUE	−0.06369	1.516129	0.35309	
AttrPseudoDocsFreq	−0.13137	1.430201	$4.65 \cdot 10^{-10}$	***
PhraseEntropy	0.14420	1.622710	0.00962	**
PhraseFreq	0.06913	7.757674	0.19069	
AdjEntropy	0.32228	3.035512	0.00702	**
AbstractnessAdj	2.19053	51.139300	0.05885	·
AdjEntropy:AbstractnessAdj	−0.40627	53.314445	0.02936	*
AdjFreq	−0.07105	1.888976	0.07110	·
AdjMorphAttr=TRUE	−1.13485	51.430755	0.14743	
AdjMorphAttr=TRUE:AdjFreq	0.09560	42.785826	0.13644	
AdjMorphAttr=TRUE:AbstractnessAdj	0.57861	7.204284	0.18576	
AdjMorphAttr=TRUE:MeasurableAttr	0.28349	2.153931	0.03379	*
NounFreq	−0.02107	1.211155	0.55214	
NounEntropy	0.48281	5.472519	0.01485	*
AbstractnessNoun	7.25282	120.654973	0.01506	*
NounEntropy:AbstractnessNoun	−0.91651	128.793619	0.03342	*
PhraseFreq:AbstractnessNoun	−0.12109	7.197317	0.17213	
PhraseFreq:AbstractnessAdj	−0.04202	3.920773	0.37507	

Table 8.9: BE-NoR model with additional interaction terms

resent for the model.

Summarizing this comparison between the BE and BE-NoR models, we observe that BE seems to favor features encoding global frequencies and shapes of distributions, whereas in BE-NoR, semantic features become considerably more prominent.

Suppressor Variables revisited. The suppression hypothesis that originally gave rise to BE-NoR can be partially confirmed, at least, as `PhraseEntropy` now turns out as a positive factor, as expected, whereas `AbstractnessAdj` still behaves counterintuitively. Note that, compared to the BE model, the intercept changed its sign as well, which licenses a much more intuitive interpretation: If all IVs in the model are incidentally zero, the predicted composed rank becomes relatively high. Given the coding scheme of the IVs involved, this implies a situation where the correct attribute is neither a property nor measurable, no training material at all is available, the adjective and the noun are both extremely concrete in nature, the corresponding vectors representing the adjective and the phrase are highly peaked. Albeit completely virtual, this configuration underlines the relative impact of the different feature groups: Even for adjective-noun phrases that are represented by *ideally* shaped C-LDA vectors, the composed rank as predicted by the BE-NoR model will be no lower than 23, on average, if the attribute to be predicted exhibits detrimental semantic properties or a lack of training data.

Exploring Interactions between Features

The final model is constructed by including interaction terms between some of the previously insignificant variables[20] as formalized in (8.13). Interaction terms are highlighted in boldface.

$$\begin{aligned}\log(AttrRankComp) \sim\ & PropertyAttr + MeasurableAttr + \\ & AdjEntropy + AbstractnessAdj + \\ & \textbf{AdjEntropy * AbstractnessAdj} + \\ & NounEntropy + AbstractnessNoun + \\ & \textbf{Noun Entropy * AbstractnessNoun} + \\ & \log(AdjFreq + 1) + \log(NounFreq + 1) + \\ & \log(PhraseFreq + 1) + AdjMorphAttr + \\ & \textbf{MeasurableAttr * AdjMorphAttr} + \\ & \textbf{AdjMorphAttr *} \log\textbf{(AdjFreq+1)} + \\ & \textbf{AbstractnessAdj *} \log\textbf{(PhraseFreq+1)} + \\ & \textbf{AbstractnessNoun *} \log\textbf{(PhraseFreq+1)} + \\ & \textbf{AbstractnessAdj * AdjMorphAttr} + \\ & \log(AttrPseudoDocsFreq + 1) + PhraseEntropy \end{aligned} \tag{8.13}$$

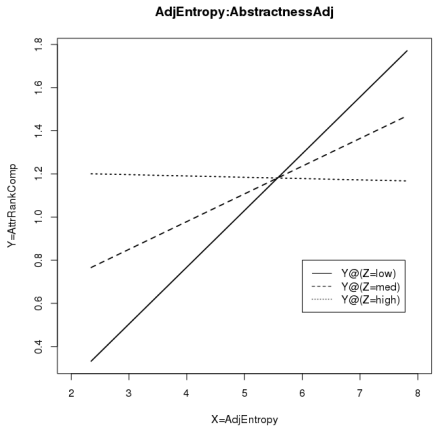
Compared to BE-NoR, this model yields slight improvements in terms of R^2_{adj} ($\Delta = 0.0114$) and S_E ($\Delta = -0.011$), while the overall significance of the model stays at the same, very high level ($p < 2.2 \cdot 10^{-16}$). This implies that the interaction terms are effective, which can also be recognized from the changes on the level of coefficients as displayed in Table 8.9 on the preceding page. We discuss these interactions in the order given in (8.13).

AbstractnessAdj moderating AdjEntropy. `AbstractnessAdj` finally turns out as a positive predictor with a rather high positive coefficient ($\beta = 2.19053$; $p = 0.05885$). Taking part in the interaction term also strengthens the impact of the `AdjEntropy` variable, both in terms of magnitude and significance ($\beta = 0.32228$; $p = 0.00702$). The interaction itself, denoted as `AdjEntropy:AbstractnessAdj` in Table 8.9, exhibits a negative impact on the DV ($\beta = -0.40627$; $p = 0.02936$).

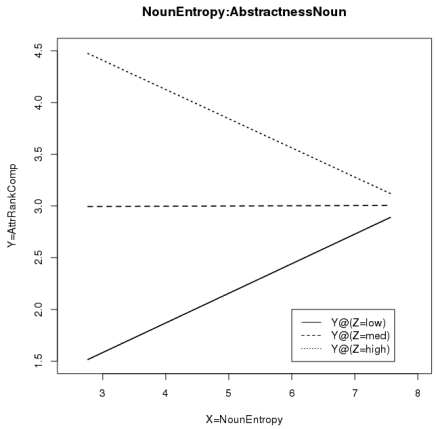
In interpreting this interaction, we follow the strategy suggested by Cohen et al. (2003), as summarized in Section 8.3.1 on page 139. Considering `AdjEntropy` as the predictor and `AbstractnessAdj` as the moderator, we select 0.2, 0.5 and 0.8 as meaningful values[21] for the latter. These values and the relevant coefficients from Table 8.9

[20] In this investigation, possible interaction terms are limited to combinations of adjective/adjective, noun/noun, adjective/attribute and noun/attribute factors.

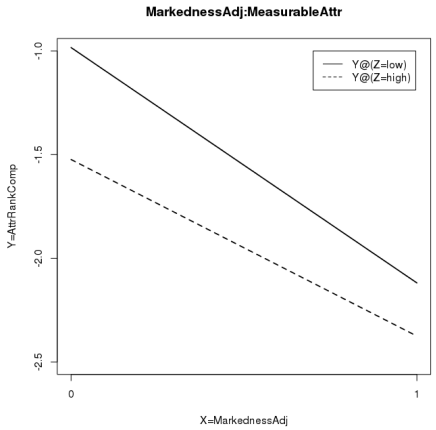
[21] These values were selected as to include representatives from the "concrete"(0.2) and the "abstract" (0.8) poles plus one from the middle (0.5) of the spectrum.



(a) Y = `AttrRankComp` on X = `AdjEntropy` at three levels of Z = `AbstractnessAdj`



(b) Y = `AttrRankComp` on X = `NounEntropy` at three levels of Z = `AbstractnessNoun`



(c) Y = `AttrRankComp` on X = `AdjMorphAttr` at two levels of Z = `MeasurableAttr`

Figure 8.3: Simple regression lines for three interaction terms

are used to compute three simple regression lines for $Z = low$, $Z = med$ and $Z = high$, with X ranging from the minimum to the maximum observed value of `AdjEntropy`. The resulting lines are displayed in Fig. 8.3a.

Most importantly, the three lines are not parallel, which allows us to accept the interaction as a valid statistical factor, as there is obviously a change in `AttrRankComp` as predicted by `AdjEntropy` that is due to different values of `AbstractnessAdj`. Moreover, comparing the lines against each other gives interesting insights into the characteristics of the interaction: If Z values are low (indicating concrete adjectives; solid line), attribute ranks in the phrase vector increase with increasing entropy. This is perfectly in line with our expectations and holds, to a slightly smaller degree, also for adjectives being located in the middle of the abstractness scale (dashed line). For highly abstract adjectives, however, we barely find any differences in composed ranks predicted based on adjective vectors of varying entropy (dotted line), which implies that in this particular segment of the data, the predictive power of individual vector entropy is drastically diminished by semantic properties of the respective adjective.

AbstractnessNoun moderating NounEntropy. The impact of `AbstractnessNoun` increases by an order of magnitude due to its interaction with `NounEntropy`, turning out as the most important factor in the model ($\beta = 7.25282$; $p = 0.01506$). `NounEntropy`, after remaining insignificant in the BE-NoR model, is now rendered a significant contributor of modest relative importance ($\beta = 0.48281$; $p = 0.01485$). Again, the impact of the interaction term on the DV is negative in its own right ($\beta = -0.91651$; $p = 0.03342$).

The simple regression lines displayed in Fig. 8.3b[22] indicate that only in case of concrete nouns (low Z values, solid lines) low vector entropy coincides with low attribute ranks in the composed vector, as would be expected. In the medium range of Z (dashed line), noun abstractness exerts no influence at all on the relationship between `NounEntropy` and `AttrRankComp`, whereas for rather abstract nouns (high Z values, dotted line), low vector entropy favors high attribute ranks. Evidently, vector quality and vector entropy drastically diverge in this segment of the data: C-LDA is still capable of producing highly peaked representations for abstract nouns; at the same time, the likelihood of going astray in promoting the correct attribute in the composed vector considerably increases.

MeasurableAttr moderating AdjMorphAttr. The effect of `MeasurableAttr` acting as a moderator in the relationship between `AdjMorphAttr` and `AttrRankComp` is definitely statistically valid, albeit rather small ($\beta = 0.28349$; $p = 0.03379$), as confirmed by the almost parellel simple regression lines in Fig. 8.3c. These lines have to be interpreted such that the correct attribute(s) are ranked lower in phrase vectors involving a morphologically marked adjective. Holding for both measurable and non-measurable at-

[22] Computing simple regression lines for the `NounEntropy:AbstractnessNoun` interaction follows the same procedure as described above for `AdjEntropy:AbstractnessAdj`.

	Coefficient	VIF	p-value	Sig.
(Intercept)	4.15515	d.n.a	$2.60 \cdot 10^{-13}$	***
AdjEntropy	0.29606	1.048864	$1.06 \cdot 10^{-5}$	***
AdjFreq	-0.09319	1.636237	0.007633	**
AdjMorphAttr=TRUE	-1.74947	37.394894	0.006045	**
AttrPseudoDocsFreq	-0.09925	1.322539	$2.38 \cdot 10^{-7}$	***
MeasurableAttr=TRUE	-1.21749	1.846985	$2.24 \cdot 10^{-15}$	***
AdjFreq:AdjMorphAttr=TRUE	0.15141	39.908664	0.010440	*
AdjMorphAttr=TRUE:MeasurableAttr=TRUE	0.93839	1.752005	0.000166	***

Table 8.10: Regression model for predicting attribute ranks in adjective vectors

tributes, the moderation effect is slightly stronger in the former case (cf. flatter slope of the dashed line).

All other interactions included in this model (cf. Equation (8.13) on page 150 again) are not significant. Apart from the factors just discussed, the variables already involved in the BE-NoR model are left largely unchanged by the interactions introduced here.

As a last finding on Table 8.9, we point out that there are differences in orders of magnitude among several coefficients, with `AbstractnessNoun` and `AbstractnessAdj` being the most influential ones by far. This corroborates that attributes provide a layer of meaning that is generally favored by concrete rather than abstract words.

Summary of Findings on Phrase Level

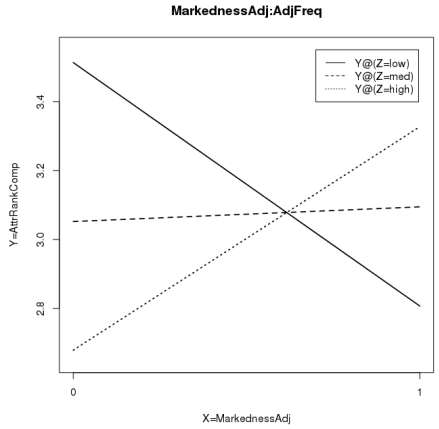
Overall, the results of regressing phrase vector quality support the conjecture that the compositional approach taken by C-LDA in order to compute attribute-based representations of adjective-noun phrases is largely reasonable. Main sources of the inadequacies becoming apparent in large-scale attribute selection are due to the construction of individual attribute-based word representations. Deepening the analysis by focussing on word vector quality in the next section, we are interested in identifying the factors that are detrimental to effective adjective and noun representations in C-LDA.

8.3.3 "Zooming in": Regression of Word Vector Quality

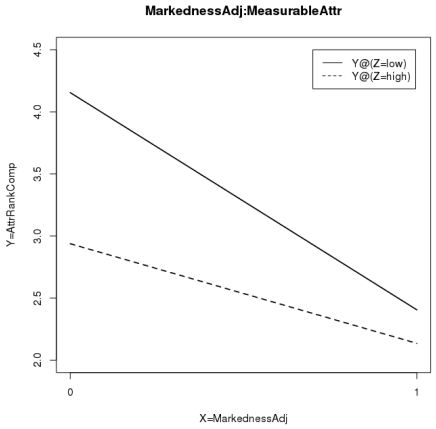
Explaining Adjective Vector Quality

Regression of attribute ranks in adjective vectors starts out with all adjective and attribute features from Table 8.1 on page 126. Removing insignificant predictors by backward elimination and including interaction terms where appropriate, we finally arrive at the model summarized in Table 8.10. This model yields an $R^2_{adj} = 0.2505$ at an extremely high level of significance ($p < 2.2 \cdot 10^{-16}$).

We identify two independent predictors of adjective vector quality that do not take part in any interaction, both of them being highly significant, albeit rather modest in



(a) Y = `AttrRankAdj` on X = `AdjMorphAttr` at three levels of Z = `AdjFreq`



(b) Y = `AttrRankAdj` on X = `AdjMorphAttr` at two levels of Z = `MeasurableAttr`

Figure 8.4: Simple regression lines for two interaction terms

magnitude: `AdjEntropy` ($\beta = 0.29606$; $p = 1.06 \cdot 10^{-5}$) and `AttrPseudoDocsFreq` ($\beta = -0.09925$; $p = 2.38 \cdot 10^{-7}$). Not surprisingly, the quality of adjective representations benefits from a sufficient amount of training data, and low entropy in adjective vectors coincides with better vector quality in terms of lower ranks of the correct attribute(s).

Moreover, two important semantic factors stand out in this model: `AdjMorphAttr` ($\beta = -1.74947$; $p = 0.006045$) and `MeasurableAttr` ($\beta = -1.21749$; $p = 2.24 \cdot 10^{-15}$), both exerting a highly significant *negative* impact on the DV. In the case of the measurability variable, the negative sign confirms the pattern already observed in our previous analysis of attribute ranks in phrase vectors (cf. Table 8.9): Obviously, adjectives denoting measurable attributes are favored by C-LDA not only on the phrase level, but already on the level of individual word representations.

The negative coefficient of `AdjMorphAttr` indicates that C-LDA is remarkably effective for *morphologically marked* adjectives, thus overcoming the inherent sparsity issues that render this subclass of adjectives a notorious challenge to corpus-based approaches. Effectiveness on morphologically marked adjectives is mitigated by two interactions, however: `AdjMorphAttr:AdjFreq` ($\beta = 0.15141$; $p = 0.010440$) and `AdjMorphAttr:MeasurableAttr` ($\beta = 0.93839$; $p = 0.000166$). These interactions are visualized in terms of simple regression lines[23] in Figs. 8.4a and 8.4b. Considering `AdjMorphAttr:AdjFreq` first, we observe that particularly infrequent, morphologically marked adjectives (low Z values, solid line) are modeled more effective than their morpholog-

[23] Recall from its definition in Section 8.1 on page 128 that `AdjMorphAttr` is actually a binary factor taking only 0 and 1 as values. Therefore, the predicted values of $\hat{Y}$ are strictly points; for the sake of better visibility, however, we decided to link these points by lines anyway.

	Coefficient	VIF	p-value	Sig.
(Intercept)	2.11095	d.n.a.	0.0226	*
NounEntropy	0.18629	1.093246	0.0268	*
NounFreq	0.12979	6.306811	0.0924	·
AbstractnessNoun	3.12733	33.337728	0.0362	*
AttrPseudoDocsFreq	-0.09730	1.227773	$1.50 \cdot 10^{-7}$	***
MeasurableAttr=TRUE	-0.52998	1.313636	$3.16 \cdot 10^{-5}$	***
NounFreq:AbstractnessNoun	-0.26440	38.045267	0.0614	·

Table 8.11: Regression model for predicting attribute ranks in noun vectors

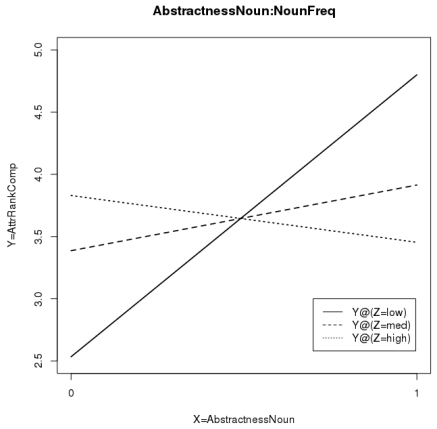


Figure 8.5: Simple regression lines for $Y =$ `AttrRankNoun` on $X =$ `AbstractnessNoun` at three levels of $Z =$ `NounFreq`

ically unmarked counterparts in this segment of the data. With increasing frequency of occurrence, however, this trend is reversed (cf. dashed and dotted line). With respect to the `AdjMorphAttr:MeasurableAttr` interaction, we find that measurability of the correct attribute does not make much of a difference in case of morphologically marked adjectives, given that both lines start at similar points for $X = 1$. Morphologically unmarked adjectives end up at higher ranks in general, while this overall trend is attenuated by adjectives denoting measurable attributes (higher Z values, dashed line).

Explaining Noun Vector Quality

Table 8.11 summarizes the regression of `AttrRankNoun` on all noun and attribute features. As before, insignificant variables were removed by backward elimination, interaction terms were included where appropriate. In this model, R^2_{adj} amounts to 0.1302 only, overall significance being extremely high, though ($p < 2.2 \cdot 10^{-16}$).

The factors `NounEntropy` ($\beta = 0.18629$; $p = 0.0268$), `AttrPseudoDocsFreq` ($\beta =$

-0.09730; $p = 1.50 \cdot 10^{-7}$) and `MeasurableAttr` ($\beta = -0.52998$; $p = 3.16 \cdot 10^{-5}$) behave as expected, i.e., in line with previous analyses of adjective and phrase vector quality (cf. Tables 8.9 and 8.10). As a notable exception, we find a positive coefficient for `NounFreq` ($\beta = 0.12979$; $p = 0.0924$), indicating that C-LDA tends to favor low-frequency over high-frequency nouns. Recurring to the relationship of high word frequency and strong polysemy that has been frequently observed in the literature (Kilgarriff, 1997; Kremer et al., 2012, i.a.)[24], we consider this an interesting tendency (even though small in impact and barely significant) suggesting that C-LDA noun representations are susceptible to multiple senses of highly frequent nouns being conflated in one vector representation.

`AbstractnessNoun` turns out as the most important predictor in this model ($\beta = 3.12733$; $p = 0.0362$). Concrete nouns are modeled much more adequately than abstract ones in general, even though this pattern is slightly mitigated by an interaction with `NounFreq` ($\beta = -0.26440$; $p = 0.0614$), as can be seen from the simple regression lines in Fig. 8.5: Low ranks are observed for concrete nouns when their frequency of occurrence is low (cf. solid line), whereas increasing frequencies of occurrence add a considerable amount of noise to this pattern (cf. dashed and dotted line).

8.3.4 Compositional Processes: Linking Word and Phrase Level

Having analyzed various factors with respect to their impact on C-LDA word and phrase vector quality, we now summarize our findings by linking the levels of word and phrase meaning to each other. We argue that this will enable interesting perspectives on the compositional processes taking place at the intersection from word to phrase meaning.

Table 8.12 presents an overview of all factors that turned out significant in at least one of the regression models on the adjective, noun or phrase level. The direction of the effect is given in terms of $+$, $-$ or 0, where 0 indicates that a significant effect on one level could not be reproduced on the other. This table enables a distinction of two types of factors according to their behavior on both levels of analysis: Variables with coefficients proven to be statistically significant and concordant (with regard to their direction) on both the word and the phrase level are denoted *consistent*, all others *inconsistent*.[25] Consistent factors are mostly directly propagated from the word to phrase

[24] In fact, the factors `NounFreq` and `NumNounSenses` (accounting for the number of different word senses according to WordNet) show a rather strong positive correlation in our data ($\rho = 0.55$; $p < 2.2 \cdot 10^{-16}$), confirming the claim put forward by Kilgarriff (1997).

[25] Recall from Section 8.3.1 that the coefficients involved in a regression function can only be interpreted in the context of all other variables in the respective model. Therefore, the overview presented in Table 8.12 might be criticized as an overgeneralization from a strictly mathematical perspective, as it implies a comparison of individual coefficients across several regression models. Nevertheless, we argue, the concentration of perspectives conducted here may still reveal interesting insights into the interplay of word and phrase meaning in C-LDA, as long as the conclusions to be drawn are not interpreted in terms of exact rules but rather general tendencies.

Feature	Adj. Level	Noun Level	Phrase Level
AdjEntropy	+	n/a	+
NounEntropy	n/a	+	+
PhraseEntropy	n/a	n/a	+
AdjFreq	−	n/a	−
NounFreq	n/a	+	0
PhraseFreq	n/a	n/a	0
AttrPseudoDocsFreq	−	−	−
MeasurableAttr	−	−	−
AdjMorphAttr	−	n/a	0
AdjMorphAttr:AdjFreq	+	n/a	0
AdjMorphAttr:MeasurableAttr	+	n/a	+
AbstractnessAdj	0	n/a	+
AbstractnessAdj:AdjEntropy	0	n/a	−
AbstractnessNoun	n/a	+	+
AbstractnessNoun:NounFreq	n/a	−	0
AbstractnessNoun:NounEntropy	n/a	0	−

Table 8.12: Impact of features on word and phrase vector quality

level as a result of multiplicative vector composition, whereas inconsistent factors shed light on interesting compositional processes underlying C-LDA. In that respect, this summary also complements the compositionality analysis in Section 8.2.

Consistent Factors

Entropy Features. Consistently across adjective, noun and phrase representations, vector entropy turns out as a reasonable approximation of vector quality: The lower its entropy, the lower the rank of the correct attribute(s) within a vector. On the other hand, vector entropy – and particularly phrase vector entropy – is clearly not sufficient to explain vector quality in its own right, as is reflected by the rather modest coefficients of the entropy variables throughout all levels of analysis: There is a considerable number of C-LDA vectors that exhibit low entropy, yet promote incorrect attributes for a given adjective-noun phrase.

Frequency Features. Densely populated pseudo-documents providing large amounts of training data for the individual attributes are generally supportive for attribute selection. For adjectives, we find that the larger their global frequency, the better both their individual vector quality and the quality of the composed phrase vectors they take part in. Only if an adjective is sufficiently frequent in the first place, there is a chance to capture its attribute meaning reliably by means of surface dependency patterns.

Attribute Semantics. In line with the results from re-training C-LDA on confined subsets of attributes (cf. Section 7.3.4), the correct attribute being measurable is strongly advantageous. If the correct attribute is *not* measurable, however, attribute selection is most likely to yield a suboptimal result for the respective adjective-noun phrase.

Interaction between Adjective and Attribute Semantics. Attribute measurability also acts as an important moderator on morphological relatedness. In general, adjective representations generated by C-LDA tend to prefer morphologically marked adjectives[26], unless the respective adjective denotes an attribute that is measurable. Note that this explains the relative advantage of L-LDA over C-LDA on the subset of measurable attributes (cf. Table 7.16 on page 119 and Fig. 7.6 on page 121). Apparently, adjectives denoting measurable attributes require a great deal of disambiguation rather than smoothing.

In a nutshell, major obstacles to attribute-based adjective representations as generated by C-LDA are due to infrequent, morphologically unmarked adjectives denoting attributes that are not measurable. As a result of multiplicative vector composition, these detrimental characteristics of adjective meaning are directly propagated from the word to the phrase representation.

Noun Semantics. Abstractness of the noun consistently stands out as the factor with the strongest impact on the noun and the phrase level, indicating that attribute selection by C-LDA stronlgy benefits from concrete nouns.

Inconsistent Factors

We now turn to the factors found to be inconsistent, i.e., to exhibit interesting variance across the word and the phrase level. From the perspective of semantic compositionality, there are two explanations for this kind of divergences:

1. If a factor is encountered in one constituent, but not on the phrase level, it must have been *compositionally overridden* by the other constituent.

2. If a factor is encountered on the phrase level, but in only one of the constituents, it must have been *compositionally introduced* by the other constituent.

In the following, we explore compositional effects that are triggered by nouns and override or introduce some aspects of adjective meaning. After that, we focus on the inverse effect, i.e., the noun being affected by compositional processes triggered by the adjective.

[26] This effect is present both on the word and the phrase level; it is not significant on the latter, though.

Adjective-based Recovery from Insufficient Nouns. Abstractness of adjectives acts as a positive factor on the phrase level, strongly in favor of concrete adjectives. On the phrase level, abstractness further acts as a moderator on the positive relationship between adjective entropy and phrase vector quality, such that low entropy coincides with lower ranks for concrete adjectives; for adjectives showing a medium or even high degree of abstractness, however, the predictive power of vector entropy considerably diminishes. On the word level, there is no evidence for abstractness as a valid main effect, and neither for the interaction between abstractness and vector entropy, which suggests that some properties in the semantics of the noun result in the appearance of adjective abstractness as a significant predictor on the phrase level. We illustrate this by way of the examples[27] given in Table 8.13 on the next page.

It turns out that, if `NounEntropy` is high (either because the noun is highly abstract in nature and tends to resist attribute-based modeling or because it is highly polysemous and collapsing the attribute profiles of its various senses into one vector representation yields an uninformative distribution), low values of `AbstractnessAdj` and `AdjEntropy` are essential in order to achieve good phrase vector quality. The example phrases *abnormal power* and *short life* as given in Table 8.13 are cases in point, where uninformative, low-quality vector representations resulting from rather abstract nouns are leveled out by concrete adjectives whose vector representations are much more selective and of excellent quality, thus empowering C-LDA to select the correct attribute in both cases. This analysis is further supported by the phrase *equal terms* as displayed in the same table: The noun exhibits similar properties with regard to abstractness and vector entropy, whereas the adjective is in itself rather abstract and poorly modeled. Consequently, the composed vector is incapable of capturing the correct attribute meaning of the phrase. We visualize this pattern in Fig. 8.6 on the following page[28].

Morphologically marked Adjectives: Disambiguation and Deterioration Effects. `AdjMorphAttr` emerges as another inconsistent factor of adjective meaning. Ranking morphologically marked adjectives lower than morphologically unmarked ones on average, C-LDA can be considered very effective in alleviating sparsity issues in adjective representations that are due to morphological relatedness. Highly frequent adjectives, however, are responsible for a reversal in this preference, which we traced back to their relatively high ambiguity potential (cf. Kremer et al., 2012). Both these effects are leveled out by vector composition, which suggests some heterogeneous influence of C-LDA noun representations.

[27] In selecting these examples, we controled for highly abstract, highly frequent and highly polysemous nouns resulting in uninformative, low-quality vector representations. The adjectives exhibit various patterns of abstractness and individual vector entropy in order to demonstrate in which cases these features of an adjective make a difference in predicting phrase vector quality.

[28] In this and the subsequent visualizations of compositional processes underlying C-LDA, nodes labeled with $+$ represent variables with a positive impact on phrase vector quality, whereas variables with a detrimental impact are labeled with $-$.

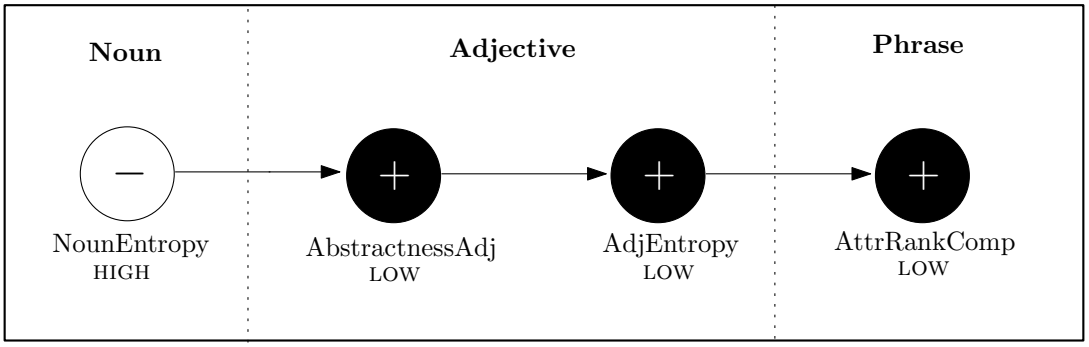


Figure 8.6: Adjective-based Recovery from Insufficient Nouns

	abnormal power (NORMALITY)	*short life* (DURATION)	*equal terms* (EQUALITY)
NounEntropy	7.361527	7.016088	7.420787
AbstractnessNoun	0.628940	0.631840	0.700700
NounFreq	96320	200215	61831
NumNounSenses	9	14	7
AdjEntropy	6.104194	4.230792	7.022189
AbstractnessAdj	0.459440	0.563920	0.608870
AttrRankAdj	1	2	77
AttrRankNoun	71	119	71
AttrRankComp	1	1	62

Table 8.13: Examples for adjective-based recovery from insufficent nouns

	high mountain (HEIGHT)	*modern history* (MODERNITY)
AdjMorphAttr	TRUE	TRUE
AdjFreq	1147823	263866
AdjEntropy	5.670378	5.599181
NounEntropy	6.858342	6.936956
AbstractnessNoun	0.281690	0.471770
NounFreq	13842	131973
NumNounSenses	2	5
AttrRankAdj	89	62
AttrRankNoun	1	1
AttrRankComp	5	6

Table 8.14: Examples for disambiguation of ambiguous morphologically marked adjectives by nouns

	evil purpose (EVIL)	*offensive remark* (OFFENSIVENESS)	*nasty trick* (NASTINESS)
`AdjMorphAttr`	TRUE	TRUE	TRUE
`NounEntropy`	6.908904	6.576570	6.488986
`AbstractnessNoun`	0.634390	0.584540	0.543240
`NounFreq`	45542	1635	4592
`NumNounSenses`	3	2	7
`AttrRankAdj`	2	5	8
`AttrRankNoun`	164	245	205
`AttrRankComp`	6	33	37

Table 8.15: Examples for deterioration of morphologically marked adjectives on the phrase level

On the one hand, the `AdjMorphAttr:AdjFreq` interaction disappears, which we interpret in terms of a positive disambiguation effect being triggered by nouns. This is shown in Table 8.14 for the adjectives *high* and *modern*: Both of them may also convey evaluative readings, which are strongly dispreferred in the context of the nouns *mountain* and *history*, however. Note that these examples nicely illustrate the view purported by Asher (2011) that the contribution of the noun constituent is predominant in the compositional semantics of adjective-noun phrases in that nouns introduce type presuppositions that are preserved throughout adjectival modification. On the other hand, the overall preference for morphologically marked adjectives on the word level may also deteriorate as a result of vector composition. We conclude that in a considerable number of cases, the contribution of the noun vectors causes C-LDA attribute selection to go astray, as illustrated in Table 8.15.

In both cases, it is hard to identify noun-specific characteristics that are responsible for the observed disambiguation or deterioration effect. We conjecture that the typicality of the respective attribute in the concept denoted by the noun might play a role, given that HEIGHT in *mountain* and MODERNITY in *history* (triggering disambiguation; cf. Table 8.14 and Fig. 8.7) are intuitively much more typical[29] than EVIL in *purpose*, OFFENSIVENESS in *remark* or NASTINESS in *trick* (all triggering deterioration; cf. Table 8.15). Given that typicality of features strongly correlates with production frequency (Mervis et al., 1976), it seems plausible that relations between a noun and attributes that are highly prominent in its meaning are more frequent in corpora, which shapes the resulting attribute-based noun vector in a more accentuated way.

Apart from typicality, we suppose that surface sparsity might also be caused by *indirect predications*. By this term we refer to adjective-noun phrases where the attribute relation does not hold between the adjective *Adj* and the noun *N* directly, but rather

[29] We are using the term *typicality* along the lines of prototype theory (Rosch, 1973) here, assuming that some attributes are more central to the meaning of a concept than others.

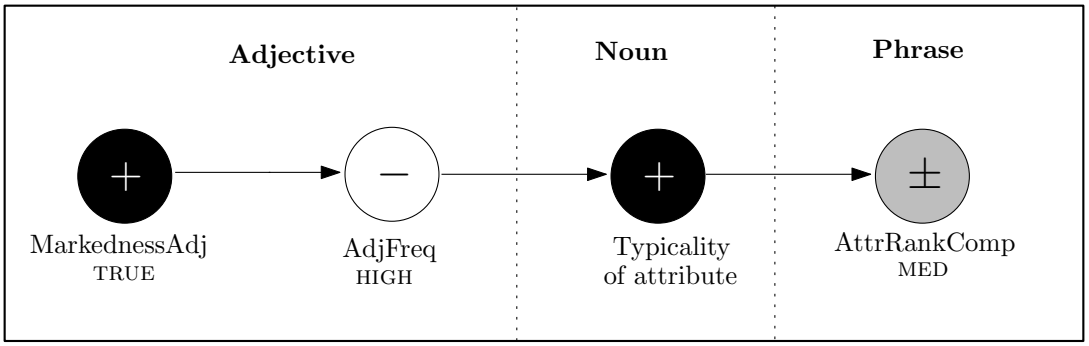


Figure 8.7: Disambiguation of morphologically marked, highly frequent adjectives by nouns

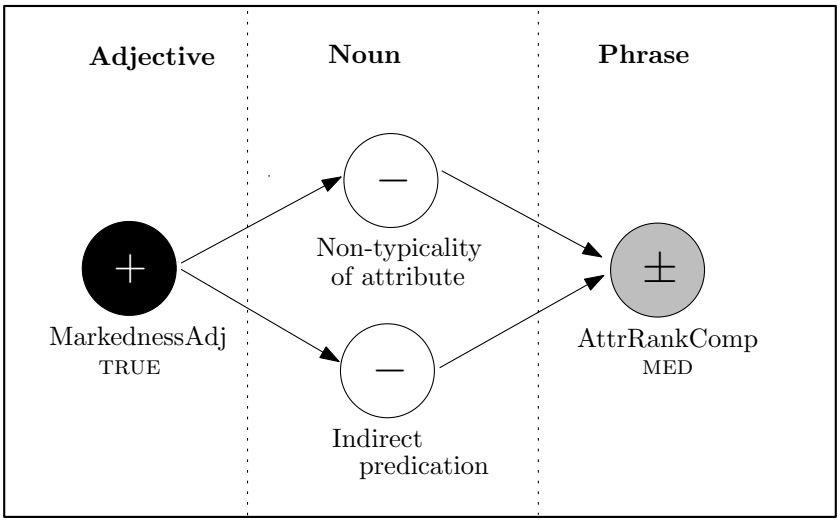


Figure 8.8: Deterioration of morphologically marked adjectives by nouns

between *Adj* and an entity or event N' that is semantically related to N. The phrase *nasty trick* (cf. Table 8.15) is a case in point, as *nasty* is not exactly a predicate over *trick*; rather the person that *does* the trick acts in a nasty way. Consequently, *nastiness* and *trick* are rather unlikely to occur in an overtly observable relation (other than, for instance, *nastiness* and *person*). Therefore, even though NASTINESS is quite well represented in the adjective vector, the noun vector is almost completely ignorant of this aspect of meaning. See Fig. 8.8 for a visualization of this deterioration process. Note that this process provides an explanation for one of the "compositionality puzzles" encountered in Section 8.2 (cf. ADJ-n-comp subset in Table 8.4 on page 135).

Noun Disambiguation by Adjectives. Inconsistent factors involving C-LDA noun representations comprise `NounFreq` (as a main effect), `AbstractnessNoun` and `NounEntropy`

	short life (DURATION)	*great work* (IMPORTANCE)	*difficult time* (DIFFICULTY)	*uncomfortable day* (COMFORT)
`NounFreq`	200215	348754	569317	257258
`NumNounSenses`	14	7	10	10
`AdjEntropy`	4.230792	7.368324	7.023659	6.145984
`AbstractnessAdj`	0.563920	0.573950	0.799080	0.593280
`AdjFreq`	418048	1225742	301267	16989
`AttrRankAdj`	2	11	18	193
`AttrRankNoun`	119	20	67	71
`AttrRankComp`	1	1	19	151

Table 8.16: Examples for disambiguation of nouns by adjectives

$$\vec{\text{life}} = \begin{pmatrix} \text{ASSURANCE} : & 0.045 \\ \text{HOLINESS} : & 0.035 \\ \text{HANDINESS} : & 0.033 \\ \text{QUALITY} : & 0.033 \\ \text{ORDINARINESS} : & 0.027 \\ \text{CHANGEABLENESS} : & 0.025 \\ \text{HAPPINESS} : & 0.025 \\ \text{SEX} : & 0.021 \\ \text{BENIGNITY} : & 0.021 \\ \text{FULLNESS} : & 0.019 \\ \vdots & \vdots \end{pmatrix}$$

(a) Word vector: *life*

$$\vec{\text{short}} \odot \vec{\text{life}} = \begin{pmatrix} \text{DURATION} : & 0.18 \\ \text{NICENESS} : & 0.12 \\ \text{CONSTANCY} : & 0.11 \\ \text{DISTANCE} : & 0.09 \\ \text{SENSATIONALISM} : & 0.08 \\ \text{LENGTH} : & 0.03 \\ \text{BENEFICENCE} : & 0.03 \\ \text{DULLNESS} : & 0.03 \\ \text{PITCH} : & 0.02 \\ \text{BENIGNITY} : & 0.02 \\ \vdots & \vdots \end{pmatrix}$$

(b) Phrase vector: *short life*

Figure 8.9: Fragment of attribute-based vector representations comprising 10 most prominent dimensions

(in interactions).

Following our analysis of individual noun vector quality, we interpret `NounFreq` as a proxy of noun polysemy. Against this background, we consider it a desirable property of the C-LDA model that high noun frequency is found as a detrimental feature on the word level that is leveled out by vector composition, as this indicates that adjective representations are largely capable of resolving ambiguity issues present in attribute-based noun representations.

Consider the examples in Table 8.16: All phrases shown are composed of highly frequent nouns, each of them exhibiting numerous senses according to WordNet. Conflating this variety of senses in one vector representation may result in a rather blurred, hardly selective attribute profile as sketched, for the noun *life*, in Fig. 8.9. Note that at least two senses of *life* coalesce in the ten most prominent dimensions of its C-LDA word vector, as the attributes ASSURANCE, HOLINESS and ORDINARINESS are related

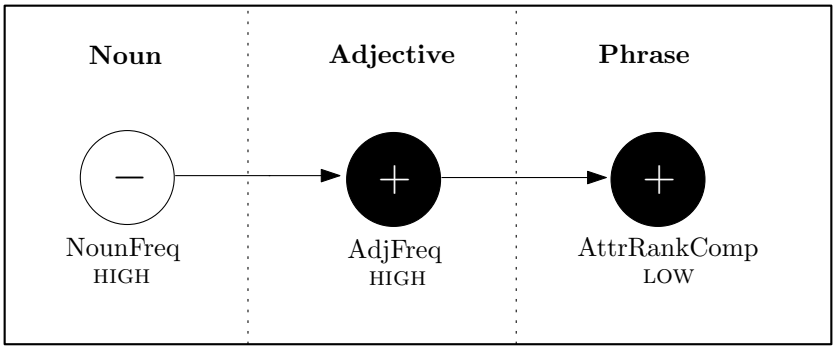


Figure 8.10: Disambiguation of nouns by adjectives

to a collective interpretation of the term, whereas QUALITY, CHANGEABLENESS, HAPPINESS and FULLNESS clearly denote aspects of individual living[30]. Thanks to the contribution of the adjective, the phrase vector as depicted in Fig. 8.9 correctly promotes the DURATION attribute to the first rank and generally tends to favour other aspects of individual living (e.g., NICENESS, CONSTANCY).

With regard to the specific characteristics of the adjectives that support disambiguation, we find that among different configurations of abstractness, morphological relatedness, vector entropy and frequency as summarized in Table 8.16 on the previous page, the latter factor is the most decisive one. Apparently, high corpus frequencies of adjectives are generally beneficial, since they increase the chance of obtaining more accentuated adjective representations, and may even contribute to the resolution of ambiguity issues in noun representations as summarized in Fig. 8.10.

Susceptibility of abstract nouns to unselective adjectives. Moreover, `AbstractnessNoun` moderates `NounEntropy` on the phrase level such that only for concrete nouns, low-entropy word representations coincide with high phrase vector quality. In case of abstract nouns, C-LDA may still produce highly selective noun vectors that tend to promote the wrong attributes in the phrase meaning, however. This interaction between `AbstractnessNoun` and `NounEntropy` has not been observed on the level of individual noun representations, which indicates a considerable proportion of inauspicious combinations of particular adjectives with abstract nouns in the data.

The most important characteristics of the adjectives involved in such combinations can be understood from the examples in Table 8.17 on the facing page. The adjective vectors involved in these examples are either not selective enough in order to reinforce the correct attribute(s) as found in the noun vector (cf. the relatively high adjective vector entropy in *manifest disapproval* and *accurate measurement*), or simply fail to capture the attribute meaning of the adjective by giving strong preferences to incorrect attributes

[30] These senses are paraphrased in WordNet as denoting "living things collectively" and "the actions and events that occur in living [...] of an individual", respectively.

	manifest disapproval (OBVIOUSNESS)	*unswerving devotion* (CONSTANCY)	*accurate measurement* (ACCURACY)
NounEntropy	6.242087	5.773019	6.358582
AbstractnessNoun	0.692660	0.725450	0.679850
AbstractnessAdj	0.767840	0.621350	0.638840
AdjEntropy	6.619185	5.079476	6.351236
AdjFreq	7290	514	87214
AttrRankAdj	198	179	46
AttrRankNoun	16	20	6
AttrRankComp	101	49	13

Table 8.17: Examples for susceptibility of abstract nouns to unselective adjectives

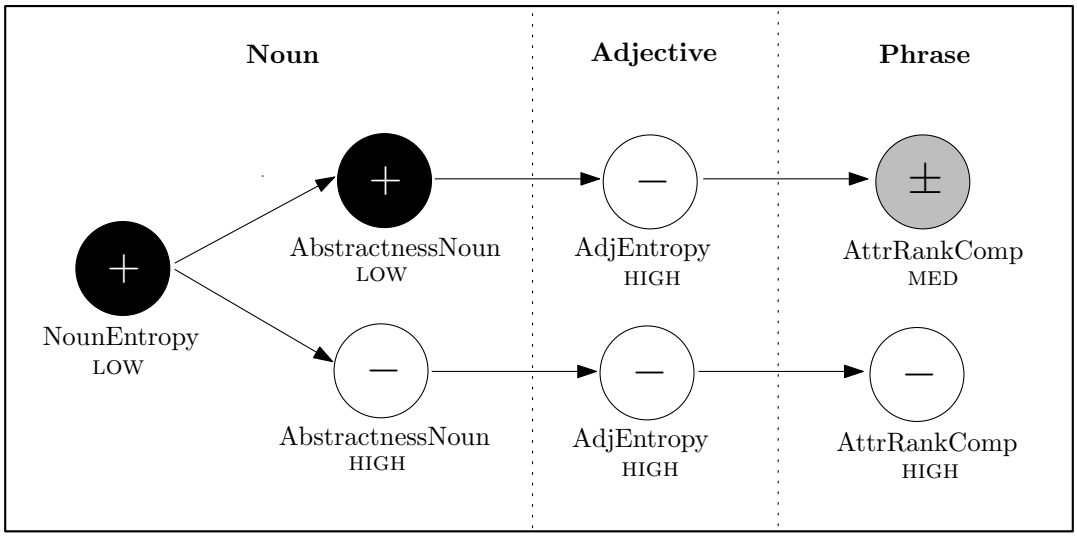


Figure 8.11: Susceptibility of Abstract Nouns

(cf. low entropy and low quality of the adjective vector in *unswerving devotion*), thus causing a deterioration of individual noun vector quality. We conclude that abstract nouns, even if represented by sufficiently selective word vectors, are particularly susceptible to poor adjective representations, as summarized in Fig. 8.11. Note that this pattern provides an explanation for another "compositionality puzzle" from Section 8.2, i.e., the adj-N-comp subset (cf. Table 8.4 on page 135).

8.3.5 Major Findings and Discussion

The analysis presented above helped to reveal some of the major strengths and weaknesses of the C-LDA attribute model. On the positive side, adjective-noun phrases most likely to be effectively modeled by C-LDA tend to exhibit a combination of features

	wide road (WIDTH)	*young people* (AGE)	*hot water* (TEMPERATURE)
AdjEntropy	6.684649	5.160293	5.903032
NounEntropy	6.751190	6.836761	6.139176
PhraseEntropy	4.440063	2.696725	2.044959
MeasurabilityAttr	TRUE	TRUE	TRUE
AttrPseudoDocsFreq	8413	73234	28751
AdjFreq	473767	642947	145770
AdjMorphAttr	TRUE	FALSE	FALSE
AbstractnessNoun	0.322410	0.445570	0.326570
AttrRankAdj	8	1	1
AttrRankNoun	1	7	3
AttrRankComp	1	1	1

Table 8.18: Examples exhibiting beneficial combinations of consistent factors

such as the following (cf. Table 8.18 for exemplary instantiations of such phrases):

- low vector entropy
- highly populated pseudo document for the correct attribute
- high corpus frequency of the adjective
- measurable attribute
- morphologically marked adjective
- concrete noun

As a particular strength of the model, we emphasize that C-LDA is capable of generating reliable adjective representations even in cases of morphological relatedness, where purely pattern- or dependency-based approaches are doomed to failure (cf. analysis of smoothing power in Section 7.3.2). In general, C-LDA adjective vectors are of a better individual quality than their counterparts representing nouns, as revealed by their respective impact on the quality of composed vectors.

On the other hand, the analysis also clearly points out the limitations of an attribute-based distributional model such as C-LDA: First and foremost, the inventory of attributes investigated in the large-scale experiment is separable along a measurability axis that strongly correlates with selection performance: Measurable attributes gently favor, while non-measurable ones tend to resist attribute selection.

Measurable attributes largely subsume *physical* or *experiential* properties. Given that learning abstract word meanings from experientially grounded data is considered to be

sense	*consequence*	*regard*	*assumption*	*attitude*
explanation	*disregard*	*understanding*	*distinction*	*sympathy*

Table 8.19: 10 nouns in HeiPLAS-Dev set with highest values in abstractness

a "fundamental challenge" (Andrews et al., 2009), it is not surprising that C-LDA faces a challenge in modelling abstract concepts. Even human subjects may find it difficult to provide succinct attribute-based descriptions for this class of concepts[31].

Against this background, the fact that C-LDA does provide adequate vector representations for a subset of the data provides supporting evidence for the view that it is – at least partially – possible to induce experientially grounded knowledge from purely textual sources (cf. Baroni et al., 2010), even at the higher granularity of attributes rather than individual properties.

Finally, we were able to identify a number of compositional processes underlying C-LDA attribute models which either (i) enable the system to recover from some of the detrimental characteristics encountered on the level of modeling word meaning (e.g., disambiguation of polysemous nouns or highly frequent, morphologically marked adjectives), or (ii) explain some of the weaknesses of C-LDA in large-scale attribute selection, e.g., the susceptibility of abstract nouns to unselective adjectives or the deterioration of morphologically marked adjectives due to indirect predications[32].

8.4 Options for Enhancing C-LDA Performance

Capitalizing on the results of the regression anaylsis just presented, we explore the prospects of optimizing C-LDA performance by improving individual word vector representations in order to yield attribute-based vector representations that are more informative with regard to the correct attribute(s) denoted by an adjective-noun phrase and, thus, better suited for attribute selection.

We frame this task as an *optimization problem* in a distributional enrichment framework[33]. In this framework, individual word vectors are enriched by means of complementary distributional information, such that a particular *objective function* yields an

[31] We encourage the reader to try this for the examples in Table 8.19 which constitute the ten most abstract nouns in the development set.

[32] The latter phenomenon is closely related to event-based adjectives which already surfaced in Chapter 5 of this thesis in the context of adjective classification. Apart from obvious parallels in their linguistic behavior, indirect predications and event-based adjectives have in common that they are very hard for human subjects to identify in corpus annotation studies: As discussed in Section 5.1.4, we were not successful in establishing a separate class of event-based adjectives on empirical grounds. Likewise, the annotators of the HeiPLAS data, despite being instructed accordingly in the annotation guidelines (cf. Appendix B), were not fully reliable in eliminating indirect predications from the WordNet examples. In fact, that is the only reason why example phrases involving indirect predications such as *nasty trick* and *warm coat* do actually appear in the data analyzed in this study.

[33] Formal details of distributional enrichment are deferred until Chapter 9.

	Phrase Entropy	Noun Entropy	Adj. Entropy
all attributes	0.21***	0.11**	0.15***
property attributes	0.17***	0.08	0.16**
measurable attributes	**0.47***	0.06	0.34***
selected attributes	0.35***	0.05	0.18**

Table 8.20: Correlation scores (Spearman's ρ) between composed ranks and entropy of word and phrase vectors

optimal value. For this task, we consider phrase vector quality as a criterion for the objective function to be maximized. More precisely, we will use the regression equations learned in regressing phrase vector quality (cf. Section 8.3.2) as objective functions in the optimization process, arguing that the impact of a particular vector update operation (i.e., whether an update results in an improved or an impaired vector representation) can be evaluated in terms of the expected changes in the dependent variable.

As discussed throughout Section 8.3, there are several options for regressing phrase vector quality which consequently result in different options to formulate an objective function based on this quantity. These are briefly explained in the following.

Oracle. An obvious strategy to optimize phrase vector quality is to formulate an objective function in terms of *AttrRankComp*:

$$\text{minimize } \mathit{AttrRankComp} \tag{8.14}$$

Note that this objective function includes an oracle as it involves information about the correct attribute, which is generally not available to the model at prediction time. Therefore, it is *not* compatible with an unsupervised setting.

No-Ranks. As a second option, the regression equation obtained from the BE-NoR model (cf. Table 8.8 on page 148) can be used as objective function. This model is less adequate in explaining phrase vector quality, but fully compatible with an unsupervised distributional setting as it does not leverage any external information about which attributes are correct for a given example phrase. Removing all insignificant factors from the original regression equation yields:

$$\begin{aligned}\text{minimize } \mathit{AttrRankComp} =& 3.14 - 0.92\,\mathit{MeasurableAttr} \\ &- 0.14\,\mathit{AttrPseudoDocsFreq} \\ &+ 0.71\,\mathit{AbstractnessNoun} \\ &+ 0.18\,\mathit{PhraseEntropy}\end{aligned} \tag{8.15}$$

Minimal. A *minimal* objective function relies on `PhraseEntropy` as the only factor to be minimized:

$$\text{minimize } AttrRankComp = PhraseEntropy \tag{8.16}$$

Apart from its parsimony[34], this objective function shows another interesting property in that `PhraseEntropy` is expected to harmonize particularly well with the measurability filter discussed above. This can be seen from Table 8.20, where correlation scores between composed ranks and vector entropy in phrase and individual word vectors, respectively, are compared for different attribute inventories.

In Chapter 9, these settings will be embedded into a novel distributional enrichment framework for systematic vector updating that aims to establish a principled strategy to improve the performance of topic-based attribute models in attribute selection.

8.5 Summary

This chapter was devoted to a linear regression analysis of various factors that determine the performance of C-LDA in the task of attribute selection from adjective-noun phrases. Our intention was three-fold: (i) to determine strengths and weaknesses of the system at the levels of lexical meaning and compositionality, (ii) to gain insights into the linguistic processes at the intersection of lexical and phrasal meaning, and (iii) to devise an optimization strategy that is targeted at improving the predictive quality of C-LDA in attribute selection.

Compositionality. The rules of compositionality are found to be largely intact and successfully mirrored in C-LDA, given that vector composition yields relative improvements in vector quality in more than 90% of the instances in the large-scale development data. We consider this a justification of our approach to treat attributes as an abstract layer of meaning that is shared between adjectives and nouns. Attribute selection can be seen as an intersective process, and the multiplicative vector composition method anchored in C-LDA mimics this intersective character of the problem quite well. With regard to the relative contribution of the constituents to the compositional meaning of the phrase, we found that the adjective is more influential than the noun. This is interesting from a theoretical perspective, as it qualifies attribute selection as a process where the adjective selects the most appropriate attribute(s) from a range provided by the noun, in line with Pustejovsky (1995).

In sum, the compositional aspect of the attribute selection problem is captured by C-LDA; sources of error mainly concern the semantic layer of word meaning and its suitability to be modeled by attribute-based vector representations, as well as certain properties of attribute meaning (to be summarized below).

[34] According to *Occam's Razor*, simple theories are generally preferable over more complex ones: "Accept the simplest explanation that fits the data" (MacKay, 2003).

Lexical Meaning and Attribute Meaning. On the level of attribute meaning, we found that measurability of the correct attribute for each example phrase is highly beneficial for attribute selection. To a lesser extent, this also holds for attributes being subclassified as properties in WordNet. On the other hand, attributes that are neither measurable nor a property pose a hard challenge to C-LDA. Due to their abstractness and fine granularity, grouping them toghether (either manually or using automatic clustering approaches) might be a promising avenue to pursue in future work. These aspects have been attested both in a regression analysis of phrase vector quality and in an experiment that evaluates C-LDA performance on different subsets of attribute inventories. On the contrary, lexical ambiguity on the level of attributes is not an issue for C-LDA.

On the level of word meaning, we found a general qualitative superiority of adjective over noun vectors. Moreover, we highlighted several lexical properties that are beneficial to attribute selection, among them morphological relatedness of adjectives (showing that C-LDA smoothing is very effective in overcoming sparsity issues due to rare occurrences of adjectives with their morphologically marked attribute nouns in linguistic surface patterns) and a limited degree of polysemy as well as semantic concreteness of nouns.

Intersection of Word and Phrase Meaning. Linking the results obtained from regressing word and phrase vector quality to one another, we identified several linguistically plausible processes at the intersection of word and phrase meaning. In particular, C-LDA shows robust capabilities in resolving lexical ambiguities. Disambiguation is found to be mostly triggered by adjectives, but – in particular circumstances – also by nouns. Moreover, confirming both the stronger compositional impact of adjectives over nouns in C-LDA and their qualitative superiority, adjectives also have the potential to help the model "recover" from insufficiently modeled noun meanings.

Optimization Strategies. Having traced back most of the deficits of C-LDA to the level of lexical modeling, we will develop a new optimization strategy for distributional semantic models using factors from regression analysis as objective function for optimization. Along these lines, a complete framework for enhancing attribute selection performance of topic-based attribute models will be formulated in the next chapter.

9 Distributional Enrichment: Improving Structured Vector Representations

Distributional semantic models are usually instantiated as *structured* or *unstructured* ones, thus enforcing either *specificity* or *density*. Depending on its intended purpose, it may seem reasonable to maximize specificity or to minimize sparsity in a distributional model: Lexical tasks such as attribute selection, for instance, clearly call for an interpretable, structured model such as C-LDA, as shown in the previous chapters of this thesis. Using the same model in a similarity judgement task for predicting phrasal similarity for pairs of adjective-noun phrases, however, turns out not successful (Hartung and Frank, 2011a).

In this chapter, we propose a novel framework for *distributional enrichment* in order to combine the specific advantages of structured and unstructured distributional models. Key to distributional enrichment is the idea to augment structured representations of individual words to centroids of their nearest neighbours, while keeping the principle of meaning representation along structured, interpretable dimensions intact and at the same time increasing the overall density of the semantic space. The selection of nearest neighbours is carried out in an *auxiliary model*, i.e., a distributional space that represents semantic information from a different perspective that is complementary to the original model. This approach aims at reducing the sparsity of the structured meaning representations, while preserving their specificity to the largest extent possible.[1]

Distributional enrichment can be instantiated in various ways and, hence, tailored to various lexical tasks. In the context of this thesis, the motivation for distributional enrichment arises from the poor performance of topic-based attribute models in large-scale attribute selection.

This chapter is structured as follows: In the next section, we provide more technical detail on the general idea of distributional enrichment, including background and motivation. In Section 9.2, we describe the notion of *auxiliary models* which is at the core of distributional enrichment. We discuss possible instantiations of auxiliary models and assess their suitability for enhancing topic-based attribute models in a systematic benchmark test against the BLESS data set (Baroni and Lenci, 2011). A formal description of the distributional enrichment framework is given in Section 9.3. Besides, this section presents concrete instantiations used in an experiment on applying the frame-

[1] Erk et al. (2010) follow a similar idea by relying on a *primary* and *secondary* corpus for selectional preference modeling. In their approach, however, the aspects of complementarity and preservance of specificity are less pronounced than in the framework proposed here.

work to C-LDA and L-LDA attribute models to improve their large-scale attribute selection performance. The results of this experiment are summarized and discussed in Section 9.4. Section 9.5 concludes the chapter by summing up the major findings.

9.1 General Idea and Overview

Similarly to recent work by Zhang et al. (2014), distributional enrichment aims at model combinations from different semantic perspectives. In cases where semantic representations generated by structured distributional models are too sparse or otherwise unreliable, distributional enrichment exploits complementary sources of distributional information and integrates them into an orginal (structured) distributional space. As a major challenge in distributional enrichment, auxiliary models have to be designed as to preserve most of the semantic information contained in the original model, while generating denser and more reliable meaning representations.

Due to their underlying complementarity, original and auxiliary models usually differ in their dimensionality. Therefore, their vector representations are not easily interoperable, i.e., vectors from the original and the auxiliary model cannot be directly combined with each other.

In addition, the distributional enrichment framework to be proposed here stipulates a vector update process which iterates between the original and the auxiliary model, as depicted in Figure 9.1. This figure contains three distributional models in their graphical representation: the **original model before the update** (which in our case is an attribute-based structured distributional model) in the top left area, an **auxiliary model** (which we posit as an unstructured distributional model) at the bottom, and the **original model after the update**[2] in the upper right area. The visualization of these models is restricted to three dimensions each. These dimensions are semantically different, which is indicated by different axis labels: In the original model(s), they correspond to attributes (denoted as a_1, a_2 and a_3), while the dimensions of the auxiliary model correspond to arbitrary context words (denoted as c_1, c_2 and c_3).

Distributional enrichment comprises the four steps of *target retrieval*, *carrier selection*, *carrier projection* and *centroid construction* (indicated in the figure by labels ① to ④), which we describe in the following by way of an informal example:

① **Target Retrieval:** Assume that the attribute-based vector representation of the noun *tractor* in the original model is unreliable. The vector representation of *tractor* in the auxiliary model is looked up.

② **Carrier Selection:** This step aims at collecting additional semantic information in the auxiliary model. To this end, a number of vectors that are located in close

[2] In fact, the update process underlying distributional enrichment *integrates* complementary distributional information provided by the auxiliary model into the original model. The difference between the originial model before and after the update is made here only for ease of presentation. Technically, they are the same object.

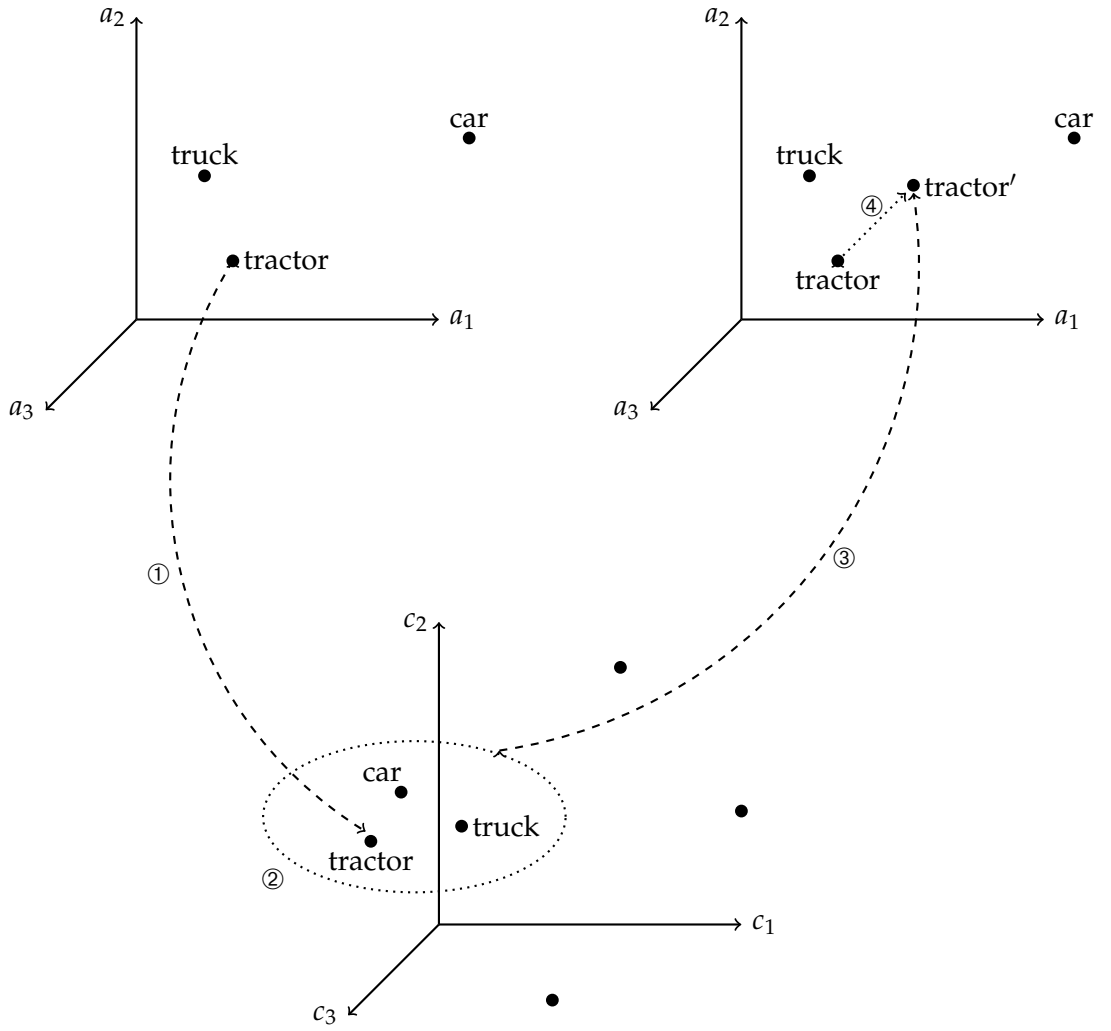


Figure 9.1: Distributional enrichment as a four-step process including: ① *target retrieval* in the auxiliary model, ② *carrier selection* in the auxiliary model, ③ *carrier projection* back into the original model, and ④ *centroid construction* in the original model

proximity to *tractor* in auxiliary space are collected. We refer to these vectors as *carriers*. In our example, *car* and *truck* are selected as carriers for *tractor*, as indicated by the dotted region in Figure 9.1. An important question is how to determine the number of carriers to be selected for a given target. This will be discussed in Section 9.3.

③ **Carrier Projection:** For each of the selected carriers, carrier projection retrieves structured vector representations from the original model. In our example, this yields attribute-based vectors for *car* and *truck*. This step can be seen as the inverse of target retrieval. Note that up to now, no additional information has been added into the original model.

④ **Centroid Construction:** Finally, the structured vector representations corresponding to the carriers are used to construct a centroid vector in the original space. In our example, the centroid, denoted as *tractor*′ is computed from the structured vectors of *tractor*, *car* and *truck*. Only at this point, the original model is updated such that the vector originally representing *tractor* is replaced by the centroid *tractor*′. The selection of carriers and their projection to a centroid vector has to be optimized in such a way that the new structured representation of *tractor* is semantically closer to *car* and *truck* (cf. dotted arrow in Figure 9.1), and thus more reliable relative to the original vector.

Assumptions. This procedure is aimed at the derivation of richer structured representations in the sense that the properties characterizing a particular vector are passed on to its semantic neighbours. Overall, this leads to a denser and more coherent semantic space. Two requirements must be be met to achieve this result: First, significant overlap between targets and carriers is required, i.e., the original model must provide structured meaning representations for a substantial proportion of carrier elements in the first place. Second, the auxiliary model must be constructed in such a way that the carriers can be expected to promote exactly the intended features, i.e., proximity in the auxiliary model needs to correlate with semantic relatedness between target and carriers.

9.2 Auxiliary Distributional Models

As discussed in the previous section, the main function of auxiliary models in distributional enrichment is to select appropriate carrier vectors from the semantic neighbourhood of the targets of enrichment. Thus, auxiliary models provide a complementary semantic perspective in the interest of higher overall density of the original structured model. Carrier selection can be performed in a paradigmatic manner by nearest neighbour search (Yianilos, 1993) or by relying on the strength of association between a target and syntagmatically related context words. In both cases, it is crucial that the carriers

are carefully selected in order to avoid semantic drift effects (McIntosh and Curran, 2009) when being projected back into the structured model and used for centroid construction. With respect to our goal of using distributional enrichment for enhancing structured attribute models, the carrier vectors need to be selected such that they are *attribute-preserving*, i.e., that their attribute profiles are semantically compatible with the target representations to be enriched.

9.2.1 Benchmarking First- and Second-order Auxiliary Models for Attribute-preserving Carrier Selection

In this section, various options for constructing auxiliary models will be subjected to a benchmark investigation in order to assess their suitability for acquiring attribute-preserving carrier vectors. In particular, we compare the prospects of *first-order* vs. *second-order* auxiliary models for this task.

Recall from Section 2.3.4 that first-order distributional models use purely syntagmatic contexts as features to describe a target word, whereas second-order models capitalize on contexts of first-order context words. Due to higher density in the resulting meaning representations, second-order distributional models can a priori be expected to be particularly effective in discovering paradigmatic neighbours.

Paradigmatic relatedness is not necessarily equivalent to an attribute-preserving relationship, though: For instance, the phrases *small boy* and *small house* establish a valid paradigmatic relation between *boy* and *house*. Obviously, however, these concepts do not share an attribute-preserving relation.[3] This leads to the question as to which distributional contexts facilitate the acquisition of attribute-preserving semantic neighbours in first- and second-order distributional models. We approach this question by subjecting various instantiations of both types of models to the BLESS benchmark (Baroni and Lenci, 2011) which has been specifically designed for comparatively evaluating distributional semantic models of different kinds. In this study, we assume that membership in the same ontological category is a sufficient condition for two concepts to be substantially compatible in their attribute meaning (cf. Baroni et al., 2010).

Benchmark data. As argued by Baroni and Lenci (2011), a thorough evaluation of vector space models should address the aspects of (i) "determining to what extent words close in semantic space are actually semantically related", and (ii) "analyzing, among related words, which type of relation they tend to instantiate". For these purposes, the same authors released the BLESS data set which is, to our knowledge, the most carefully designed and most comprehensive data set currently available that covers both these aspects.

The BLESS data consists of manually selected triples of target *concepts*, their *relata* and the particular semantic *relation* that holds between them. Consider Table 9.1 for

[3] Similar arguments can be brought up for the case of syntagmatic relations as well.

Concept	Relation	Relatum
alligator-n	COORD	crocodile-n
dishwasher-n	HYPER	appliance-n
snake-n	MERO	tongue-n
castle-n	EVENT	defend-v
coat-n	ATTRI	short-j
bottle-n	RANDOM-N	microphone-n
plum-n	RANDOM-V	tour-v
table-n	RANDOM-J	fast-j

Table 9.1: Examples of triples taken from BLESS data set (Baroni and Lenci, 2011)

	Cardinality		
Relation	**min.**	**avg.**	**max.**
COORD	6	17.1	35
HYPER	2	6.7	15
RANDOM-N	16	32.9	67

Table 9.2: Distribution of COORD, HYPER and RANDOM-N relations in the BLESS data set in terms of minimal, average and maximal number of relata per target word (taken from Baroni and Lenci (2011))

instructive examples. As can be seen from the table, BLESS features six types of relations in total[4]. Overall, the data set comprises 200 unique target concepts[5] associated to 5676 unique relata in 26654 different relations. For additional information concerning further statistics or details of the construction of the resource, the reader is referred to Baroni and Lenci (2011).

For the purpose of using BLESS as a benchmark for evaluating auxiliary distributional models with respect to their capacity of acquiring attribute-preserving relata (in the terminology of BLESS), we are mainly interested in those triples involving COORD and HYPER relations, assuming that concepts being linked by one of these relations are most likely to share similar attribute profiles. For comparison, we consider the RANDOM-N relation as well. Relevant statistics about the distribution of these types of

[4] As for the relation types, COORD, HYPER and MERO are relations between nouns, denoting co-hyponymy, hypernymy and meronymy, respectively. EVENT denotes a relation between a noun and a typical event the concept is involved in or affected by. ATTRI stands for a relation between a noun and an adjective denoting one of its attributes (without any further specification of the attribute itself). Finally, RANDOM-* is included in order to control for nouns, verbs and adjectives that are *semantically unrelated* to the target. (Baroni and Lenci, 2011)

[5] Baroni and Lenci (2011) point out that their focus was on semantically concrete concepts when they compiled the data set.

relations in the BLESS data are reported in Table 9.2.

Procedure. Our benchmarking methodology is as follows:

1. For each target concept in BLESS, we use the vector representations in V_{aux} to predict its nearest neighbours at ranks $k = 1$, $k = 3$ and $k = 5$, based on computing the cosine vector similarity (as introduced in Chapter 2) over all pairs of nouns contained in V_{aux}.

2. For each of these neighbours, we look up the semantic relation that holds between the target and the neighbour according to BLESS.

3. If this relation is one of COORD or HYPER, the prediction counts as `positive`; if the relation is RANDOM-N (or any other BLESS relation), the prediction counts as `negative`. If the predicted neighbour is not covered by BLESS, we check whether it is a synonym according to WordNet[6]. In this case, the prediction counts as `positive` as well, otherwise as `unknown`.

4. V_{aux} is assigned overall precision-at-rank scores[7] (P@k) computed as the fraction of `positive` predictions among all (i.e., `positive` + `negative` + `unknown`) predictions (Manning et al., 2008).

5. In order to assess the quality of nearest neighbours that are unknown to BLESS, we consult the WordNet taxonomy as an additional resource, computing the average *inverse path distance* (IPD)[8] between the target and the nearest neighbour over all `unknown` and (for comparison) `positive` cases.

9.2.2 Benchmark Results

First-order models. We constructed various instantiations of first-order distributional models along contextual paths[9] inspired from previous work (Padó and Lapata, 2007;

[6] For this purpose, we look up all synsets of which the target noun is a member, irrespective of any sense distinctions. Every noun in each of these synsets is considered as a synonym of the target noun.

[7] Lenci and Benotto (2012) use *Average Precision* (AP; Kotlerman et al. (2010)) as a metric for evaluating ranked neighbour lists reflecting the BLESS semantic relations. We do not adhere to this metric here, as deviations from the ideal ranking are *averaged* over all ranks by AP. For assessing the prospects of an auxiliary model in distributional enrichment, however, we are mainly interested in the performance on the first k ranks, which is why we consider precision-at-rank more meaningful for our purposes.

[8] Intuitively, the shorter a path linking two nodes in the taxonomy, the higher their semantic relatedness. (Budanitsky and Hirst, 2006) IPD accounts for this intuition by computing the inverse of the *shortest* of all possible paths between two lemmas, taking all their synsets into account. Thus, IPD scores range from 0 to 1, with larger values indicating a higher degree of relatedness. An IPD score of 1 is awarded only if two words are members of the same synset. For computing IPD, we rely on the `Wordnet::Similarity` package (Pedersen et al., 2004).

[9] The cooccurrences instantiating the dependency paths were extracted from the pukWaC corpus (Baroni et al., 2009).

Parameter	Setting
Num. Dimensions	2000
Component Weighting Function	PosPMI (Niwa and Nitta, 1994)
Stopword Filtering	active
Lemmatization	active
Frequency Thresholds	$\theta_{bow} \geq 10$; $\theta_{dep} \geq 1$
BOW context window	5 words left, 5 words right
Part-of-Speech Tags	Penn Treebank, coarse-grained

Table 9.3: Best parameter settings in first-order auxiliary models as determined in BLESS benchmark; part-of-speech tags from the Penn Treebank (Marcus et al., 1993) are mapped to their base class (i.e., all different noun tags to `N`, all verb tags to `V`, etc.); frequency thresholds θ_{bow} and θ_{dep} are applied to relations extracted from bag-of-words or dependency paths, respectively.

	P@1	P@3	P@5	Avg. IPD (`unknown`)	Avg. IPD (`positive`)
`N:COORD:N`	0.38	**0.34**	**0.30**	**0.15**	0.28
`N:SBJ:V:OBJ1:N`	0.03	0.01	0.01	0.08	0.27
`N:SBJ:V`	0.02	0.01	0.01	0.08	0.31
`N:SBJ:M:VC1:V`	0.02	0.02	0.01	0.08	0.27
`N:OBJ:V`	0.09	0.08	0.06	0.08	0.29
`N:OBJ:M:VC1:V`	0.00	0.00	0.00	0.10	0.23
`N:SBJ:V:PRD1:J`	0.02	0.01	0.01	0.07	0.15
`N:SBJ:M:VC1:V:PRD1:J`	0.02	0.01	0.01	0.09	0.16
`N:SBJ:V:IOBJ1:N`	0.00	0.00	0.00	0.09	0.26
`N:SBJ:M:VC1:V:IOBJ1:N`	0.01	0.01	0.01	0.11	0.23
`N:NMOD1:I:PMOD1:N`	0.13	0.07	0.05	0.08	**0.36**
`N:NMOD1:N`	0.16	0.09	0.08	0.06	0.35
`N:SBJ:V:PRD1:N`	0.12	0.10	0.07	0.10	0.23
Top-4	0.36	0.27	0.22	0.06	0.34
Top-2	0.33	0.26	0.22	0.09	0.32
all	**0.40**	0.30	0.24	0.09	0.30
BOW	0.32	0.25	0.21	0.11	0.35

Table 9.4: Benchmark results for best instantiations of first-order auxiliary models (cf. Table 9.3) on BLESS data

Rothenhäusler and Schütze, 2009) and subjected them to the benchmark methodology described above. The best overall parameterization is given in Table 9.3. The performance values corresponding to these settings are displayed in Table 9.4 which is structured as follows: The area above the first horizontal line contains the performance of each contextual path (as given in the first column[10]) when being used to construct an individual model. The area below this line contains three configurations with in which *all*, the *top-2* and the *top-4* contextual paths (according to their individual precision) are collapsed into one model. For comparison, the performance of a model constructed along bag-of-words contexts[11] is given below the second horizontal line.

As can be seen from the table, the best precision at $k = 1$ ($P@1 = 0.40$) is obtained from the configuration that combines all individual dependency paths into one model. Due to its linguistically more principled way of extracting contextual material, this model clearly outperforms the BOW configuration ($P@1 = 0.32$). Breaking down the overall score into the contribution of individual paths, we find that overall performance is largely determined by the `N:COORD:N` relation. Given that this relation can be considered as the syntagmatic manifestation of a paradigmatic functional relation, this result is not surprising. Moreover, the predictions of `N:COORD:N` are most stable on subsequent ranks, as becomes evident from a comparative inspection of the P@3 and P@5 scores.

Interestingly, there seems to be no complementarity among the best-performing individual paths, as can be seen from the fact that the combinations of `N:COORD:N` with `N:NMOD1:N` (i.e., the *Top-2* configuration) and with `N:NMOD1:N`, `N:NMOD1:I:PMOD1:N` and `N:SBJ:V:PRD1:N` (*Top-4*), respectively, yield a performance inferior to `N:COORD:N` on its own. This effect is also confirmed by investigating IPD for predicted neighbours unknown to BLESS: Average IPD amounts to 0.15 for `N:COORD:N` in these cases, which is clearly the best result across all configurations (including the combined models).

In the face of these results, the capacity of first-order models to be used as auxiliary models in distributional enrichment of C-LDA vector representations must be considered questionable. Even though there might be some potential in neighbours classified as unknown by BLESS, an overall chance of less than 50% for acquiring an attribute-preserving auxiliary vector obviously poses a risk for distributional enrichment.

[10] Our notation for specifying dependency paths is as follows: Each path begins with the part-of-speech tag of the target word and ends with the part-of-speech tag of the context word. In between, there is either exactly one dependency edge label (in case of a path containing only one edge) as generated by the Malt parser (Nivre et al., 2007) or an alternating sequence of edge labels and part-of-speech tags of intermediary nodes (in case of longer paths). Edge labels point from the syntactic dependent to the head by default, inverse edges pointing to the dependent are marked by `1` (as in `SBJ1`, for instance). All elements within a path description are separated by a colon.

[11] The window used to extract context words for a target element in this setting comprises five words to the left and five words to the right, which is a common setting in the literature (cf. Mitchell and Lapata (2010), among others). For full comparability, all other parameters are in concordance with the settings used for the dependency-based models as given above.

Parameter	Setting
Num. Dimensions	2000
Component Weighting Function	PosPMI (Niwa and Nitta, 1994)
Stopword Filtering	active
Lemmatization	active
Frequency Thresholds	*bow+bow*: $\theta_{bow} \leq 10$ *dep+dep*: $\theta_{dep} \geq 1$ *combined*: $\theta_{bow} \geq 5$; $\theta_{dep} \geq 1$
Part-of-Speech Tags	Penn Treebank, coarse-grained

Table 9.5: Best parameter settings in second-order auxiliary models as determined in BLESS benchmark; part-of-speech tags from the Penn Treebank (Marcus et al., 1993) are mapped to their base class (i.e., all different noun tags to `N`, all verb tags to `V`, etc.); frequency thresholds θ_{bow} and θ_{dep} are applied to relations extracted from bag-of-words or dependency paths, respectively.

	P@1	P@3	P@5	Avg. IPD (`unknown`)	Avg. IPD (`positive`)
`N:OBJ:V/V:OBJ1:N`	0.30	0.23	0.20	0.12	0.27
`N:SBJ:V/V:SBJ1:N`	0.07	0.05	0.04	0.09	0.23
`N:COORD:N/N:COORD1:N`	0.52	0.48	**0.46**	**0.16**	0.30
`N:SBJ:V:OBJ1:N/N:OBJ:V:SBJ1:N`	0.07	0.07	0.05	0.12	**0.47**
`N:NMOD:N/N:NMOD1:N`	0.15	0.13	0.12	0.13	0.34
`N:NMOD1:I:PMOD1:N/N:PMOD:I:NMOD:N`	0.13	0.09	0.08	0.11	0.29
`N:SBJ:V:PRD1:N/N:PRD:V:SBJ1`	0.20	0.02	0.02	0.15	0.36
all-inv	0.29	0.23	0.19	0.12	0.29
sel-inv	0.37	0.27	0.21	0.13	0.27
`N:BOW5:N`/*sel-dep*	0.19	0.15	0.15	0.12	0.29
sel-dep/`N:BOW5:N`	0.33	0.26	0.20	0.13	0.28
`N:BOW5:N/N:COORD:N`	0.24	0.22	0.19	0.12	0.34
`N:COORD:N/N:BOW5:N`	**0.55**	**0.50**	0.45	**0.16**	0.29
`N:BOW5:N/N:BOW5:N`	0.20	0.14	0.11	0.12	0.31

Table 9.6: Benchmark results for best instantiations of second-order auxiliary models (cf. Table 9.5) on BLESS data

Second-order models. As defined in Section 2.3.4, second-order distributional models are characterized by target and context words being linked along two contextual paths. Each path can be specified in terms of a syntactic dependency or a bag-of-words relation.[12] In a search for the best instantiation of second-order models for distributional enrichment, we explored three types of configurations: (i) models constructed from two inverse dependency paths[13] (*dep+dep*), (ii) models being based on two bag-of-words relations (*bow+bow*), and (iii) *combined* models being constructed from one dependency and one bag-of-words path. The best parameterizations for these models are given in Table 9.5.

Table 9.6 shows performance figures as achieved by benchmarking various instantiations of second-order models based on these parameterizations against the BLESS data. The table is sub-divided into three sections: The upper part, above the first horizontal line, contains purely dependency-based models. The setting *all-inv* refers to the combination of *all* first-order paths from Table 9.4 being extended to yield a second-order path by their individual inverse. The setting *sel-inv* refers to the combination of all second-order paths listed explicitly in the upper part of Table 9.6. The middle part of the table contains a selection of combined models, where *sel-dep* refers to the set of first-order paths taking part in the previously mentioned *sel-inv* setting. The bottom part, below the second horizontal line, contains one *BOW* model[14] to be considered as a baseline for comparison.

Interpreting these results, we observe moderate individual performance of most inverse second-order dependency paths. `N:COORD:N/N:COORD1:N` is a notable exception, as it yields the best individual performance by far ($P@1 = 0.52$). As a major advantage, this model performs relatively robust at $k = 3$ and $k = 5$ as well, contrary to the other inverse models. Similarly to first-order dependency models (cf. Table 9.4 on page 178), combinations of individual inverse second-order models (*all-inv*, *sel-inv*) are detrimental: They clearly outperform the second-order *BOW* baseline, but do not meet the performance of the best individual model.

As for the combined models, the results show that combinations using a dependency path in the first and a *BOW* path in the second step are considerably more effective than vice versa. The overall best result is achieved by the combination of `N:COORD:N/N:BOW5:N` ($P@1 = 0.55$). Apparently, the best compromise in second-order

[12] We formalize second-order contextual relations using the notation `<path1>/<path2>`, where `<path1>` and `<path2>` can be described as given in Footnote 10.

[13] Dependency paths can be *inverted* by making each of its constitutive arcs point from the dependent to the head. For instance, the inverse of the path `N:SBJ:V` is `V:SBJ1:N`, the inverse of `N:SBJ:V:PRD1:J` is `J:PRD:V:SBJ1:N`. Combining a dependency path and its inverse path results in a second-order path that can be considered as linking a target to a context word sharing similar properties, which is highly desirable for the task of detecting attribute-preserving neighbours. For example, all context words extracted for a target word w along the second-order path `N:OBJ:V/V:OBJ1:N` are characterized by occurring as syntactic objects of the same head nouns (cf. Thater et al., 2010).

[14] Throughout the entire discussion, `BOW5` refers to contextual paths extracted from a bag-of-words context window of five words to the left and five words to the right of the target word.

	P@1 (BLESS)	P@1 (manual, strict)	P@1 (manual, lenient)
1st-order: *all*	0.39	0.51	0.54
2nd-order: N:COORD:N/N:BOW5:N	0.54	0.73	0.79

Table 9.7: Results after manual assessment of nearest neighbours predicted by best first-order and second-order model

distributional modeling can be achieved from being as restrictive as necessary in the first step and as permissive as possible in the second.

Compared to their respective first-order counterparts, most of the inverse second-order paths yield relative gains in their individual performance[15], at small margins, though. However, N:COORD:N/N:COORD1:N is the only dependency-based second-order setting outperforming the first-order *BOW* model ($P@1 = 0.52$ vs. $P@1 = 0.32$; cf. Table 9.4). This clearly shows that second-order distributional models are not prima facie superior to plain first-order *BOW* models, if particular semantic relations are in focus (hypernyms and co-hyponyms, in our case). With careful, linguistically principled context selection, however, their larger potential becomes clearly visible.

Discussion. The generally high quality of the BLESS data set notwithstanding, all benchmark results reported above have to be considered in light of the large proportions of nearest neighbour predictions that could not be evaluated due to coverage issues of the gold standard. In case of the best-performing second-order model, for instance, the proportion of unknown nearest neighbours amounts to more than 42%. As a first attempt to estimate the quality of these predictions, we computed path distance metrics for unknown predictions in all models. These scores, as reported in Tables 9.4 and 9.6, can be seen as rough estimates of the quality of unknown predictions. Given the general limitations of path-based distance metrics computed from the WordNet taxonomy[16], they are certainly not fully accurate.

In order to obtain a more solid estimation, we manually evaluated the unknown nearest neighbour predictions of the best-performing first-order (configuration *all*; cf. Table 9.4) and second-order model (N:COORD:N/N:BOW5:N; cf. Table 9.6) with respect to whether target words and their nearest neighbours according to the auxiliary model can be expected to exhibit similar attribute profiles. This criterion was operationalized in two settings: In the *strict* setting, only members of the same semantic category

[15] Except for N:NMOD:N/N:NMOD1:N.

[16] Most importantly, WordNet has been found to exhibit varying granularity across different areas of the taxonomy (Leacock and Chodorow, 1998) (e.g., the shortest path between *rice* and *potato* comprises three nodes, while the one linking *freezer* and *microwave* has length 7), and concepts whose semantic relatedness is due to aspects different from hyponymy are often widely dispersed across the taxonomy (e.g., *tennis*, *ball* and *net*). The latter issue is known as the "tennis problem" (Fellbaum, 1998).

were accepted (i.e., co-hyponyms and hypernyms), assuming that their attribute profiles should be characterized by *significant overlap* (e.g., *cauliflower–leek*, *giraffe–rhino*). In the *lenient* setting, only *partial overlap* in the attribute profiles was required, i.e., nearest neighbours were also accepted if their category membership differs from the target, while they still share some aspects of meaning that indicate agreement of individual attributes. This situation may be caused by concordant *telic roles* (Pustejovsky, 1995) as in *spinach–sauce*, for instance.

The results of this study (in terms of P@1) are shown in Table 9.7. Both models show a considerable boost in performance relative to the BLESS setting, which indicates that a high proportion of nearest neighbours not covered by the BLESS gold standard is actually attribute-preserving. Given the difficulty of the task – each nearest neighbour is picked from a total set of 5676 candidates –, we consider the overall precision of the `N:COORD:N/N:BOW5:N` model ($P@1 = 0.79$) very satisfactory. Among the remaining errors of this model, there is a high proportion of "near misses" due to topical relationships between the target and the predicted nearest neighbour (e.g., *pub–supermarket*, *fighter–battle*).

Conclusions. From these benchmark results, we draw the following conclusions: First, distributional enrichment of attribute-based vector representations should capitalize on second-order auxiliary models built from a combination of contextual paths that incorporate syntactic coordinates in the first and their syntagmatically related nouns in the second step (`N:COORD:N/N:BOW5:N`). Second, first-order models are not suitable for detecting attribute-preserving nearest neighbours, given their low performance in our tailored benchmarks. Nevertheless, they may still be useful for distributional enrichment in a purely syntagmatic setting where the context words being extracted from contextual paths are directly taken as (syntagmatic) neighbours at face value, rather than being used as features in paradigmatic nearest neighbour acquisition.

9.3 Distributional Enrichment for Attribute Selection

Overview. In this section, we apply distributional enrichment to structured distributional attribute models. Following a formal definition of the framework, we devise concrete instantiations to be used for enhancing topic-based attribute models. Our goal is to test whether distributional enrichment is capable of improving large-scale attribute selection performance, thus minimizing the gap between C-LDA and semi-supervised approaches (cf. Section 7.3.5 on page 122).

Formal definition of the framework. Let w be a target word, $\vec{w_{attr}}$ its vector representation in the structured distributional model V_{attr} spanning attributes A as dimensions, and $\vec{w_{aux}}$ its vector representation in the auxiliary model V_{aux} with context words C as dimensions. We define the following parameters:

- **structured distributional model V_{attr}:**

$$V_{attr} = \left\{ \vec{w_{attr}} \;\middle|\; \vec{w_{attr}} = \sum_{a \in A} \omega(w, a) \cdot \vec{e_a} \right\}, \forall w \in W_{attr} \tag{9.1}$$

- **auxiliary model V_{aux}:**

$$V_{aux} = \left\{ \vec{w_{aux}} \;\middle|\; \vec{w_{aux}} = \sum_{c \in C} \omega(w, c) \cdot \vec{e_c} \right\}, \forall w \in W_{aux} \tag{9.2}$$

- **targets of enrichment I:**

$$I \subseteq W_{attr} \tag{9.3}$$

  defines whether distributional enrichment is applied to all targets in W_{attr} or only a subset of them.

- **target retrieval function η:**

$$\eta : I \to V_{aux} \tag{9.4}$$

  is used to retrieve the vector representation $\vec{i_{aux}}$ from V_{aux} for a particular target of enrichment $i \in I$, based on string identity.

- **carrier elements J:**

$$J \subseteq \begin{cases} W_{aux} & \text{(in paradigmatic setting)} \\ C_{aux} & \text{(in syntagmatic setting)} \end{cases} \tag{9.5}$$

  Being represented in both V_{aux} and V_{attr}, carrier elements function as the bridge between structured and auxiliary models. Depending on the concrete instantiation of distributional enrichment, J contains adjectives or nouns. An ideal carrier $j \in J$ provides a strong, attribute-preserving relation to the target of enrichment i within V_{aux}. Thus, j can be used to *carry* complementary semantic information from V_{aux} to V_{attr} in order to enrich the structured representation of i, $\vec{i_{attr}}$. The vector representing j in the structured model, $\vec{j_{attr}}$, can be retrieved by the *carrier projection function* defined below.

- **carrier retrieval function θ:**

$$\theta : J \to V_{attr} \tag{9.6}$$

  is used to retrieve the structured vector representation $\vec{j_{attr}}$ from V_{attr} for a particular carrier element $j \in J$, based on string identity.

- **relatedness function r:**

$$r : I \times J \rightarrow \mathbb{R} \tag{9.7}$$

determines the strength of relatedness between a target of enrichment $i \in I$ and a carrier $j \in J$ in V_{aux}. Depending on a paradigmatic or syntagmatic setting, r can either be based on paradigmatic similarity or a score measuring the strength of the syntagmatic relation between i and j.[17]

- **ordered set of carriers C for a target i:**

$$C(i) = \{\langle j_k, \lambda_k \rangle\} \text{ with } 1 < k < |J|,$$

where $\lambda_k = r(i, j_k)$ and a partial ordering is imposed on the pairs $\langle j_k, \lambda_k \rangle \in C(i)$ such that

$$\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_k. \tag{9.8}$$

Thus, $C(i)$ arranges the carriers for each target i in descending order according to their strength of relatedness. In the following, we will use the notation $C_k(i)$ in order to refer to the pair in $C(i)$ at position k.

- **carrier projection function μ:**

$$\mu : J \times \mathbb{R} \rightarrow V_{attr}$$

is used to map a pair from $C(i)$, consisting of a carrier element and an associated weight, to a structured vector representation $\vec{j_{attr}}$ in V_{attr}. μ can be instantiated in a *weighted* and an *unweighted* setting, such that

$$\vec{j_{attr}} = \begin{cases} \mu_{weighted}(j_k, \lambda_k) = \lambda_k \cdot \theta(j_k) \\ \mu_{unweighted}(j_k, \lambda_k) = \theta(j_k) \end{cases} \tag{9.9}$$

Here, θ is used to map the carrier element j_k back into the structured attribute space. The retrieved vector may be weighted according to its relatedness to j_k, using λ_k as a scaling factor.

- **objective function** *predict-quality*:

$$\textit{predict-quality} : V_{attr} \rightarrow \mathbb{R} \tag{9.10}$$

This function predicts the estimated quality of an attribute-based phrase representation based on different regressor variables estimated during linear regression of phrase vector quality (cf. Section 8.4).

[17] Accordingly, we will instantiate r with cosine similarity and PosPMI scores, respectively, in different model instantiations.

Distributional Enrichment as Centroid Construction:

$$\begin{aligned}\vec{w'_{attr}}(k) :=& \vec{w_{attr}} \oplus \sum_{k=1}^{K} \mu(C_k(w)) \\ =& \vec{w_{attr}} \oplus \left(\lambda_1 \cdot \vec{j_{1_{attr}}} \oplus \ldots \oplus \lambda_K \cdot \vec{j_{K_{attr}}}\right)\end{aligned}$$

for the smallest number $1 \leq K \leq |J|$ such that

$$\text{predict-quality}(\vec{w'_{attr}}(k)) \leq \text{predict-quality}(\vec{w'_{attr}}(k-1)),$$

where initially

$$\text{predict-quality}(\vec{w'_{attr}}(0)) = \text{predict-quality}(\vec{w_{attr}}).$$

Figure 9.2: Formal statement of Distributional Enrichment; $\vec{w'_{attr}}(k)$ denotes the centroid replacing the original structured vector $\vec{w_{attr}}$ in the k-th iteration of the procedure.

In distributional enrichment, *predict-quality* serves the purpose of an objective function, i.e., to assess whether a particular noun vector $\vec{w_{attr}} \in V_{attr}$ can be expected to contribute to an enhanced attribute-based phrase representation when being added to the centroid. Based on the regression functions for maximizing phrase vector quality as given in Equations (8.14)–(8.16) on page 169, we explore three different instantiations of *predict-quality* for the task of improving structured noun representations in a distributional attribute model:

$$\textit{predict-quality}_{oracle} = \texttt{AttrRankComp} \tag{9.11}$$

$$\begin{aligned}\textit{predict-quality}_{no-ranks} =& \; 0.71\,\texttt{AbstractnessNoun} + 0.18\,\texttt{PhraseEntropy} \\ & - 0.92\,\texttt{MeasurableAttr} - \; 0.14\,\texttt{AttrPseudoDocsFreq}\end{aligned} \tag{9.12}$$

$$\textit{predict-quality}_{minimal} = \texttt{PhraseEntropy} \tag{9.13}$$

Smaller values of *predict-quality* correspond to an expected higher quality of $\vec{w_{attr}}$. Note that all regressor variables in the original regression functions that are independent of the noun vector to be enhanced (e.g., factors related to adjective vectors) are omitted in these instantiations, as they can be regarded as constants for the purposes of *predict-quality*.

1: **function** UPDATE($V_{attr}, V_{aux}, I, J, \eta, \theta, C, \mu, r$, predict-quality)
2: **for all** $i \in I$ **do**
3: $\vec{centroid} \leftarrow \vec{w_{i_{attr}}}$
4: *vectorList* $\leftarrow$ empty list
5: *vectorList*.append($\vec{w_{i_{attr}}}$)
6: *objValue* $\leftarrow$ predict-quality($\vec{w_{i_{attr}}}$)
7: **for** $1 \leq k \leq |J|$ **do**
8: *vectorList*.append($\mu(C_k(i))$)
9: $\vec{tmp} \leftarrow$ buildCentroid(*vectorList*)
10: *currentObjValue* $\leftarrow$ predict-quality($\vec{tmp}$)
11: **if** *currentObjValue* $\leq$ *objValue* **then**
12: $\vec{centroid} \leftarrow \vec{tmp}$
13: *objValue* $\leftarrow$ *currentObjValue*
14: **else**
15: **break**
16: **end if**
17: **end for**
18: $\vec{w_{i_{attr}}}' \leftarrow \vec{centroid}$
19: **end for**
20: **end function**

Figure 9.3: Algorithm for Iterative Vector Updating

Formal statement. Based on these parameters, the goal of distributional enrichment of V_{attr} is to replace $\vec{w_{attr}}$ by a centroid $\vec{w'_{attr}}$ which is constructed from the top-K carriers $\langle j_k, \lambda_k \rangle \in C(w)$, as displayed in Fig. 9.2.

Iterative update algorithm. Thus, distributional enrichment can be seen as a minimization problem which we solve by means of an iterative *vector update algorithm*. Apart from different assignments of the parameters introduced above (which will be discussed in Sections 9.3.1 and 9.3.2 below), the core procedure is generic across all instantiations. It is implemented as given in Fig. 9.3:

- For each target of enrichment $i \in I$, the procedure yields an attribute-based vector representation $\vec{w_{i_{attr}}}'$ which is set to the centroid of all carrier vectors satisfying the objective criterion (cf. line 18). Initially, $\vec{centroid}$ is set to $\vec{w_{i_{attr}}}$ (cf. line 3). Thus, in the special case that no carrier can be found that satisfies the objective criterion, $\vec{w_{i_{attr}}}'$ equals $\vec{w_{i_{attr}}}$, i.e., no update is performed.

- Carrier vectors that are candidates to become members of the centroid are collected in a list *vectorList* which initially contains $\vec{w_{i_{attr}}}$ as its only element (cf. lines 4 and 5). In each step of the inner loop (lines 7–17), the kth-best carrier vector from $C(i)$ is added to *vectorList* (cf. line 8).

- The objective criterion is assessed in a "looking-ahead" approach (cf. lines 9–16):

An intermediate centroid is computed from the current elements of *vectorList* and stored in $\vec{tmp}$. By applying the objective function to $\vec{tmp}$ (cf. line 10), the relative quality of the intermediate centroid can be evaluated: An equal or lower score of *currentObjValue* compared to *objValue* (which stores the objective value previously determined from $\vec{centroid}$) indicates an expected improvement of $\vec{tmp}$ over the previous centroid. In this case, $\vec{centroid}$ is set to $\vec{tmp}$ (cf. line 12) and the inner loop continues with the next carrier at rank $k + 1$. Otherwise, the previous centroid remains unchanged and the procedure is aborted[18] (cf. line 15).

Note that this procedure licenses either (i) to *keep* the original structured vector or (ii) to *replace* it by a centroid that may be constructed from an arbitrary number of carrier vectors, depending on the lower bound of semantic relatedness between carriers and targets of enrichment that is introduced by superimposing the objective function onto the iteration cycle over ordered carriers. In the latter case, the resulting centroid is assumed to capture more attribute-specific information than $\vec{w_{attr}}$ in the sense that the contribution of semantically related carriers will promote coherent sets of attributes in the structured vector, thus increasing their chances in attribute selection.

Unsupervised nature of the framework. It is important to emphasize, however, that information about the correct attribute(s) is generally not available to the objective function at prediction time. Therefore, there is no guarantee that attributes being promoted by distributional enrichment result in overall adequacy of the distributional profile in an updated semantic representation. In other words, the enrichment framework proposed here is *unsupervised* in the sense that it does not include any means in order to control for the *correct* attributes being promoted *directly*. The only means to facilitate this desired effect is to instantiate the carrier elements in such a way that they exhibit an attribute-preserving semantic relation to the respective target of enrichment.

Key aspects of distributional enrichment. In the following, we discuss several instantiations of the framework. These instantiations will be compared with regard to the following *key aspects of distributional enrichment* (cf. Table 9.8) which we consider the major determinants for enhancing structured vector representations by distributional enrichment:

- **functional relationship between target and carrier:** This basically determines whether the carrier elements stand in a syntagmatic or paradigmatic relation to the target words.
- **semantic relation between target and carrier:** What are the (syntactic) correlates that facilitate attribute-preserving semantic relations between target and carrier?

[18] Recall from the definiton of $C(i)$ that subsequent carriers are guaranteed to be less similar to i, which is why they can be ignored.

Variant	Functional Relationship Target – Carrier	Semantic Relation Target – Carrier	Complementarity
ParaDisE	paradigmatic	NN:COORD:NN/ NN:BOW5:NN	low
ParaDisE-Adj	paradigmatic	JJ:PRD:VB:SBJ1:NN/ NN:SBJ:VB:PRD1:JJ	low
SynDis-Co	syntagmatic	NN:COORD:NN	low
SynDis-Mo	syntagmatic	NN:SBJ:VB:PRD1:JJ	high

Table 9.8: *Key aspects of distributional enrichment* in different instantiations of the framework; first column contains an abbreviation used to refer to the respective model throughout the text

- **complementarity:** To what extent can the semantic information that is used to construct the centroid vector be considered as complementary in the sense that it extends the model with *additional* information beyond the one initially contained in V_{attr}?

9.3.1 Paradigmatic Distributional Enrichment

Paradigamtic distributional enrichment may be used to enhance attribute-based vector representations of nouns or adjectives, as detailed in the following.

Paradigmatic Distributional Enrichment of Noun Vectors (ParaDisE-Noun)

In this setting, *paradigmatic neighbours* are used for iteratively updating attribute-based noun vectors. To this end, the algorithm is initialized with all nouns from V_{attr} as targets of enrichment and all target words from V_{aux} as potential carriers:

$$I := \{w \in W(V_{attr}) | w \text{ is a noun}\} \qquad J := W(V_{aux})$$

The relatedness function r is set to return the strength of association between each target i and carrier j in terms of the cosine similarity of their vectors in the auxiliary model:

$$r(i,j) := cos(\vec{w_{i_{aux}}}, \vec{w_{j_{aux}}})$$

Thus, the ordered set of weighted carriers for target i, $C(i)$, encompasses the semantic neighbours of i in V_{aux} in descending order of their spatial proximity. As V_{aux} inherits all parameters from the second-order model that performed best in the BLESS benchmark (cf. Section 9.2.1), we expect that a large proportion of the nearest neighbours predicted for i exhibit an attribute-preserving semantic relation of either *co-hyponymy*, *hypernymy*,

or even *(near-)synonymy*, so that their most prominent attributes are likely to dominate the attribute profile of the resulting centroid.

Note, however, that ParaDisE-Noun is characterized by *low complementarity* of information sources: Given that the attribute information that enters the centroid is not externally acquired but merely pooled from previously existing noun vectors in V_{attr}, there remains a risk of acquiring attribute-preserving neighbours that are insufficiently represented in V_{attr} themselves. Depending on whether insufficent representations in V_{attr} turn out as an issue that is limited to singular vectors or affects entire semantic regions, this lack of complementarity might undermine the prospects of ParaDisE-Noun.

Paradigmatic Distributional Enrichment of Adjective Vectors (ParaDisE-Adj)

Apart from nouns, paradigmatic neighbours can also be used for enhancing attribute-based adjective representations. In such an instantiation, all adjectives from V_{attr} become the targets of enrichment:

$$I := \{w \in W(V_{attr}) | w \text{ is an adjective}\} \qquad J := W(V_{aux})$$

Analogously to ParaDisE-Noun, the relatedness function r yields the strength of association between each target i and carrier j in terms of their cosine similarity in V_{aux}:

$$r(i,j) := cos(\vec{w_{i_{aux}}}, \vec{w_{j_{aux}}})$$

Note, however, that V_{aux} must be constructed such that it contains adjectives as target words. We implement V_{aux} as a second-order adjective model. The relation between target and context words is defined by a combination of inverse dependency paths: `JJ:PRD:VB:SBJ1:NN/NN:SBJ:VB:PRD1:JJ`. Along these contextual paths, a target adjective is linked to all other adjectives that are observed as modifiers of the same nouns in a predicative syntactic construction. Thus, distributional descriptors of the meaning of property-denoting adjectives in this model are properties that coincide in the same concepts. For instance, (i) an *insurance* may be designated as being *expensive* or *costly* at different occurrences in a corpus, (ii) different instances of *persons* may be described as *young* or *old*, and (iii), apart from being *calm*, a *hotel* may also be qualified as *luxury* (coincident properties). Being used in a paradigmatic setting, we expect these descriptors to produce semantic neighbours that are highly attribute-preserving. Nevertheless, similarly to ParaDisE-Noun, paradigmatic enrichment of adjective vectors is also characterized by *low complementarity* of information sources, because the adjective centroid is accumulated from previously existing adjective vectors from V_{attr}.

9.3.2 Syntagmatic Distributional Enrichment

For comparison, we propose two further instantiations of distributional enrichment based on *syntagmatic relations* between targets of enrichment and carriers. The syntagmatic character of this relation has an important implication: While paradigmatically

related carriers are selected by taking the entire distribution over all dimensions of an auxiliary vector into account (via the cosine metric), syntagmatic relata are assumed to be meaningful by themselves due to their *in praesentia* status[19]. In that respect, syntagmatically acquired carriers might serve as an instrument in order to preserve attribute meaning without relying on the potentially error-prone detour via paradigmatic neighbours.

Distributional Enrichment of Noun Vectors based on Syntagmatic Coordinates (SynDis-Co)

Contrary to the ParaDisE approach described above, the set of carrier elements equals the *context words* from V_{aux} now, while the targets of enrichment encompass all nouns from V_{attr} again:

$$I := \{w \in W(V_{attr}) | w \text{ is a noun}\} \qquad J := C(V_{aux})$$

The relatedness function r is set to return the component value of $\vec{w_{i_{aux}}}$ that measures the strength of the relationship between i and the context word j in V_{aux}:

$$r(i,j) := \omega_{aux}(w_i, w_j)$$

In this instantiation, V_{aux} is framed as a first-order distributional model with dimensions being selected along the dependency path `N:COORD:N`[20]. Thus, all context words in V_{aux} are nouns that are syntagmatically related to a target of enrichment i as its syntactic coordinates. We readily utilize these nouns as carrier elements, expecting them to share a high degree of semantic properties with i and hence to be useful for substituting the insufficient attribute vector $\vec{i_{attr}}$ in V_{attr}.

Note that this instantiation of the update algorithm requires substantial overlap between $C(V_{aux})$ and $W(V_{attr})$, as only those nouns that are already represented in V_{attr} can serve as effective carriers. Therefore, low complementarity of information sources might still be an issue in this model.

Distributional Enrichment of Noun Vectors based on Syntagmatic Predicative Modifiers (SynDis-Mo)

This variant of the update procedure capitalizes on syntagmatic carrier selection as well. Analogously to SynDis-Co, J is set to the context words of V_{aux} and r inherits the component weights from the auxiliary vectors:

$$I := \{w \in W(V_{attr}) | w \text{ is a noun}\} \qquad J := C(V_{aux})$$
$$r(i,j) := \omega_{aux}(w_i, w_j)$$

[19] Provided that the particular syntagmatic relation has been carefully selected.

[20] All other settings are in concordance with the models performing best in the BLESS benchmark (cf. Section 9.2.1).

The major characteristics of V_{aux}, however, crucially differ from SynDis-Co. The cooccurrences of target and context words in this instantiation of V_{aux} are exclusively extracted along the first-order dependency path `N:SBJ:V:PRD1:J`. As a result, the context words of V_{aux} comprise only adjectives that are observed as *predicative modifiers* of the target nouns.[21]

Using adjectives as carrier elements for updating insufficient noun vectors can be motivated from the compositionality of (property-denoting) adjectives and nouns in language, given that attribute selection in adjective-noun phrases is an intersective process, i.e., the attributes that are highlighted in the compositional semantics of a phrase are a subset of the ones denoted by an adjective in isolation (cf. Pustejovsky, 1995). In our own analysis in Section 8.2, we have found a considerable proportion of adjective-noun phrases in which the adjective triggers *compositional gains* by promoting the correct attributes in the compositional phrase vector representation, which supports Pustejovsky's claim that adjectives can be used in language as selectors of properties in the deep lexical structure of a noun. Consequently, it should be possible to approximate the total range of the noun's possible attributes by aggregating a large number of selectors (in terms of a centroid over adjectives). This is the main idea underlying the design of the SynDiS-Mo model. Restricting the carrier elements to predicative adjectives only is motivated by our finding that predicative use is a robust indicator of attribute-denoting adjectives (cf. Table 5.11 on page 72).

Using adjectives as carriers has positive implications on the complementarity of this model: Contrary to the previously described update procedures via paradigmatic neighbours and syntagmatic nominal coordinates, it is the only instantiation that incorporates substantial *external* information in order to complement the noun vectors already present in V_{attr} beforehand. Given that adjective vectors tend to exhibit stronger peaks in their distribution over attributes, we avoid the risk of running into an "insufficiency regress" (i.e., constructing the centroid over vectors that are insufficient themselves) that is inherent in both ParaDisE and SynDis-Co.

9.3.3 Joint Distributional Enrichment of Adjective and Noun Vectors

In this instantiation of the framework, distributional enrichment is applied in order to enhance adjective and noun vector representations simultaneously. To this end, the ParaDisE-Noun, ParaDisE-Adj, SynDis-Co and SynDis-Mo models are combined in a joint model that essentially implements an iterative wrapper procedure around singular update steps of the indvidual models. At each step of the process, the joint model considers one update proposal of each individual model and selects the best one according to the objective function. In more detail, the joint model for distributional model is summarized in Fig. 9.4. In the following, we point out the most important aspects of the procedure:

[21] Again, all other settings are in concordance with the models performing best in the BLESS benchmark (cf. Section 9.2.1).

```
1: function JOINTDISTRIBUTIONALENRICHMENT(D, M, V_attr, predict-quality)
2:     for all phrases p ∈ D do
3:         objValue ← predict-quality(adj⃗_attr ⊙ noun⃗_attr)
4:         adjCentroid ← adj⃗
5:         nounCentroid ← noun⃗
6:         adjsInCentroid ← empty list
7:         adjsInCentroid.append(adj⃗)
8:         nounsInCentroid ← empty list
9:         nounsInCentroid.append(noun⃗)
10:        while true do
11:            adjLookAheadList ← adjsInCentroid
12:            nounLookAheadList ← nounsInCentroid
13:            for all model instances m ∈ M do
14:                cand ← m.getNextCarrier()
15:                if m is adj model then
16:                    adjLookAheadList.append(cand)
17:                    adjLookAheadCentroid ← computeCentroid(adjLookAheadList)
18:                    lookAheadResults ← predict-quality(adjLookAheadCentroid ⊙ nounCentroid)
19:                else
20:                    nounLookAheadList.append(cand)
21:                    nounLookAheadCentroid ← computeCentroid(nounLookAheadList)
22:                    lookAheadResults ← predict-quality(adjCentroid ⊙ nounLookAheadCentroid)
23:                end if
24:            end for
25:            bestProposal ← arg min(lookAheadResults)
26:            if bestProposal.getObjValue() ≤ objValue then
27:                bestCarrier ← bestProposal.getCarrier()
28:                if bestProposal.getModelInstance() is a noun model then
29:                    nounsInCentroid.append(bestLookAheadResult.getWordVector())
30:                    nounCentroid ← computeCentroid(nounsInCentroid)
31:                else
32:                    adjsInCentroid.append(bestLookAheadResult.getWordVector())
33:                    adjCentroid ← computeCentroid(adjsInCentroid)
34:                end if
35:                objValue ← bestProposal.getObjValue()
36:                bestProposal.getModelInstance().removeFromCarriers(bestCarrier)
37:            else
38:                break
39:            end if
40:        end while
41:        adj⃗_attr′ ← adjCentroid
42:        noun⃗_attr′ ← nounCentroid
43:    end for
44: end function
```

Figure 9.4: Algorithm for Joint Distributional Enrichment

- Parameters of the procedure (cf. line 1) are a collection of adjective-noun phrases D[22], a set of individual model instantiations M, a structured attribute model V_{attr} and an objective function *predict-quality* as defined in (9.10) on page 185.

- Two centroids, *adjCentroid* and *nounCentroid*, are constructed simultaneously (cf. lines 4–9), in order to replace the original structured vectors $\vec{adj}_{attr}$ and $\vec{noun}_{attr}$ in V_{attr} (cf. lines 41 and 42). If the objective function does not license any update throughout the procedure, the original structured vector ($\vec{adj}_{attr}$ or $\vec{noun}_{attr}$ in V_{attr}, respectively) remains unchanged effectively (cf. initializations in lines 4 and 5).

- At each step of the iterative update procedure (cf. lines 10–40), the individual model instances $m \in M$ are consulted in order to propose a carrier as a candidate to become a member of one of the centroids (cf. line 14). All proposals collected within one iteration (in the list variable *lookAheadResults*) are individually evaluated in a "look-ahead" manner, i.e., they are tentatively added to the centroid and evaluated by the objective function (cf. lines 15–23). Only the best-performing carrier that minimizes the objective value in the current iteration (stored in *bestCarrier*, cf. line 25) actually enters one of the centroids (cf. lines 30 and 33), provided that it also outperforms the current global objective value (cf. line 26).

- At the transition from one iteration step to the next, the singular *bestCarrier* is removed from the carrier list C_m of the respective enrichment model instance m (cf. line 36). All other carriers from instantiations $m' \in M$ (with $m' \neq m$) that were *not* selected as globally optimal are kept in their respective local carrier lists $C_{m'}$[23]. Thus, they are still available for selection in further steps and it is guaranteed that the entire space of carriers (across all model instantiations) can be explored.

- As long as at least one of the model instances in the current iteration yields a carrier that improves the objective criterion, the procedure continues with the next iteration; otherwise, it aborts immediately without any further update of the centroids (cf. line 38).

The parameterizations of individual model instances $m \in M$ entering the joint model follow the specifications given in Sections 9.3.1 and 9.3.2 above.

[22] Contrary to the individual enrichment models, the joint model is not applied to individual adjectives or nouns, but to adjective-noun phrases in the first place.

[23] Recall from definition (9.8) on page 185 that carrier lists within an indvidual instantiation m of enrichment models M are always arranged in decreasing order with respect to their expected degree of attribute preservance.

9.4 Experiment 6: Large-scale Attribute Selection after Distributional Enrichment

The effectiveness of distributional enrichment in large-scale attribute selection is evaluated by measuring the impact of applying the enrichment variants introduced above to topic-based attribute models.

9.4.1 Experimental Settings

In these experiments, V_{attr} is instantiated with a topic-based attribute model, i.e., either a C-LDA or an L-LDA attribute model in their best-performing configuration as determined in Experiment 5 (cf. Section 7.3). This implies using multiplicative vector composition and ESel as attribute selection function. Each instance of these attribute models comprises a total of 12,987 noun vectors and 3,984 adjective vectors as targets. Auxiliary spaces are constructed as given below. In all instantiations of V_{aux}, targets of enrichment are set to the nouns from V_{attr} which occur in phrase vectors with ranks greater than 1, and the carrier projection function μ is used in an *unweighted* manner.

ParaDisE-Noun. A ParaDisE model for noun enrichment is constructed according to the settings that achieved best performance throughout the second-order benchmarks in Section 9.2.2: 2000 dimensions of meaning populated along the second-order paths `N:COORD:N/N:BOW5:N`, using positive pointwise mutual information (PosPMI; Niwa and Nitta (1994)) as component weighting function, stopword filtering and lemmatization being active, frequency thresholds of $\theta_{dep} \geq 1$ for dependency relations and $\theta_{bow} > 5$ for bag-of-words relations, as well as a coarse-grained set of part-of-speech tags with all tags from the Penn Treebank (Marcus et al., 1993) being mapped to their base class (i.e., all different noun tags were mapped to `N`, all verb tags to `V`, etc.). For comparison, we test a strictly inverse second-order model along the paths `N:COORD:N/N:COORD1:N`, with all other settings being equal. The vocabulary represented in both these models comprises all 922 nouns from the HeiPLAS data set. Overlap between the target words in V_{attr} and the carrier elements in each of these instantiations of V_{aux} amounts to 92%.

ParaDisE-DM. For comparison, an additional paradigmatic auxiliary model is constructed from the *Distributional Memory* (DM; Baroni and Lenci, 2010) which is presumably the largest and most versatile distributional semantic resource currently openly available. We use *TypeDM*, an instantiation of DM that contains 30,693 first-order vector representations over more than 25,336 dimensions of meaning populated from a mixture of lexicalized and syntactic patterns. In this setting, lexical overlap between the target words in V_{attr} and the carriers selected from V_{aux} amounts to 100%. Vector components are weighted according to the degree of variation in the surface realizations of a pattern (Baroni and Lenci, 2010). For building an auxiliary space from DM,

Parameter	Setting
Num. Dimensions	2000
Component Weighting Function	PosPMI (Niwa and Nitta, 1994)
Stopword Filtering	active
Lemmatization	active
Frequency Threshold	$\theta_{dep} \geq 1$
Part-of-Speech Tags	Penn Treebank, coarse-grained
Contextual Paths	`JJ:PRD:VB:SBJ1:NN/NN:SBJ:VB:PRD1:JJ`

Table 9.9: Parameter settings of second-order auxiliary model used for paradigmatic enrichment of adjective vectors (ParaDisE-Adj)

we rely on pre-compiled lists of the 10 nearest neighbours for each TypeDM target as provided by the authors[24]. This model allows for an interesting comparison against our own *ParaDisE* instantiation as it (i) is purely based on first-order co-occurrences, (ii) it contains a much larger vocabulary (which enables exploration of a larger semantic space for nearest neighbour selection, on the one hand, but also renders this task more difficult, on the other), (iii) it is of a much higher dimensionality, and (iv) the dimensions are constructed in a different way, relying on a mixture of lexical and syntactic linguistic patterns. We are mainly interested in understanding how a second-order auxiliary model being tailored to attribute preservation as much as possible compares to the large-scale and much more general approach underlying TypeDM in a paradigmatic distributional enrichment scenario.

SynDis-Co. This model contains all nouns from the HeiPLAS data set as target elements, while carriers are selected along (i) the singular first-order dependency path `N:COORD:N` or (ii) the combination of all first-order paths as given in Table 9.4 on page 178 (which is equivalent to the best first-order setting in the BLESS benchmark). Again, the set of part-of-speech tags is reduced to a coarse-grained inventory, with stopword filtering and lemmatization being active. Co-occurrence counts are weighted by PosPMI and the 2000 nouns resulting in the highest PosPMI values over all targets are selected as context words (i.e., as carrier elements). In these instantiations, the target/carrier overlap between V_{attr} and V_{aux} amounts to 23% (all paths) and 97% (`N:COORD:N`).

SynDis-Mo. In this model, the carrier elements comprise only adjectives found as predicative modifiers of the target words in the HeiPLAS data set, being extracted from pukWaC along the dependency path `N:SBJ:V:PRD:J` (using coarse-grained part-of-speech tags). After stopword filtering and lemmatization, co-occurrence counts are weighted by PosPMI and the 2000 adjectives resulting in the highest PosPMI values

[24]The list has been constructed by Partha Pratim Talukdar and is freely available from `http://clic.cimec.unitn.it/dm/materials/ri.w-lw_nn_10.txt.gz`.

over all targets are selected as carrier elements. In this instantiation, the target/carrier overlap between V_{attr} and V_{aux} amounts to 55%.

ParaDisE-Adj. Motivated by our finding that attribute-based adjective vectors have a stronger impact on phrase vector quality than noun vectors have on average (cf. Table 8.7 on page 144), we apply distributional enrichment to adjective representations as well. We construct a second-order auxiliary model for adjectives according to the settings given in Table 9.9. The vocabulary being represented in this model comprises 4213 adjectives, the target/carrier overlap between V_{attr} and V_{aux} amounts to 91%.

EnJoiDis. A joint enrichment model is implemented as a combination of the ParaDisE-Noun, ParaDisE-Adj, SynDis-Co and SynDis-Mo models as detailed above, following the algorithm outlined in Fig. 9.4 on page 193. The EnJoiDis model is applied to all adjective-noun phrases in the HeiPLAS data.

Evaluation metrics. The impact of distributional enrichment on attribute selection is measured by comparing the performance of the original C-LDA models and the instantiations of distrubutional enrichment just described, in terms of precision, recall and F_1 score (computed as micro averages over all attributes considered).

9.4.2 Experimental Results

Experimental results on the HeiPLAS development set are summarized in Table 9.10 on the following page. Results of distributional enrichment of C-LDA attribute models are reported in the upper part of the table, results of distributional enrichment of L-LDA models in the lower part. In the last row of each part, the attribute selection performance achieved by the original structured attribute models before distributional enrichment (C-LDA and L-LDA, respectively) are repeated as baselines. Preceding rows compare the performance of the ParaDisE[25], SynDis-Co[26], SynDis-Mo, EnJoiDis, ParaDisE-Adj and ParaDisE-DM enrichment models, where *minimal*, *no-ranks* and *oracle* denote the objective function being used. The columns refer to precision, recall and F_1 scores; statistical significance over the baseline (as determined by iterative re-sampling (Yeh, 2000; Padó, 2006)) is given in the last column.

Our main interest in analyzing these results is on comparing (i) the effectiveness of distributional enrichment across C-LDA, L-LDA and the various attribute inventories, (ii) the individual performance of the different objective functions, and (iii) the impact of joint modelling in distributional enrichment compared to individual model instantiations.

[25] The setting based on `N:COORD:N/N:BOW5:N` consistently outperforms `N:COORD:N/N:COORD1:N`. Therefore, only the former is reported here and discussed in the following.

[26] The setting based on `N:COORD:N` consistently outperforms the one based on the conjunction of all first-order paths, which is why only the former is reported here and discussed in the following.

		all attrs.			property attrs.			measurable attrs.			selected attrs.		
		P	R	F	P	R	F	P	R	F	P	R	F
ParaDisE	minimal	0.08	0.05	**0.07**	0.18	0.16	0.17 *	**0.27**	0.22	0.24	0.42	0.41	0.41
	no-ranks	0.08	0.05	**0.07**	0.19	0.16	0.17 *	**0.27**	0.22	0.24	0.42	0.41	**0.42**
	oracle	0.09	0.06	0.07*	0.24	0.20	0.22	*0.31*	0.25	*0.27***	0.45	0.44	0.45*
SynDis-Co	minimal	0.08	0.05	0.06	0.19	0.17	0.18$^{\cdot}$	**0.27**	0.23	**0.25**	0.42	0.40	0.41
	no-ranks	0.08	0.05	0.06	0.19	0.17	0.18$^{\cdot}$	**0.27**	0.23	**0.25**	0.42	0.40	0.41
	oracle	*0.10*	*0.07*	*0.08***	0.22	0.20	0.21	0.29	0.24	0.26*	*0.46*	0.45	*0.46***
SynDiS-Mo	minimal	0.08	0.05	0.06	**0.28**	0.18	**0.22**	**0.27**	0.22	0.24	**0.43**	0.41	**0.42**
	no-ranks	0.08	0.05	0.06	**0.28**	0.18	**0.22**	**0.27**	0.22	0.24	**0.43**	0.41	**0.42**
	oracle	0.09	0.06	0.07*	*0.29*	0.19	*0.23*	0.28	0.23	0.25$^{\cdot}$	0.43	0.42	0.43
ParaDisE-Adj	minimal	0.08	0.05	**0.07**	0.19	**0.19**	0.19	0.25	0.24	**0.25**	0.39	0.42	0.40
	no-ranks	0.08	0.05	**0.07**	0.19	**0.19**	0.19	0.25	0.24	**0.25**	0.39	0.42	0.40
	oracle	0.09	0.06	0.07*	0.23	*0.21*	0.22	0.29	*0.26*	*0.27**	0.44	*0.46*	0.45*
EnJoiDis	minimal	**0.09**	**0.06**	**0.07***	0.21	**0.19**	0.20	0.26	**0.25**	**0.25**	0.40	**0.43**	0.41
	no-ranks	**0.09**	**0.06**	**0.07**$^{\cdot}$	0.21	**0.19**	0.20	0.26	**0.25**	**0.25**	0.41	**0.43**	**0.42**
	oracle	0.08	0.05	0.06	0.21	0.19	0.20	0.28	0.24	0.26	0.39	0.42	0.40
ParaDisE-DM	minimal	0.08	0.05	0.06	0.19	0.16	0.17$^{\cdot}$	0.26	0.21	0.23	0.4	0.41	0.40
	no-ranks	0.08	0.05	0.06	0.19	0.16	0.17$^{\cdot}$	0.26	0.21	0.23	0.41	0.40	0.41
	oracle	0.09	0.06	0.07**	0.24	0.20	0.22	0.26	0.21	0.23	0.45	0.45	0.45*
C-LDA		0.08	0.05	0.06	0.23	**0.19**	0.21	0.26	0.21	0.23	0.41	0.39	0.40
ParaDisE	minimal	0.13	0.02	0.03	0.19	0.06	0.09*	0.27	0.22	0.24	0.49	0.33	0.40
	no-ranks	0.15	0.02	0.03	0.22	0.04	0.07	0.27	0.22	0.24	0.48	0.33	0.39
	oracle	0.15	0.02	0.04	*0.31*	0.07	0.11**	0.28	0.23	0.25	0.44	0.33	0.38
SynDis-Co	minimal	0.10	0.03	0.04	0.22	0.05	0.09$^{\cdot}$	0.28	0.22	0.25	0.51	**0.35**	**0.42**
	no-ranks	0.09	0.02	0.03	0.19	0.04	0.06	**0.29**	0.23	0.25	0.50	**0.35**	0.41
	oracle	0.16	0.03	0.05*	0.29	0.05	0.09$^{\cdot}$	0.29	0.23	0.26	0.52	0.36	0.42
SynDis-Mo	minimal	0.12	**0.03**	**0.05***	**0.26**	**0.18**	**0.21*****	0.28	0.23	0.25	0.50	**0.35**	0.41
	no-ranks	0.12	**0.03**	**0.05***	**0.26**	**0.18**	**0.21*****	0.28	0.23	0.25	0.50	**0.35**	0.41
	oracle	*0.17*	0.03	*0.05**	*0.31*	*0.21*	*0.25***	0.28	*0.26*	0.27	0.52	0.35	0.42
ParaDisE-Adj	minimal	**0.17**	0.02	0.04	**0.26**	0.05	0.08	0.27	**0.25**	**0.26**	**0.52**	**0.35**	**0.42**
	no-ranks	**0.17**	0.02	0.04	0.24	0.04	0.07	0.27	**0.25**	**0.26**	**0.52**	**0.35**	**0.42**
	oracle	*0.17*	0.02	0.04	0.26	0.06	0.09$^{\cdot}$	*0.30*	0.25	*0.28**	0.52	0.36	*0.43*
EnJoiDis	minimal	0.15	**0.03**	0.04	0.22	0.04	0.07	0.27	**0.25**	**0.26**	0.48	**0.35**	0.40
	no-ranks	0.13	**0.03**	**0.05**	0.25	0.07	0.11**	0.28	**0.25**	**0.26**	0.47	**0.35**	0.40
	oracle	0.12	0.02	0.04	0.23	0.05	0.08*	0.29	0.25	0.27$^{\cdot}$	*0.53*	0.34	0.41
ParaDisE-DM	minimal	0.10	0.02	0.04	0.20	0.05	0.08	0.28	0.22	0.25	**0.52**	0.33	0.40
	no-ranks	0.14	0.02	0.03**	0.24	0.03	0.06	0.28	0.23	0.25	0.50	0.34	0.40
	oracle	0.12	0.02	0.03	0.26	0.05	0.08	0.29	0.25	0.27	0.50	*0.37*	*0.43*
L-LDA		0.15	0.02	0.03	**0.26**	0.03	0.06	0.28	0.22	0.25	0.51	0.33	0.40

Table 9.10: Attribute selection performance on HeiPLAS development set after distributional enrichment of attribute models based on C-LDA (upper part) and L-LDA (lower part) attribute models; significance over original C-LDA and L-LDA attribute models, respectively. Best results highlighted in boldface (unsupervised settings) and italics (oracle setting).

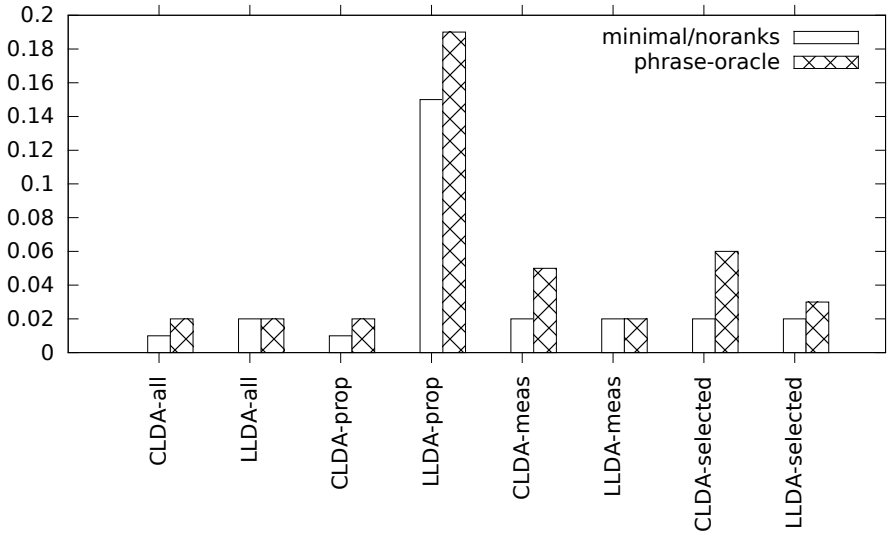


Figure 9.5: Largest improvements in F_1 score (y-axis) over original structured C-LDA and L-LDA models achieved by distributional enrichment in various inventories of attributes from HeiPLAS-Dev data (x-axis). Texture of bars indicates different objective functions used (cf. legend).

Overall effectiveness of distributional enrichment. Distributional enrichment is mildly effective in enhancing the large-scale attribute selection capabilities of both C-LDA and L-LDA models: Small, but significant improvements of $+0.01$ and $+0.02$ points in F_1 score over the original C-LDA and L-LDA models can be observed, leading to a maximum performance of $P = 0.09$, $R = 0.06$, $F_1 = 0.07$ for C-LDA and $P = 0.17$, $R = 0.03$, $F_1 = 0.05$ for L-LDA. In smaller subsets of attributes, however, the improvement rate tends to increase, up to $+0.19$ points in F_1 score relative to the original L-LDA model on the *property* attribute inventory. On this subset of the data, distributional enrichment exhibits a strong preference for L-LDA vector representations; in all other cases, the size of improvements achieved by distributional enrichment does not seem to be strongly influenced by the original attribute model being based on C-LDA or L-LDA. These findings are graphically summarized in Fig. 9.5 which shows the largest improvements in F_1 score over original C-LDA and L-LDA models that can be achieved in individual instantiations of distributional enrichment across the various attribute inventories. Distributional enrichment does not affect the overall pattern that L-LDA attribute vectors enable attribute selection at higher precision, whereas C-LDA representations yield an attribute selection performance that is more balanced between precision and recall (cf. discussion in Section 7.3.5).

Impact of objective functions. With respect to the three objective functions being evaluated here, we find that individual models favour the *minimal* setting, whereas joint models tend to perform better with *no-ranks*. In most cases, the gap between unsupervised settings (i.e., *minimal* and *no-ranks*) and the supervised upper bound (*oracle*) is rather small (cf. plain vs. crossed bars in Fig. 9.5). This indicates that objective functions merely based on phrase entropy (i.e., *minimal*) and additional semantic factors (i.e., *no-ranks*) are reasonable approximations of the factor to be optimized during distributional enrichment, i.e., the rank of the correct attribute in the composed phrase vector. On the other hand, the type of information that unsupervised objective functions have at their disposal is not always sufficient for enhancing phrase vector quality, as a decrease in vector entropy may also be due to suboptimal updates which either erroneously promote an incorrect attribute or get stuck after relative improvements of the correct attribute, for instance. These detrimental effects can be observed in some of the paradigmatic enrichment models on the *property* subset, for instance, which are significantly inferior to the original C-LDA attribute model. Yet, in very few configurations involving joint models, it is even more beneficial to use one of *minimal* or *no-ranks* instead of *oracle*. In these cases, the supervised update process guided by `AttrRankComp` runs into local optima that can be bypassed by relying on an objective criterion based on phrase vector entropy.

Impact of joint modelling. Our results clearly show that, taken in isolation, all instantiations of distributional enrichment evaluated here are reasonable, given that each of them tends to develop individual strengths on a particular subset of the data. Furthermore, distributional enrichment generally benefits from combining individual instances of enrichment models into a joint model: First, nearly all instantiations of joint models across the various attribute inventories outperform the respective original attribute model in either precision or recall.[27] Second, the joint models are mostly superior to the individual model instantiations they are composed from, which indicates that the joint approach effectively exploits complementary aspects of meaning in carrier vectors in order to construct attribute-preserving centroids.

Example. Fig. 9.6 shows an update process produced by a joint distributional enrichment model[28] for the phrase *scarce vegetables* (correct attribute: QUANTITY). The original C-LDA model yields an incorrect prediction for this phrase, namely SMELL. The update process visualized in the figure results in an enhanced structured vector representation from which the correct attribute is selected.

The example demonstrates that the carrier vectors provided by the individual enrichment models are individually plausible and capture different aspects of semantic

[27] The C-LDA *property* setting poses the only exception to this pattern.

[28] In this example, the original structured model is a C-LDA model in *property* attribute space; the update procedure is run using *no-ranks* as objective function.

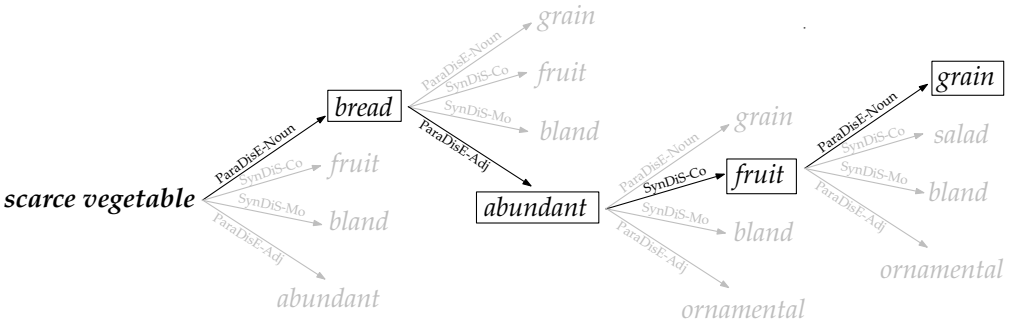


Figure 9.6: Example of an update process produced by joint distributional enrichment, visualized as state transition network. States are depicted as nodes, transitions as labeled edges. Possible transitions correspond to carrier elements provided by an individual enrichment model. Nodes framed in black denote the carrier elements that have been selected in the respective transition.

relatedness: Carriers proposed by ParaDisE-Noun and SynDiS-Co tend to be members of the same ontological category or at least thematically related (e.g., *vegetable–fruit/salad/bread*). ParaDisE-Adj and SynDiS-Mo propose adjectives with a tendency to be located on the same semantic scale (e.g., *scarce–abundant*) or selecting typical properties in the meaning of the modified noun (e.g., *vegetable–bland*), respectively.

Fig. 9.6 also illustrates the potential of joint distributional enrichment models to construct an enhanced phrase vector representation from various semantic sources which mutually complement each other. In the given example, the semantic information that enters the centroid is a mixture of adjectives from the same scale and a collection of nouns from the same and related ontological categories.

Moreover, the example points out how the impact of a particular carrier element may vary, depending on the current state of the update process. The contribution of the carrier *grain* as proposed by the ParaDisE-Noun model, for instance, is found relevant only after three carriers from two different sources have been previously used to shape the centroid towards the intended attribute profile. We consider it another advantage of joint distributional enrichment that the procedure of combining different semantic sources in order to construct an enhanced structured representation is implemented in a way that is aware of possible path depencies. Thus, joint models also bear the potential of bypassing local optima that may occur in individual models.

9.4.3 Evaluation on Test Set

In a last experiment, the attribute selection capacities of the attribute models based on C-LDA and L-LDA, after being subjected to distributional enrichment, are evaluated on the previously held-out HeiPLAS test set. The results are reported in Table 9.11 for all attribute inventories, using the configurations previously optimized on the devel-

	all attrs.			property attrs.			measurable attrs.			selected attrs.		
	P	R	F	P	R	F	P	R	F	P	R	F
ParaDisE	**0.09**	0.05	**0.07**$^{\cdot}$	0.18	**0.15**	0.16	**0.24**	0.18	0.20$^{\cdot}$	**0.39**	0.37	**0.38**
SynDis-Co	**0.09**	**0.06**	**0.07****	0.18	**0.15**	0.16	**0.24**	0.18	**0.21**$^{\cdot}$	0.36	0.34	0.35
SynDis-Mo	0.08	0.05	0.06	**0.24**	0.14	**0.17**	0.23	0.17	0.19	0.37	0.33	0.35
Adj-ParaDisE	0.08	0.05	0.06	0.17	**0.15**	0.16	0.20	0.17	0.19	0.38	**0.39**	**0.38**$^{\cdot}$
EnJoiDis	**0.09**	**0.06**	**0.07***	0.17	**0.15**	0.16	0.22	**0.20**	**0.21**	0.36	**0.39**	**0.38**$^{\cdot}$
ParaDisE-DM	**0.09**	**0.06**	**0.07***	0.19	**0.15**	**0.17**	**0.24**	0.18	**0.21**$^{\cdot}$	0.37	0.35	0.36
C-LDA	0.08	0.05	0.06	0.18	0.14	0.16	0.22	0.16	0.19	0.38	0.33	0.35
ParaDisE	**0.16**	0.02	0.04	0.24	0.07	0.11*	0.22	0.17	0.19	0.41	0.26	0.32
SynDis-Co	0.08	0.02	0.03***	0.24	0.06	0.09$^{\cdot}$	**0.25**	**0.20**	**0.22***	0.45	0.26	0.33
SynDis-Mo	0.09	0.02	0.03	0.23	**0.14**	**0.18*****	0.22	0.18	0.20	0.43	0.24	0.31
Adj-ParaDisE	0.13	0.02	0.03	0.22	0.04	0.07	0.21	0.18	0.19	**0.48**	**0.31**	**0.37**
EnJoiDis	0.13	**0.03**	**0.05****	**0.28**	0.08	0.12*	0.21	**0.20**	0.20	0.42	0.30	0.35
ParaDisE-DM	0.09	0.02	0.03***	0.25	0.05	0.08	0.23	0.18	0.20	0.44	0.26	0.33
L-LDA	0.14	0.02	0.03	**0.28**	0.04	0.07	0.22	0.17	0.19	0.47	0.26	0.33

Table 9.11: Attribute selection performance on HeiPLAS test set after distributional enrichment of attribute models based on C-LDA (upper part) and L-LDA (lower part) attribute models; significance over original C-LDA and L-LDA attribute models, respectively

opment set. This implies vector composition by multiplication, entropy-based attribute selection (ESel) and distributional enrichment in the *minimal* (individual enrichment models) or *no-ranks* setting (joint models). Significance codes used in the table refer to differences over the performance of the original C-LDA or L-LDA model (given in the last rows of the upper and lower part of the table, respectively).

By and large, the results on the test set confirm the observations during development: Distributional enrichment is generally effective, given that, for both C-LDA and L-LDA, significant improvements over the original attribute models can be obtained in all subsets of the data.

More in detail, our large-scale evaluation on the entire HeiPLAS test set results in small, but significant improvements (+0.01 F_1 over C-LDA, +0.02 over L-LDA); on smaller subsets of the data, however, more substantial improvements can be obtained. The strongest growth rate is achieved by applying SynDis-Mo enrichment to an L-LDA property space (+0.11 F_1; p=0.004). In absolute numbers, the best overall result amounts to F_1=0.38 (P=0.38, R=0.39) as obtained from ParaDisE-Adj over C-LDA on the *selected* subset. This corresponds to an increase of +0.03 points in F_1 score (p=0.098). The general tendency of L-LDA favouring precision over recall has been confirmed during testing once again. Therefore, if precision is in focus, it is most advisable to rely on an L-LDA attribute model. The best overall precision is yielded by ParaDisE-Adj over L-LDA in the *selected* subset (P=0.48). This particular configuration of distributional enrichment

also exhibits considerable smoothing power, which results in a recall of R=0.31 (+0.05 over the original L-LDA model; p=0.24) and thus comes very close (F_1=0.37) to the best overall model.

Comparing our specifically designed enrichment models to the generic ParaDisE-DM model, we cannot attest a systematic advantage to the latter: ParaDisE-DM tends to be on a par with more specific models based on C-LDA (except for the *selected* attributes sample, where it lags in recall), whereas it is systematically inferior to at least one specific model throughout all configurations that are based on an L-LDA model.

9.4.4 Discussion

Relationship between attribute inventories and types of enrichment models. Our experimental results shed light on characteristic relationships between types of enrichment models and particular subsets of the data: On the entire large-scale data set, joint distributional enrichment generates the most adequate representations, which is intuitively plausible under the assumption that the entirety of the HeiPLAS data covers a variety of phenomena (cf. Section 8.3.4) for which it is hard to devise a unified enrichment strategy in advance. On the *property* subset, we find a strong preference for SynDis-Mo enrichment. This is also obvious given that this model has been specifically tailored to *property*-denoting adjectival modifiers based on insights from adjective classification (cf. Table 5.11 on page 72). Nevertheless, in view of the relatively small target/carrier overlap in this model instance (cf. Section 9.4.1), the large growth rates obtained are remarkable. These results clearly support the hypothesis that attribute-based meaning representations can be approximated from syntagmatically related predicative adjectives at least for a fraction of attributes. On the measurable subset, the picture is less clear: Among the L-LDA configurations, a preference towards the SynDis-Co model can be observed, whereas for C-LDA, SynDis-Co, ParaDisE, ParaDisE-DM and Joint are almost on par. Finally, the sample of selected attributes clearly demands for enrichment of adjective vectors, given that ParaDisE-Adj stands out here (most clearly visible in L-LDA).

Key aspects of distributional enrichment. These findings are also instructive with respect to the key aspects of distributional enrichment as introduced in Table 9.8 on page 189: In view of the particular effectiveness of individual enrichments models in certain sub-spaces of attribute meaning and the fact that there is no singular configuration of enrichments models that turns out globally most effective, we conclude that factors relating to the functional and semantic relationship between targets and carriers have a stronger impact on the capacities of distributional enrichment models than complementarity and overlap, as the former license to tailor the semantic properties of an auxiliary model to the specific requirements and phenomena in the data. This is corroborated by the observation that, despite its much larger data basis and target/carrier

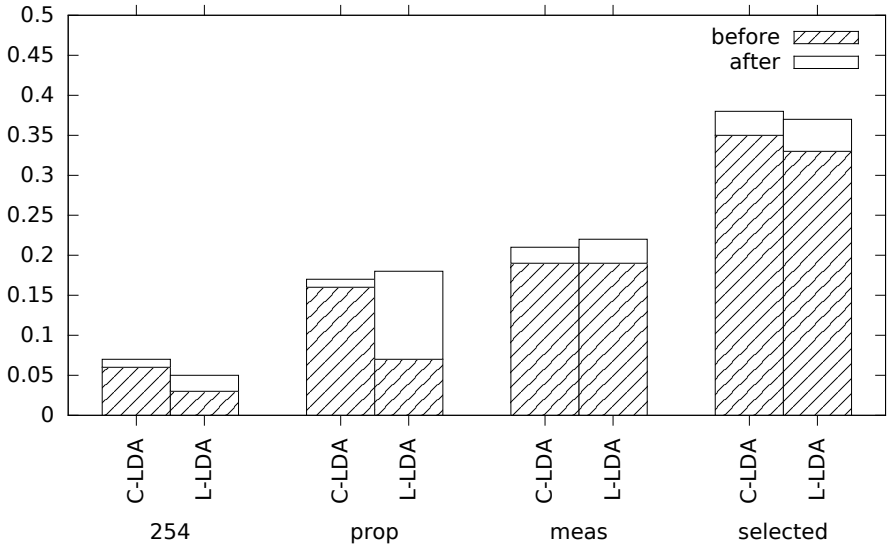


Figure 9.7: Performance of C-LDA and L-LDA attribute models in F_1 score before (shaded bars) and after (plain bars) distributional enrichment

overlap, the generic ParaDisE-DM model is found to be on a par with more specific enrichment models in some configurations, but predominantly inferior.

Comparing C-LDA and L-LDA attribute models after distributional enrichment. After distributional enrichment, the differences in attribute selection performance that have been observed in the original attribute models (cf. Fig. 7.6 on page 121) are largely leveled out. C-LDA models are still slightly superior in the large-scale challenge and on the *selected* attributes subset (and also exhibit a more balanced precision-recall ratio). However, on the *property* and the *measurable* inventories, distributional enrichment causes the best L-LDA models to outperform their C-LDA counterparts.[29] These effects are graphically summarized in Fig. 9.7.

On the one hand, this suggests that in smaller, semantically more confined attribute spaces, it may be reasonable to use less densely populated, more accentuated word representations of the L-LDA type as point of departure, thus leaving the needs for smoothing entirely to the distributional enrichment process. On the other hand, the gradual assimilation of C-LDA and L-LDA, taken together with the relatively small gains obtained from incorporating an oracle (cf. development results in Table 9.10 on page 198), may also be indicative of a plateau effect suggesting that distributional enrichment has exploited its full potential in distributionally enhancing topic-based attribute models in their current state.

[29] These differences between C-LDA and L-LDA models after distributional enrichment are not statistically significant, though.

Conclusion. Taken together, our results show that unsupervised distributional attribute models such as C-LDA benefit from distributional enrichment in terms of enhanced performance in the challenging task of large-scale attribute selection. The improvements are not substantial enough to reduce the disparity between unsupervised and semi-supervised approaches, though. Nevertheless, there is a clear potential in refining structured meaning representations by distributional enrichment. In doing so, tailoring auxiliary models to particular types of distributional information may add an additional gain, compared to using a large-scale distributional resource of lower specificity as auxiliary model.

The positive impact of distributional enrichment notwithstanding, attribute selection crucially depends on appropriately restricted sub-spaces of attribute meaning. Altogether, the attribute inventory offered by WordNet turns out too heterogeneous to provide a solid foundation for attribute-based distributional semantic modeling. The fact that our attribute models yield considerable improvements when being adapted to regions of attribute meaning that are confined by semantic principles (measurable attributes, property attributes) is encouraging for the general prospects of structured distributional models, however. With respect to large-scale attribute selection, automatically grouping related attributes together by hierarchical clustering methods (Manning et al., 2008), for instance, might be a promising avenue to explore in future work.

9.5 Summary

In this chapter, we have introduced a novel framework for distributional enrichment, with the intended purpose of enhancing structured distributional models with recourse to auxiliary distributional models. These are designed to encode complementary semantic information in order to facilitate the reduction of sparsity in the original structured model, while preserving most of the specific semantic information it contains. At the core of distributional enrichment is the idea of substituting structured vector representations that are considered insufficient by a centroid of so-called *carrier vectors*, i.e., other structured vectors that are found in close proximity to the insufficient ones in auxiliary space. The members of the centroid are determined by a greedy update algorithm based on a partial ordering on the carrier vectors with respect to their degree of relatedness and an objective function to assess whether a particular centroid can be expected to contribute to a more adequate structured meaning represenation. The framework licenses different instantiations of this objective function capitalizing on different degrees of supervision.

As proof-of-concept, we have applied distributional enrichment to the attribute-based word representations in C-LDA and L-LDA attribute models. Owing to the insights derived from the analyses in Chapter 8 that attribute selection performance should, for the main part, benefit from improved noun vectors, we devised three variants of auxiliary models for distributional enrichment of structured noun vectors. Relying on the BLESS

data set for benchmarking, these auxiliary models have been tailored to yield attribute-preserving carriers, capitalizing on paradigmatic neighbours and predicative modifiers as carrier elements. These instantiations were supplemented by an additional auxiliary model for distributional enrichment of attribute-based adjective vectors and a model combination for enhancing adjective and noun vectors in a joint approach.

In comparison to the attribute selection performance of the original C-LDA and L-LDA attribute models on the HeiPLAS test set, distributional enrichment yields significant improvements of up to +0.11 points in F_1 score. Moreover, in the majority of attribute spaces investigated (*all*, *measurable*, *property* and *selected* attributes), one instantiation of distributional enrichment is found to outperform a baseline capitalizing on the *Distributional Memory* (DM; Baroni and Lenci, 2011) as auxiliary model, which demonstrates the general potential underlying distributional enrichment for enhancing structured meaning representations. At the same time, these results emphasize the benefits of tailoring auxiliary models to particular semantic relations of interest. From comparing our distributional enrichment models against the DM baseline, we can conclude that the latter point is even more important than other aspects such as the sheer size of the auxiliary space or target/carrier overlap across structured and auxiliary space.

10 Conclusions

The main research focus of this thesis has been on the automated acquisition of attribute knowledge from textual data. For this purpose, we have relied on adjective-noun phrases, as they are a ubiquitous source of this type of ontological knowledge in text corpora. However, this goal implies the challenges that (i) not all adjective-noun phrases convey attribute knowledge, due to different semantic classes of adjectives, (ii) lexical types of adjectives and nouns may be highly ambiguous with respect to their individual attribute meaning, and (iii) attribute meaning is not overtly expressed at the linguistic surface of an adjective-noun phrase, but implicitly embedded in its compositional semantics.

In the face of these challenges, our approach to attribute learning from text has capitalized on corpus-based distributional models. In the form of structured distributional models, they offer a principled account to the unsupervised corpus-based acquisition of distributional information of a specific type or relation. At the same time, structured distributional models tend to induce rather sparse representations of meaning. Therefore, capturing the attribute relation between adjectives and nouns in such a model demands for a practical compromise between specificity and sparsity.

10.1 Contributions of this Thesis

Against this background, this thesis contributes to the state of the art in distributional semantics and knowledge acquisition from text in several ways.

Distributional attribute models between specificity and sparsity. We have developed different variants of distributional attribute models that aim at reconciling the conflicting goals of specificity and sparsity in distributional modeling. For their empirical evaluation, we have established the novel task of attribute selection from adjective-noun phrases for which we have created two manually annotated gold standards covering ten *core attributes* and a *large-scale* inventory of more than 260 attributes, respectively.

All attribute models are based on a structured distributional model which represents target adjectives and nouns in relation to attribute nouns as dimensions of meaning. Two vector mixture operations, viz. vector multiplication and vector addition, are used to compose attribute-based meaning representations of adjective-noun phrases from their constituents. Thus, the complexity of the corpus-based acquisition task is substantially reduced from triples of adjectives, nouns and attributes to pairs of adjectives

and attributes or nouns and attributes, respectively. On top of this distributional core, the attribute models are equipped with unsupervised attribute selection functions for predicting the attribute(s) elicited by an adjective-noun phrase from its composed vector representation.

In the **pattern-based** attribute model, a small, focused set of lexico-syntactic patterns specifically tailored to the task is used to populate attribute-based vector representations of adjectives and nouns. On the core attributes gold standard, this model outperforms previous work on attribute selection from adjectives by wide margins. Extending the task to a linguistically more adequate scenario in which attributes are selected from adjective-noun phrases, our models have achieved robust performance, with precision scores well above 0.60. Moreover, we have empirically demonstrated that our approach of complexity reduction throughout the corpus-based acquisition of meaning representations is essential in order to circumvent fundamental sparsity issues in purely pattern-based distributional models. Otherwise, the acquisition of sufficiently dense attribute-based vector representations is intractable, which renders the attribute selection task infeasible.

Dependency-based models are built from a range of manually devised dependency paths. Compared to lexico-syntactic patterns, these paths (i) provide more flexibility in capturing sparse surface relations between attributes and adjectives/nouns, and (ii) aim at a semantically richer distributional representation of attribute meaning by considering additional paths linking attributes and verbs. Dependency-based models have been found to achieve a substantial advantage over pattern-based ones in terms of recall, while their precision lags behind in most of the cases.

Moreover, we have proposed **topic-based attribute models** which represent attribute meaning via attribute-specific topics induced from weakly supervised variants of Latent Dirichlet Allocation, C-LDA and L-LDA. In their underlying distributional contents, topic-based attribute models are equivalent to the dependency-based ones. However, C-LDA and L-LDA augment purely distributional dependency-based representations of attribute meaning with probabilistic techniques geared towards smoothing and disambiguation. Throughout all experiments, topic-based attribute models incorporating attribute-specific topics have consistently outperformed purely distributional approaches in terms of F_1 score, also offering good trade-offs between precision and recall. A focussed evaluation on particularly sparse vectors has revealed that C-LDA, due to its excellent smoothing capacities, is best prepared among all models to alleviate sparsity issues.

Distributional enrichment. In Chapter 9, we have proposed a novel framework of distributional enrichment that is designed to augment structured distributional models with complementary distributional information. We have applied distributional enrichment to attribute models in order to improve the semantic expressivity of attribute-based noun representations. Contrary to representing each target word by one attribute

vector, distributional enrichment yields centroids of attribute-based noun representations which are acquired from auxiliary models, i.e., distributional sources complementary to the ones taken into account by previous attribute models. In order to enrich attribute-based adjective and noun representations, we have tailored our auxiliary models to provide distributionally similar nouns or semantically related adjectives.

Large-scale attribute selection. In Chapter 7, we have carried out attribute selection on a large scale, presenting the system with more than 260 attributes as possible candidates to select for a given adjective-noun phrase. Our results show that this poses a great challenge to our systems across all settings and configurations. However, restricting the attribute inventory to ontologically confined sets (i.e., *measurable* and *property-denoting* attributes) renders the large-scale selection task more tractable.

Large-scale attribute selection over measurable and property-denoting attributes has served as an empirical testbed for distributional enrichment of C-LDA and L-LDA attribute models. In a contrastive evaluation, we have found that distributional enrichment achieves small, but significant improvements over both types of topic-based attribute models for different configurations of auxiliary models. Using the *Distributional Memory* (Baroni and Lenci, 2011) as auxiliary model for comparison, we have been able to show that distributional enrichment benefits more from the specificity of the complementary information than from the wide coverage of a multi-purpose auxiliary model.

Adjective-noun compositionality. In our experiments, we have found that attribute selection is most effectively modeled as an intersective compositional process, using vector multiplication in order to combine adjective and noun representations. Analyzing the interface of word and phrase meaning in our C-LDA model in Chapter 8, we have identified a recurrent pattern of adjectives having a stronger impact on the prominence of the correct attribute(s) in the phrase representation than nouns. We consider this as supporting evidence for the view that nouns tend to offer a wider range of attributes in their semantics, from which the adjective selects the most appropriate one(s) in the given phrasal combination (Pustejovsky, 1995).

From the perspective of modeling compositional aspects of phrasal semantics in distributional vector space models, we believe that our approach poses an interesting contrast to distributional models involving linear mappings and functional application (Baroni and Zamparelli, 2010; Baroni et al., 2014, inter alia). Contrary to the latter approaches, our models are not geared towards the full compositional semantics of an adjective-noun phrase. Rather, we intend to capture particular aspects of phrasal semantics (i.e., attribute meaning) along interpretable dimensions of meaning which provide an easily interoperable interface to knowledge bases and ontologies.

In contrast to purely pattern-based methods, the compositional approach taken by our attribute models is also capable of resolving ambiguities on the level of word meaning, as in *short hair* vs. *short flight*, for instance. Ambiguities on the phrase level, as in

hot soup or *green plant*, however, are out of the scope of the current model which does not account for context beyond the adjective-noun phrase.

Adjective Classification. In Chapter 5, we have presented a machine learning approach for automatically classifying adjectives into *property-denoting* and *relational* types. In a classification experiment based on a weakly supervised training regime, the classifier separates property-denoting adjectives at excellent performance levels beyond 90% precision. Analyzing the impact of the features used by the classifier, we have identified the occurrence in predicative contextual patterns as a criterion that is highly informative for detecting property-denoting adjectives and can be smoothly integrated into unsupervised corpus-based models of adjective meaning. In the experiments carried out in this thesis, this criterion has proven its effectiveness for improving a pattern-based attribute model in Chapter 7 and for constructing a distributional enrichment model based on property-denoting adjectives in Chapter 9.

10.2 Conclusions and Perspectives

Linguistic insights. Adjective-noun phrases most likely to be successfully modeled by a topic-based attribute model such as C-LDA share the following features in support of concise attribute profiles: low vector entropy, densely populated pseudo-document for the correct attribute, high corpus frequency of the adjective. Moreover, measurability of the attribute and concreteness of the noun are highly beneficial. On the downside, abstract nouns pose severe problems to the C-LDA approach (and presumably other approaches as well).

As a particular strength of the model, thanks to the smoothing capacities of attribute-specific topic modelling, C-LDA is largely capable of generating reliable adjective representations even in cases of markedness, where purely pattern- or dependency-based approaches are shown to fail. In general, C-LDA adjective vectors are of a better individual quality than their counterparts representing nouns, as revealed by their respective impact on the quality of composed vectors and contrastive evaluation of attribute selection from word representations.

Generality of attribute inventories. From our findings on large-scale attribute selection, we conclude that robust performance can only be expected from unsupervised models when being restricted to particularly confined attribute inventories. In that respect, this thesis has also touched a "deep and difficult issue" from the early days of knowledge representation, i.e., which particular attributes, relations and constraints should be established in structures representing general, multi-purpose commonsense knowledge (cf. Barsalou, 1992). In fact, our findings give rise to the conclusion that acquisition of domain- and task-independent attribute knowledge beyond a tightly restricted set of core attributes in the tradition of Almuhareb (2006) is hard for unsuper-

vised attribute models. Attribute selection models should rather be specifically adapted to the particular application or task at hand, as is already common practice in, e.g., aspect-based sentiment analysis (McAuley et al., 2012; Liu, 2015).

If large-scale attribute acquisition is in focus, it is most advisable to present the system with ontologically coherent samples of attributes. In our experiments, a subset of *measurable* attributes proves most feasible. In the coarser granularity of the *Generative Lexicon*, this subset roughly corresponds to aspects in the FORMAL quale which "distinguishes an object within a larger domain", such as orientation, magnitude, shape, dimensionality, color or position (Pustejovsky, 1995).

In this thesis, attributes have been treated as independent and non-hierarchical. Therefore, exploiting possible subordination or inclusion relations between attributes may bear further potential for enhancing the models proposed. Automatically grouping related attributes together, e.g., by hierarchical clustering methods (Manning et al., 2008), might be a feasible alternative to restricting the inventory.

Merits of supervision. In recent work on automatically populating the common-sense knowledge base *WebChild* with triples of attributes, nouns and adjectives, Tandon et al. (2014) achieve very good attribute selection performance beyond C-LDA in a semi-supervised setting, using explicitly encoded adjective-attribute relations from WordNet as seed material. Their approach yields highly accurate triples for a restricted set of 19 attributes. Thus, the coverage limitations of large-scale attribute selection seems to be a more general problem, not only affecting the particular approach taken by C-LDA, but (semi-)supervised methods as well.

Given the superior performance of *WebChild*, it seems worthwhile to explore further prospects of supervision by treating attribute selection as a multi-label classification problem within a discriminative training regime (Lacoste-Julien et al., 2009). Alternatively, the task of predicting the most salient attributes from an adjective-noun phrase could also be formulated in a recursive neural network architecture along the lines of Socher et al. (2013). Such an approach bears the advantage of integrating the aspects of word-level representation, compositionality and attribute prediction in a uniform learning framework. However, it requires substantial amounts of labeled input-output pairs for training (i.e., adjective-noun phrases with their correct attributes).

Implications for learning implicit ontological knowledge from text. Using attribute selection from adjective-noun phrases as an example case, we have been able to show that learning implicit ontological knowledge from textual sources is generally feasible, provided that an effective strategy to overcome the sparsity of overt linguistic cues to the targeted knowledge structures in textual data can be found. To this end, structured distributional models as proposed in this thesis render a valuable service, as they (i) are sufficiently versatile to be tailored to particular types of semantic knowledge to be acquired, (ii) are sufficiently powerful to capture compositional linguistic processes,

(iii) are compatible with a variety of complementary distributional sources that can be harnessed for distributional enrichment of individual structured word representations and (iv) can be extended to incorporate probabilistic smoothing techniques.

With respect to the acquisition of attribute knowledge from adjective-noun phrases, abstractness in meaning turns out to be a major obstacle. Impeding abstractness may be encountered in attributes themselves (in terms of non-measurable attributes, for instance) or in the noun. Both cases are detrimental to distributional attribute models: Abstract attributes are often highly variable in their meaning (Borghi and Binkofski, 2014), which leads to very diverse contextual profiles, limiting the prospects of corpus-based distributional models in the first place. Abstract nouns, being "distant from immediate perception" (Turney et al., 2011), are often hard to describe in terms of concrete attributes which can be seen as approximations of perceptually or experientially grounded information (Silberer and Lapata, 2012).

This is consistent with a line of research in cognitive psychology that postulates a separation between experiential knowledge that is learned through experience with the physical world, on the one hand, and distributional knowledge that is acquired from language, on the other (cf. Andrews et al., 2009). However, the finding that distributional attribute models provides adequate vector representations for a considerable and clearly confined subset of the data provides supporting evidence for the view that it is, at least partially, possible to approximate experiential knowledge from purely textual sources (cf. Baroni et al., 2010), without the need to elicit it from human subjects in costly procedures (McRae et al., 2005; Fountain and Lapata, 2010).

Further applications. Apart from attribute selection, we believe that the models and techniques developed in this thesis may find further application in knowledge acquisition or different areas of natural language processing.

For the attribute selection task, we have focussed on adjective-noun phrases involving property-denoting adjectives. Structured distributional models also bear the potential to treat **relational adjectives** in a similar way in order to detect implicit role-like structures in their semantics (cf. Section 5.1.1) and make them explicit. Due to the denominal characteristics of relational adjectives, the problem may be framed as an instance of classifying implicit semantic relations between nominals (Girju et al., 2007). Assuming that noun-noun relations such as *cause–effect* or *product–producer* tend to follow stable lexicalization patterns that can be observed in corpora (analogously to the attribute relation between adjectives and nouns), this seems a promising field of application for structured distributional models.

With respect to adjective classification, the distinction between **intensional vs. non-intensional adjectives** is still an important open issue which we also have ignored in this thesis. Boleda et al. (2013) hypothesize a connection between the typicality of attributes and non-intensionality, arguing that "the more typical the attribute described by an adjective is for the sort of thing the noun denotes, the closer the phrase vector

is to both its adjective and its noun vector components". In their work, typicality of attributes is operationalized using association metrics based on surface co-occurrences between adjectives and nouns. We believe that compositional attribute-based meaning representations as provided by our attribute models may be more instructive clues to typicality and thus more informative for separating the two classes. Applying C-LDA to this task will also provide interesting insights as to what extent vector representations lacking a distinctive attribute profile are meaningful in their own right or artifacts of weaknesses in the model.

As another promising avenue to explore, our attribute models may be applied to **learning qualia structures** from text, given that properties of target concepts expressed in terms of adjective-noun phrases have not been taken into account in previous work in this area (Cimiano and Wenderoth, 2007; Katrenko and Adriaans, 2008). Using attribute models for qualia learning requires an explicit mapping between qualia and attributes which, at the current state of the art, has to be manually provided. Provided such a mapping is available, attributes serve as an intermediate layer between singular properties and overly abstract qualia roles. In the face of the lessons learned in this thesis, we expect that acquisition methods operating on this middle ground offer better prospects than purely pattern-based approaches targeting singular properties.

Distributional attribute models find an increasingly popular field of application in **aspect-based sentiment analysis**. Until recently, the goal of sentiment analysis has been to assign sentiment labels to large textual spans such as sentences, paragraphs or even entire documents. On the contrary, as defined by Pontiki et al. (2014), aspect-based sentiment analysis aims at the identification of aspects of given target entities and the sentiment expressed for each of these aspects in so-called *subjective phrases* (Klinger and Cimiano, 2013).

(46) The [battery life]$_{aspect}$ of [this camera]$_{target}$ is [too short]$_{subjective}$.

From the example in (46), taken from Klinger and Cimiano (2013) in slightly adapted form, it becomes clear that aspects correspond to attribute nouns (not necessarily ontological attributes, though) and subjective phrases often contain adjectives. Consequently, in the context of sentiment analysis, attribute selection from adjective-noun phrases as pursued in this thesis can also be regarded as *aspect identification*. Harnessing distributional attribute models for aspect-based sentiment analysis requires the following adaptations: (i) The attribute inventory should be tailored to the target domain, as discussed above, (ii) the model must be extended so as to include the polarity of adjectives (by linking them to existing resources such as SentiWordNet (Baccianella et al., 2010), for instance), and (iii) there is a need to broaden the scope of attribute selection beyond the current focus on adjective-noun phrases in order to account for different linguistic realizations of subjective phrases. The latter issue may be addressed in a distributional approach along the lines of Schulte im Walde (2010), using second-order distributional models to propagate implicit attributes from adjective-noun phrases to contextually related words or phrases.

(47) Happy children laugh all day long.

Considering the example in (47), for instance, the attribute EMOTIONALITY as invoked by the phrase *happy children* may be spread to its governing verb *laugh*.

Even at their current state of development, we believe that our topic-based attribute models may provide valuable information for supervised machine learning approaches. For example, Klinger and Cimiano (2013) propose to model targets, aspects and subjective phrases in a joint inference architecture. From the underperformance of their model on the task of predicting *relations* between targets, aspects and subjective phrases, the authors conclude that more informative features are needed. In future work, we will investigate whether the semantic knowledge provided by topic-based attribute models contributes to better performance of structured prediction models in aspect-based sentiment analysis.

In summary, we have shown several interesting applications of distributional attribute models in such diverse areas as formal semantics, knowledge acquisition and, finally, NLP applications. Despite having not explored these avenues within the scope of this thesis, we still believe that distributional semantic models of attribute meaning bear the potential for closing the life cycle of knowledge being automatically induced from textual data in order to flow back into practical applications.

A Different Attribute Inventories

A.1 Core Attributes

AGE	DIRECTION	SIZE	SPEED	TEMPERATURE
COLOR	DURATION	SMELL	TASTE	WEIGHT

Table A.1: Subset of 10 core attributes due to Almuhareb (2006)

A.2 Property Attributes

ABSORBENCY	DEGREE	MODERATION	SHARPNESS
ABSTEMIOUSNESS	DEPTH	MODERNITY	SIZE
ACQUISITIVENESS	DESTRUCTIBILITY	MUSICALITY	SMELL
AGE	DISPOSITION	NUMEROUSNESS	SOLIDITY
ANCESTRY	DISTANCE	OBVIOUSNESS	SPEED
ANIMATENESS	DULLNESS	PERMANENCE	STALENESS
ANIMATION	DURATION	PITCH	STATURE
APPETIZINGNESS	FAIRNESS	POSITION	STRENGTH
ATTENTION	FRESHNESS	POWER	TEMPERATURE
AUDIBILITY	FULLNESS	QUALITY	TEXTURE
BOLDNESS	HARDNESS	QUANTITY	THICKNESS
BREAKABLENESS	HEIGHT	REASONABLENESS	TIMING
COLOR	IMMEDIACY	REGULARITY	VOLUME
COMPLEXION	LENGTH	SENIORITY	WEIGHT
CONSISTENCY	LIGHT	SENSITIVITY	WIDTH
CONTINUITY	LUMINOSITY	SENTIENCE	WILDNESS
CUBICITY	MAGNITUDE	SERIOUSNESS	
CURLINESS	MAJORITY	SEX	
CURRENTNESS	MINORITY	SHAPE	

Table A.2: Subset of 73 property attributes according to WordNet 3.0

A.3 Measurable Attributes

ABSORBENCY	DURATION	MOTION	SIZE
AGE	EFFECTIVENESS	NUMEROUSNESS	SMELL
AIRWORTHINESS	EFFICACY	OPACITY	SOCIABILITY
AUDIBILITY	EQUALITY	PITCH	SOLIDITY
CLARITY	FERTILITY	POSITION	SPEED
CLEANNESS	FRESHNESS	PRICE	STRENGTH
CLEARNESS	HARDNESS	PURITY	TASTE
COLOUR	HEALTH	QUALITY	TEMPERATURE
COMPLEXION	HEIGHT	QUANTITY	TEXTURE
COMPLEXITY	INTELLIGENCE	REPULSION	THICKNESS
CONSISTENCY	LENGTH	SEAWORTHINESS	TYPICALITY
CONSTANCY	LIGHT	SENTIENCE	VALENCE
DEHISCENCE	LIKELIHOOD	SEX	VOLUME
DEPTH	LOGICALITY	SHAPE	WEIGHT
DIFFERENCE	LUMINOSITY	SHARPNESS	WETNESS
DIRECTION	MAGNETISM	SIGNIFICANCE	WIDTH
DISTANCE	MATURITY	SIMILARITY	

Table A.3: Subset of 65 measurable attributes due to manual selection

A.4 WebChild Attributes

ABILITY	FEELING	SENSITIVITY	STRENGTH
APPEARANCE	LENGTH	SHAPE	TASTE
BEAUTY	MOTION	SIZE	TEMPERATURE
COLOR	SMELL	SOUND	WEIGHT
EMOTION	QUALITY	STATE	

Table A.4: Subset of 19 attributes used by Tandon et al. (2014)

A.5 Large-scale Attribute Data Set

ABILITY	CRITICALITY	INTROSPECTIVENESS	PROLIXITY
ABSORBENCY	CUBICITY	INTROVERSION	PROPRIETY
ABSTEMIOUSNESS	CURLINESS	INTRUSIVENESS	PURITY
ABSTRACTNESS	CURRENTNESS	INWARDNESS	QUALITY
ACCURACY	CYCLICITY	KINDNESS	QUANTITY
ACQUISITIVENESS	DEGREE	LAWFULNESS	READINESS
ACTION	DEHISCENCE	LEGALITY	REALITY
ACTIVENESS	DEPTH	LENGTH	REASONABLENESS
ACTUALITY	DESTRUCTIBILITY	LIGHT	REASSURANCE
ADEQUACY	DIFFERENCE	LIKELIHOOD	RECOGNITION
ADMISSIBILITY	DIFFICULTY	LIKENESS	REGULARITY
AFFECTEDNESS	DIRECTION	LITERACY	REPULSION
AGE	DIRECTNESS	LIVELINESS	REPUTE
AIRWORTHINESS	DISPENSABILITY	LOGICALITY	RESPONSIBILITY
ALARM	DISPOSITION	LOYALTY	RIGHTNESS
AMBITION	DISTANCE	LUMINOSITY	SAMENESS
ANCESTRY	DIVERSENESS	MAGNETISM	SARCASM
ANIMATENESS	DOMESTICITY	MAGNITUDE	SEAWORTHINESS
ANIMATION	DORMANCY	MAJORITY	SENIORITY
APPETIZINGNESS	DRAMA	MALEFICENCE	SENSATIONALISM
APPROPRIATENESS	DULLNESS	MALIGNITY	SENSITIVITY
ASSURANCE	DURATION	MANDATE	SENTIENCE
ASTRINGENCY	EASE	MATERIALITY	SEPARATION
ATTENTION	EFFECTIVENESS	MATURITY	SERIOUSNESS
ATTENTIVENESS	EFFICACY	MEASURE	SEX
ATTRACTIVENESS	EMOTIONALITY	MIND	SHAPE
ATTRIBUTION	EQUALITY	MINDFULNESS	SHARPNESS
AUDIBILITY	ESSENTIALITY	MINORITY	SIGNIFICANCE
AUSPICIOUSNESS	EVENNESS	MODERATION	SIMILARITY
AWARENESS	EVIL	MODERNITY	SINCERITY
BEAUTY	EXCITEMENT	MODESTY	SIZE
BEING	EXPLICITNESS	MORALITY	SMELL
BENEFICENCE	EXTINCTION	MOTHERLINESS	SOCIABILITY
BENIGNITY	FAIRNESS	MOTION	SOCIALITY
BOLDNESS	FAMILIARITY	MUSICALITY	SOLIDITY
BREAKABLENESS	FATHERLINESS	NASTINESS	SPEED
CAPABILITY	FEAR	NATURALNESS	STALENESS
CAREFULNESS	FELICITY	NATURE	STATURE
CERTAINTY	FERTILITY	NECESSITY	STATUS
CHANGEABLENESS	FIDELITY	NICENESS	STRENGTH
CHEERFULNESS	FINALITY	NOBILITY	SUBSTANTIALITY
CIVILITY	FOREIGNNESS	NORMALITY	SUCCESS
CLARITY	FORMALITY	NUMERACY	SUFFICIENCY
CLEANNESS	FREEDOM	NUMEROUSNESS	SUSCEPTIBILITY
CLEARNESS	FRESHNESS	OBEDIENCE	TAMENESS
COLOUR	FRIENDLINESS	OBVIOUSNESS	TASTE
COMFORT	FULLNESS	OFFENSIVENESS	TEMPERATURE
COMMERCE	FUNCTION	OPACITY	TEXTURE
COMMONALITY	GENERALITY	ORDINARINESS	THICKNESS
COMMONNESS	GENEROSITY	ORIGINALITY	THOUGHTFULNESS
COMPLETENESS	GLUTTONY	ORTHODOXY	TIMIDITY
COMPLEXION	GOOD	OTHERNESS	TIMING
COMPLEXITY	GREGARIOUSNESS	OUTWARDNESS	TRACTABILITY
COMPREHENSIVENESS	HANDINESS	PASSIVITY	TRUTH
CONCRETENESS	HAPPINESS	PERFECTION	TYPICALITY
CONFIDENCE	HARDNESS	PERMANENCE	ULTIMACY
CONNECTION	HEALTH	PERMISSIVENESS	UNFAMILIARITY
CONSISTENCY	HEIGHT	PIETY	USUALNESS
CONSPICUOUSNESS	HOLINESS	PITCH	UTILITY
CONSTANCY	HONESTY	PLAYFULNESS	VALENCE
CONTINUITY	HONORABLENESS	PLEASANTNESS	VIRGINITY
CONVENIENCE	HUMANENESS	POLITENESS	VIRTUE
CONVENTIONALITY	HUMANNESS	POPULARITY	VOLUME
CONVERTIBILITY	HUMILITY	POSITION	WARINESS
CORRECTNESS	IMMEDIACY	POSSIBILITY	WEIGHT
CORRUPTNESS	IMPORTANCE	POTENCY	WETNESS
COURAGE	INDEPENDENCE	POTENTIAL	WIDTH
COURTESY	INDIVIDUALITY	POWER	WILDNESS
COWARDICE	INTEGRITY	PRACTICALITY	WORTHINESS
CREATIVITY	INTELLIGENCE	PRESENCE	
CREDIBILITY	INTENTIONALITY	PRICE	
CRISIS	INTEREST	PRIDE	

Table A.5: Entire set of all attributes as extracted from WordNet-3.0

B Annotation Instructions for Acquisition of HeiPLAS Gold Standard

B.1 Background and Task Definition

The task we ask you to perform is concerned with classifying adjective-noun phrases according to their **attribute meaning**.

Attributes are terms that **denote properties**. For instance, the attribute SPEED denotes properties such as *fast*, *slow* or *swift*, while COLOR denotes *black*, *red*, etc.

AN phrase	**Property**	**Head Noun**	**Attribute**
fast car	*fast*	*car*	SPEED
slow boat	*slow*	*boat*	SPEED
grey house	*grey*	*house*	COLOR

Table B.1: Examples of attributes and properties in adjective-noun (AN) phrases

Table B.1 exemplifies attribute meaning on the phrase level: If a property-denoting adjective is combined with a head noun to form an adjective-noun (AN) phrase such as *fast car*, this phrase makes a statement about a particular attribute of the entity denoted by the head noun. To be explicit, in *fast car*, *fast* makes a statement about the SPEED of a *car*.

In the course of the experiment, you will be presented attributes together with a number of AN phrases. Your task is to decide, for each AN phrase, whether it makes a statement about that particular attribute or not, in the way just illustrated.

For your convenience, we developed a graphical user interface for you to perform the task (see Figure B.1). Before providing you with more detailed guidelines on how to complete the task in Section B.3, we first give a brief description of the functionality of this interface.

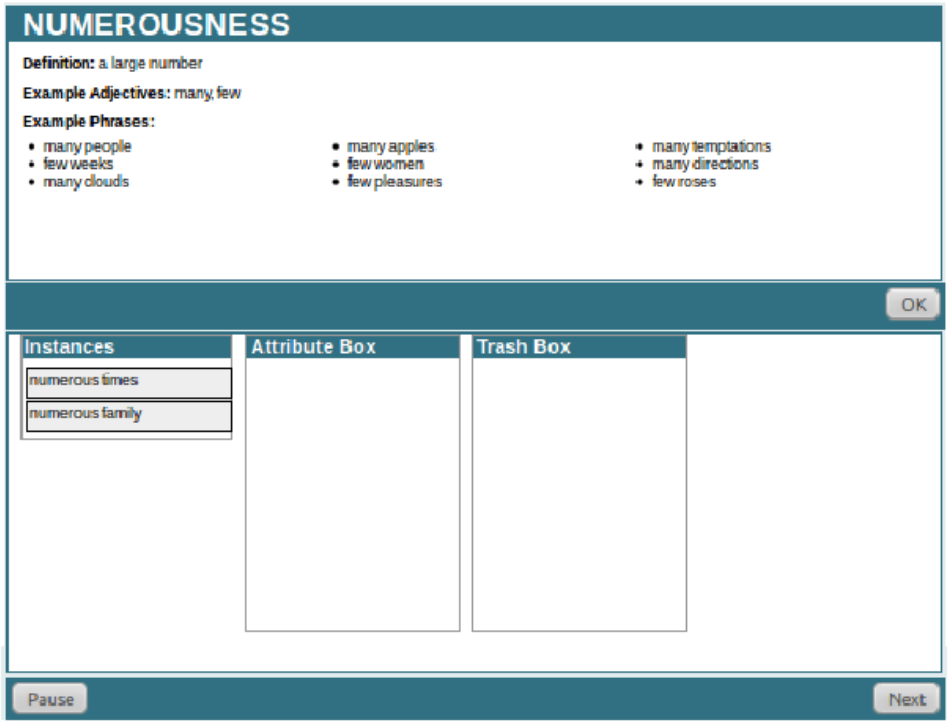


Figure B.1: GUI for completing the classification task

B.2 Functionality of the User Interface

Please direct your web browser to (...)[1] and enter a unique user name[2] into the dialog window showing up. Afterwards, you should be able to see an interface consisting of two frames. In the upper frame, you are given the following information:

Attribute: name of the attribute
Definition: definition of the attribute
Example Adjectives: some typical example adjectives that denote properties belonging to this attribute
Example Phrases: some adjective-noun phrases composed of example adjectives and carefully selected nouns in order to introduce you to the spectrum of the attribute's meaning

Clicking the `OK` button below changes the appearance of the lower frame into one that should be similar to Figure B.1.

[1] Original URL omitted.
[2] Please make sure to remember the name you entered for the case that you want to interrupt the experiment at some point to resume it later on or technical problems occurring.

The lower frame includes three boxes labeled with `Instances`, `Attribute Box` and `Trash Box`. The leftmost box contains all AN phrases to be classified wrt. this attribute. Each of these phrases can be "dragged" (i.e., moved by the mouse, with left mouse key pressed) and "dropped" into one of the other boxes (release left key when the mouse pointer is over the intended target box).

After having dropped all phrases into one of the boxes available, you may proceed to the next attribute by clicking on the `Next` button in the bottom bar. The position of an item in one of the boxes is permanently saved no earlier than this, i.e. before proceeding to the next attribute you are free to move an item back and forth between all boxes.

If you want to interrupt the task, you can do so by clicking the `Pause` button in the bottom bar. The current state of your work will be saved and automatically restored when you come back and identify yourself correctly with the same user name you entered before. Note that only attributes that are completed will be reliably restored. Therefore, we recommend that you make use of the save function only **after** having finished an attribute by clicking "Next". Our apologies for this inconvenience.

B.3 Classification Guidelines

B.3.1 General Instructions

In order to complete the classification task, please proceed as follows:

1. Read carefully through the definition and the examples provided for the attribute.

2. Once you are familiar with the meaning of the attributes and comprehend the examples, click `OK`.

3. Classify each AN phrase by dropping it into either the `Attribute Box` or the `Trash Box`, subject to the test presented in the following section.

Note that many attributes are **scalar**, i.e. the properties they denote can be ordered wrt. different degrees of intensity or occurrence. SPEED, for example, establishes a scale ranging from *slow* to *fast*. In some cases, however, the definition of an attribute and/or the example phrases provided do not reflect the full scale of properties, but merely one end of the entire spectrum. In deciding whether an AN phrase denotes an attribute, please try to always **consider the full scale**, including properties that indicate degrees of low intensity (such as *slow* in the case of SPEED) or even zero intensity of the attribute. The example adjectives provided might give you a hint, as they usually cover both ends of the spectrum.

Note also that the data set may contain several attributes **without** any AN phrases to be dropped into the `Trash Box`.

B.3.2 Classification Test

Can you say that

⟨AN phrase⟩ makes a statement about the ⟨attribute⟩ of the ⟨head noun⟩?

slow boat makes a statement about the SPEED of the *boat*?

If your answer to this test is

- "yes", then put the AN phrase into the `Attribute Box`.
- "no", then put the AN phrase into the `Trash Box`.

In applying this test, you should focus on (i) whether the AN phrase expresses an attribute meaning in the sense given by the definition, and (ii) whether this attribute meaning also applies to the head noun, i.e. whether the phrase really makes a statement about the attribute **of the noun**. Both these aspects are illustrated in the examples below.

Example 1
Attribute: SPEED
Definition: "a rate (usually rapid) at which something happens"
Example Adjectives: slow, fast
AN Phrases to be judged: slow boat, fast lane
Comment: This test is designed to check whether the respective attribute meaning is expressed both in the entire AN phrase and the head noun as well. Thus, it can certainly be said that *slow boat* "makes a statement about the SPEED of the boat", whereas it sounds odd to say that *fast lane* "makes a statement about the SPEED of the *lane*", as the meaning of SPEED does not apply to *lane* directly. In fact, it would be correct to say that in the latter case a statement is made "about the SPEED of the *cars* driving on that *lane*".
Judgements: slow boat → `Attribute Box`; fast lane → `Trash Box`

Example 2
Attribute: AGE
Definition: "how long something has existed"
Example Adjectives: new, immature, young, old, mature
AN Phrases to be judged: little girl, newfound star
Comment: It holds true that *little girl* "makes a statement about the age of the girl" (in addition to a statement about another attribute, namely SIZE).

The statement made in *newfound star* is not "about the age of the star" (in the sense of the definition) but about the period of time that has passed since the star

was found.
Judgements: little girl → `Attribute Box`; newfound star → `Trash Box`

Example 3
Attribute: CHEERFULNESS
Definition: "the quality of being cheerful and dispelling gloom"
Example Adjectives: cheerful, depressing, uncheerful, cheerless
AN Phrase to be judged: blue morning
Comment: Your judgement about the attribute meaning of an AN phrase should not only be confined to its literal meaning, but may also take possible metaphorical interpretations into account. In this example, the adjective *blue* denotes COLOR in its literal meaning, which does certainly not apply to the noun *morning* on its own. In a metaphorical reading, however, *blue* can also be interpreted as *dark* or *gloomy*, thus making a statement "about the CHEERFULNESS of the *morning*" that is bound to the negative end of the spectrum provided by CHEERFULNESS.
Judgement: blue morning → `Attribute Box`

Example 4
Attribute: SUCCESS
Definition: "a state of prosperity or fame"
Example Adjectives: successful, unsuccessful
AN Phrase to be judged: self-made millionaire
Comment: In this example, the notion of SUCCESS in the "state of prosperity" sense provided by the definition is already inherent in the word meaning of the noun *millionaire*. However, this is not sufficient for claiming that the entire phrase including the adjective "makes a statement about the SUCCESS of the *millionaire*". The contribution of *self-made* is such that the **circumstances of achieving** the particular "state of prosperity" are in focus. In fact, it would be more appropriate to paraphrase the meaning of *self-made millionaire* as "making a statement about the circumstances of having become a millionaire". Hence, in this case we observe a situation where the noun does express the attribute meaning on its own, while the adjective contributes to the phrase meaning by bringing some other meaning aspect into focus.
Judgement: self-made millionaire → `Trash Box`

C "Compositionality Puzzles": Examples from HeiPLAS Development Data

Phrase	Attribute	Rank Adj.	Rank Noun	Rank Phrase
*bitter quinine**	TASTE	1	24	24
uneven color	EVENNESS	5	72	15
offensive remark	OFFENSIVENESS	5	245	33
*soothing ointment**	COMFORT	6	239	239
actual beating	REALITY	10	87	11
cardinal rule	IMPORTANCE	8	130	15
high opinion	DEGREE	6	150	12
superior wisdom	QUALITY	5	170	62
even application	EVENNESS	10	196	30
common sailor	COMMONNESS	9	258	26
thick smoke	CONSISTENCY	2	110	16
common man	COMMONNESS	9	125	14
unlikely story	LIKELIHOOD	5	169	35
fearless reporter	BOLDNESS	10	200	34
*peppery salsa**	TASTE	1	24	24
*coarse weave**	TEXTURE	1	22	22
tall ship	STATURE	6	185	30
nasty trick	NASTINESS	8	205	37
unlikely butcher	LIKELIHOOD	5	235	12
common nuisance	COMMONNESS	9	264	44
*short smokestack**	STATURE	3	32	32
*faithful patriot**	FIDELITY	8	174	174
high hope	DEGREE	6	226	56
thick fog	CONSISTENCY	2	106	11
*sacred mosque**	HOLINESS	7	155	155
uneven ground	EVENNESS	5	163	18
*lowly corporal**	SENIORITY	10	51	51
late evening	TIMING	8	114	21
good secretary	QUALITY	5	180	20
meager fare	SUFFICIENCY	4	274	48
high point	DEGREE	6	192	19
nasty accident	NASTINESS	8	243	73

Table C.1: ADJ-n-comp subset from HeiPLAS-Dev data (32 items; cf. Section 8.2); asterisks indicate OOV terms.

Phrase	Attribute	Rank Adj.	Rank Noun	Rank Phrase
dispensable item*	DISPENSABILITY	200	7	200
inclined plane*	DIRECTION	202	2	202
short ration	QUANTITY	45	10	11
affable smile*	FRIENDLINESS	168	10	168
dramatic rescue	DRAMA	207	7	23
competent performance	ADEQUACY	181	8	35
colorful autumn*	COLOR	240	1	240
brave man	COURAGE	136	6	15
brusque manner*	COURTESY	218	3	218
high building	HEIGHT	89	3	12
abominable workmanship*	QUALITY	66	1	66
illusory promise	REALITY	96	7	26
coarse-grained wood*	TEXTURE	22	4	22
impish laughter*	PLAYFULNESS	82	9	82
lifelong study	DURATION	167	3	29
ethereal form	SUBSTANTIALITY	219	5	43
curt reply*	COURTESY	218	3	218
straight line	SHAPE	44	10	11
intractable metal	TRACTABILITY	210	9	25
uncivil tongue*	CIVILITY	244	7	244
atypical behavior	TYPICALITY	15	8	15
voracious shark*	GLUTTONY	163	7	163
vehement defense*	STRENGTH	30	6	30
determinate answer	FINALITY	209	8	33
uncomfortable chair	COMFORT	193	1	19
efficacious medicine*	EFFICACY	189	1	189
admissible evidence*	ADMISSIBILITY	275	3	275
odorless flower*	SMELL	38	2	38
accurate measurement	ACCURACY	46	6	13
critical shortage	CRISIS	147	6	23
glaring error	CONSPICUOUSNESS	216	4	21
untouchable resource	HANDINESS	253	9	30
little boy	AGE	206	1	33
courageous example	COURAGE	177	5	19
stale bread	STALENESS	234	2	12
treble voice	PITCH	164	5	32
odorous bread*	SMELL	38	6	38
attentive suitor	ATTENTION	278	5	14
prismatic light*	COLOR	240	5	240
hallucinatory dream*	REALITY	63	3	63
separate church	SEPARATION	136	3	14
gluttonous appetite*	GLUTTONY	163	6	163
acknowledged accomplishment	RECOGNITION	127	10	31
challenging task	DIFFICULTY	141	10	35
practical application	PRACTICALITY	93	6	15
knotty problem*	COMPLEXITY	233	9	233
true story	TRUTH	85	9	15
scarce vegetable	QUANTITY	111	6	11

Table C.2: adj-N-comp subset from HeiPLAS-Dev data (47 items; cf. Section 8.2)

Phrase	Attribute	Rank Adj.	Rank Noun	Rank Phrase
unconventional dress	CONVENTIONALITY	23	22	9
insufficient fund	QUANTITY	13	30	4
responsible cabinet	RESPONSIBILITY	17	17	3
deep concentration	DEPTH	14	54	5
uncommon flood	COMMONNESS	17	27	5
black deed	EVIL	42	18	6
rough life	PLEASANTNESS	24	14	5
heavy fog	THICKNESS	15	40	8
right hand	POSITION	19	19	8
big business	SIZE	15	21	5
great work	IMPORTANCE	11	20	1
massive sculpture	SIZE	36	17	8
ample waistline	SIZE	22	25	7
direct exposure	IMMEDIACY	38	15	4
speedy resolution	SPEED	18	25	4
pure tone	PURITY	24	33	6
sharp point	SHARPNESS	23	16	6
potent toxin	POTENCY	17	29	10
pure air	PURITY	24	17	6
fundamental revolution	SIGNIFICANCE	14	19	5
dirty work	CLEANNESS	35	15	7
amiable villain	NATURE	14	23	7
potent weapon	POWER	12	18	5
genuine emotion	SINCERITY	13	22	10
responsible captain	RESPONSIBILITY	17	16	2
low furniture	HEIGHT	17	15	5
right bank	POSITION	19	12	5
miniature camera	SIZE	11	31	9
right side	POSITION	19	18	7

Table C.3: adj-n-COMP subset from HeiPLAS-Dev data (29 items; cf. Section 8.2)

Bibliography

Agichtein, E. and Gravano, L. (2000). Snowball. Extracting Relations from Large Plain-Text Collections. In *Proceedings of the 5th International Conference on Digital Libraries*.

Agirre, E., López de Lacalle, O., and Soroa, A. (2014). Random Walks for Knowledge-Based Word Sense Disambiguation. *Computational Linguistics*, 40(1):57–84.

Alfonseca, E., Pasca, M., and Robledo-Arnuncio, E. (2010). Acquisition of Instance Attributes via Labeled and Related Instances. In *Proceedings of the 33rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR-10)*, pages 58–65.

Allan, J. and Kumaran, G. (2003). Stemming in the Language Modeling Framework. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 455–456, New York, NY, USA. ACM.

Almuhareb, A. (2006). *Attributes in Lexical Acquisition*. Ph.D. Dissertation, Department of Computer Science, University of Essex.

Almuhareb, A. and Poesio, M. (2004). Attribute-based and Value-based Clustering. An Evaluation. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, Barcelona, Spain, pages 158–165.

Amoia, M. and Gardent, C. (2007). A First Order Semantic Approach to Adjectival Inference. In *Proceedings of the ACL Workshop on Textual Entailment and Paraphrasing*, pages 185–192.

Amoia, M. and Gardent, C. (2008). A Test Suite for Inference Involving Adjectives. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, pages 631–637.

Andrews, M., Vigliocco, G., and Vinson, D. (2009). Integrating Experiential and Distributional Data to Learn Semantic Representations. *Psychological Review*, 116:463–498.

Asher, N. (2011). *Lexical Meaning in Context. A Web of Words*. Cambridge University Press.

Baayen, R., Piepenbrock, R., and Gulikers, L. (1996). *CELEX2*. Linguistic Data Consortium, Philadelphia.

Baayen, R. H. (2008). *Analyzing Linguistic Data. A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baccianella, S., Esuli, A., and Sebastiani, F. (2010). SentiWordNet 3.0. An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 2200–2204, Valletta, Malta.

Bakhshandeh, O. and Allen, J. F. (2015). From Adjective Glosses to Attribute Concepts. Learning Different Aspects That an Adjective Can Describe. In *Proceedings of the 11th International Conference on Computational Semantics (IWCS)*, pages 23–33, London, UK.

Baldwin, T. (2006). Data-driven Methods for Acquiring Lexical Semantics. In *Lecture Notes from the Foundational Course on Data-driven Methods for Acquiring Lexical Semantics*, ESSLLI 2006, Malaga, Spain.

Banko, M., Cafarella, M. J., Soderland, S., Broadhead, M., and Etzioni, O. (2007). Open Information Extraction for the Web. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 2670–2676.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014). Frege in Space. A Program for Compositional Distributional Semantics. *Linguistic Issues in Language Technology*, 9:5–110.

Baroni, M., Bernardini, S., Ferraresi, A., and Zanchetta, E. (2009). The WaCky Wide Web. A Collection of Very Large Linguistically Processed Web-crawled Corpora. *Journal of Language Resources and Evaluation*, 43(3):209–226.

Baroni, M. and Lenci, A. (2010). Distributional Memory. A General Framework for Corpus-based Semantics. *Computational Linguistics*, 36:673–721.

Baroni, M. and Lenci, A. (2011). How We BLESSed Distributional Semantic Evaluation. In *Proceedings of the EMNLP Workshop on Goometrical Models of Natural Language Semantics (GEMS)*, pages 1–10, Stroudsburg, PA, USA. Association for Computational Linguistics.

Baroni, M., Murphy, B., Barbu, E., and Poesio, M. (2010). Strudel. A Corpus-based Semantic Model of Based on Properties and Types. *Cognitive Science*, 34:222–254.

Baroni, M. and Zamparelli, R. (2010). Nouns are Vectors, Adjectives are Matrices. Representing Adjective-Noun Constructions in Semantic Space. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, East Stroudsburg, PA, pages 1183–1193.

Barsalou, L. W. (1992). Frames, Concepts and Conceptual Fields. In Lehrer, A. and Kittay, E., editors, *Frames, Fields and Contrasts*, pages 21–74. Lawrence Erlbaum Associates, Hillsday, NJ.

Barsalou, L. W. (2010). Grounded Cognition. Past, Present, and Future. *Topics in Cognitive Science*, 2:716–724.

Batista, G., Prati, R., and Monard, M. C. (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *SIGKDD Explorations*, 6:20–29.

Beesley, K. R. (1982). Evaluative Adjectives as One-Place Predicates in Montague Grammar. *Journal of Semantics*, 1:195–249.

Belsley, D. A., Kuh, E., and Welsch, R. E. (1980). *Regression Diagnostics. Identifying Influential Data and Sources of Collinearity*. Wiley, New York.

Bergsma, S., Lin, D., and Goebel, R. (2008). Discriminative Learning of Selectional Preference from Unlabeled Text. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 59–68, Honolulu, Hawaii. Association for Computational Linguistics.

Bing, L., Lam, W., and Wong, T.-L. (2013). Wikipedia Entity Expansion and Attribute Extraction from the Web. In *Proceedings of the 6th ACM International Conference on Web Search and Data Mining (WSDM'13)*, pages 567–576, Rome, Italy.

Blackburn, P. and Bos, J. (2005). *Representation and Inference for Natural Language. A First Course in Computational Semantics*. CSLI Publications, Stanford, CA.

Blei, D. M. and Lafferty, J. D. (2009). Topic Models. In *Text Mining. Classification, Clustering, and Applications*, volume 10 of *Data Mining and Knowledge Discovery Series*. Chapman & Hall/CRC Press, Boca Raton, FL.

Blei, D. M., Ng, A., and Jordan, M. (2003). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 3:993–1022.

Boleda, G. (2006). *Automatic Acquisition of Semantic Classes for Adjectives*. Ph.D. Dissertation, Pompeu Fabra University, Barcelona.

Boleda, G., Baroni, M., Pham, T. N., and McNally, L. (2013). Intensionality was only Alleged. On Adjective-Noun Composition in Distributional Semantics. In *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)*, pages 35–46. Association for Computational Linguistics.

Boleda, G., Vecchi, E. M., Cornudella, M., and McNally, L. (2012). First Order vs. Higher Order Modification in Distributional Semantics. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1223–1233. Association for Computational Linguistics.

Borghi, A. M. and Binkofski, F. (2014). *Words as Social Tools. An Embodied View on Abstract Concepts*. Springer Briefs in Cognition. Springer.

Bos, J. (2008). Wide-coverage Semantic Analysis with Boxer. In *Proceedings of the 2008 Conference on Semantics in Text Processing*, STEP '08, pages 277–286, Venice, Italy. Association for Computational Linguistics.

Brants, T. and Franz, A. (2006). *Web 1T 5-gram Version 1 LDC2006T13*. Linguistic Data Consortium, Philadelphia.

Brin, S. (1999). Extracting Patterns and Relations from the World Wide Web. In *Selected Papers from the International Workshop on the World Wide Web and Databases*, WebDB '98, pages 172–183, London, UK. Springer.

Brown, E., Epstein, E., Murdock, J. W., and Fin, T.-H. (2013). Tools and Methods for Building Watson. Technical Report RC25356 (WAT1302-021), IBM Research Division, Yorktown Heights, NY.

Budanitsky, A. and Hirst, G. (2006). Evaluating WordNet-based Measures of Lexical Semantic Relatedness. *Computational Linguistics*, 32(1):13–48.

Buitelaar, P., Cimiano, P., and Magnini, B. (2005). Ontology Learning from Text. An Overview. In Buitelaar, P., Cimiano, P., and Magnini, B., editors, *Ontology Learning from Text. Methods, Evaluation and Applications*, pages 3–12. IOS Press, Amsterdam.

Bullinaria, J. A. and Levy, J. P. (2007). Extracting Semantic Representations from Word Co-occurrence Statistics. A Computational Study. *Behavior Research Methods*, pages 510–526.

Carnap, R. (1947). *Meaning and Necessity. A Study in Semantics and Modal Logic*. Chicago University Press, Chicago, IL.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J. L., and Blei, D. M. (2009). Reading Tea Leaves. How Humans Interpret Topic Models. In Bengio, Y., Schuurmans, D., Lafferty, J., Williams, C., and Culotta, A., editors, *Advances in Neural Information Processing Systems 22*, pages 288–296. Curran Associates, Inc.

Charniak, E. (1996). *Statistical Language Learning*. MIT Press.

Chen, S. F. and Goodman, J. (1999). An Empirical Study of Smoothing Techniques for Language Modeling. *Computer Speech and Language*, 13(4):359–394.

Christ, O., Schulze, B. M., Hofmann, A., and König, E. (1999). The IMS Corpus Workbench: Corpus Query Processor. Technical report, IMS, University of Stuttgart.

Church, K. W. and Hanks, P. (1990). Word Association Norms, Mutual Information, and Lexicography. *Computational Linguistics*, 16:22–29.

Cimiano, P. (2006). *Ontology Learning and Population from Text. Algorithms, Evaluation and Applications*. Springer.

Cimiano, P., Unger, C., and McCrae, J. (2014). *Ontology-based Interpretation of Natural Language*, volume 24 of *Synthesis Lectures on Human Language Technologies*. Morgan & Claypool Publishers.

Cimiano, P. and Wenderoth, J. (2007). Automatic Acquisition of Ranked Qualia Structures from the Web. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 888–895.

Cohen, J., Cohen, P., West, S. G., and Aiken, L. S. (2003). *Applied Multiple Regression and Correlation Analysis for the Behavioral Sciences*. Lawrence Erlbaum Associates, 3rd edition.

Corley, C. and Mihalcea, R. (2005). Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18. Association for Computational Linguistics.

Dagan, I., Marcus, S., and Markovitch, S. (1993). Contextual Word Similarity and Estimation from Sparse Data. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics (ACL'93)*, pages 164–171, Stroudsburg, PA. Association for Computational Linguistics.

Darlington, R. B. (1968). Multiple Regression in Psychological Research and Practice. *Psychological Bulletin*, 69:161–182.

de Melo, G. and Bansal, M. (2013). Good, Great, Excellent. Global Inference of Semantic Intensities. *Transactions of the Association for Computational Linguistics (TACL)*, 1:279–290.

de Saussure, F. (2011 [1916]). *Course in General Linguistics*. Columbia University Press.

Deerwester, S., Dumais, S., Landauer, T., Furnas, G., and Harshman, R. (1990). Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*, 41:391–407.

Dowty, D. R. (1979). *Word Meaning and Montague Grammar*. Reidel, Dordrecht.

Erk, K. (2009a). Representing Words as Regions in Vector Space. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 57–65, Boulder, Colorado. Association for Computational Linguistics.

Erk, K. (2009b). Supporting Inferences in Semantic Space. Representing Words as Regions. In *Proceedings of the 8th International Conference on Computational Semantics (IWCS-8)*, pages 104–115, Tilburg, The Netherlands. Association for Computational Linguistics.

Erk, K. (2012). Vector Space Models of Word Meaning and Phrase Meaning. A Survey. *Language and Linguistics Compass*, 6(10):635–653.

Erk, K. and Padó, S. (2008). A Structured Vector Space Model for Word Meaning in Context. In *Proceedings of EMNLP*, Honolulu, HI.

Erk, K. and Padó, S. (2009). Paraphrase Assessment in Structured Vector Space. Exploring Parameters and Datasets. In *Proceedings of the EACL Workshop on Geometrical Methods for Natural Language Semantics (GEMS)*, pages 57–65, Athens, Greece.

Erk, K., Padó, S., and Padó, U. (2010). A Flexible, Corpus-Driven Model of Regular and Inverse Selectional Preferences. *Computational Linguistics*, 36(4):723–764.

Etzioni, O., Banko, M., Soderland, S., and Weld, D. S. (2008). Open Information Extraction from the Web. *Communications of the ACM*, 51:68–74.

Fang, Y. and Chang, K. C.-C. (2011). Searching Patterns for Relation Extraction over the Web. Rediscovering the Pattern-Relation Duality. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining*, pages 825–834, New York. ACM.

Fellbaum, C., editor (1998). *WordNet. An Electronic Lexical Database*. MIT Press.

Fernández, M., Cantador, I., López, V., Vallet, D., Castells, P., and Motta, E. (2011). Semantically Enhanced Information Retrieval. An Ontology-based Approach. *Web Semantics. Science, Services and Agents on the World Wide Web*, 9(4):434–452.

Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E. T. (2013). Watson. Beyond Jeopardy. *Artificial Intelligence*, 199–200:93–105.

Fillmore, C. (1968). The Case for Case. In Bach, E. and Harms, R., editors, *Universals in Linguistic Theory*, pages 1–88. Holt, Rinehart and Winston, New York.

Firth, J. R. (1957). Modes of Meaning. In *Papers in Linguistics 1934–1951*, pages 190–215. Longmans, London, UK.

Fleiss, J. L. (1971). Measuring Nominal Scale Agreement Among Many Raters. *Psychological Bulletin*, 76:378–382.

Fountain, T. and Lapata, M. (2010). Meaning Representation in Natural Language Categories. In Ohlsson, S. and Catrambone, R., editors, *Proceedings of the 32nd Annual Conference of the Cognitive Science Society*, pages 1916–1921.

Fox, J. (1997). *Applied Regression Analysis, Linear Models and Related Models*. SAGE Publications.

Fox, J. and Weisberg, S. (2011). *An R Companion to Applied Regression*. Sage, Thousand Oaks, CA.

Frank, A. and Padó, S. (2012). Semantics in Computational Lexicons. In Maienborn, C., von Heusinger, K., and Portner, P., editors, *Semantics. An International Handbook of Natural Language Meaning*, HSK Handbooks of Linguistics and Communication Science Series, pages 2887–2917. Mouton de Gruyter.

Frege, G. (1892). Über Sinn und Bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Frisson, S., Pickering, M. J., and McElree, B. (2011). The Difficult Mountain. Enriched Composition in Adjective-Noun Phrases. *Psychonomic Bulletin & Review*, 18:1172–1179.

Gamerschlag, T. (2008). Stative Dimensional Verbs. In *The 18th International Congress of Linguistics*, Seoul, Korea. Handout.

Girju, R., Badulescu, A., and Moldovan, D. (2006). Automatic Discovery of Part-Whole Relations. *Computational Linguistics*, 32(1):83–135.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2007). SemEval-2007 Task 04. Classification of Semantic Relations between Nominals. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007)*, pages 13–18. Association for Computational Linguistics.

Girju, R., Nakov, P., Nastase, V., Szpakowicz, S., Turney, P., and Yuret, D. (2009). Classification of Semantic Relations between Nominals. *Language Resources and Evaluation*, 43(2):105–121.

Grefenstette, E. and Sadrzadeh, M. (2011). Experimental Support for a Categorical Compositional Distributional Model of Meaning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 1394–1404. Association for Computiational Linguistics.

Grefenstette, E., Sadrzadeh, M., Clark, S., Coecke, B., and Pulman, S. (2014). Concrete Sentence Spaces for Compositional Distributional Models of Meaning. In Bunt, H., Bos, J., and Pulman, S., editors, *Computing Meaning*, volume 4, pages 71–86. Springer.

Gruber, T. R. (1993). A Translation Approach to Portable Ontology Specifications. *Knowledge Acquisition*, 5:199–220.

Gruber, T. R. and Cohen, P. R. (1987). Design for Acquisition. Principles of Knowledge-System Design to Facilitate Knowledge Acquisition. *International Journal of Man-Machine Studies*, 26(2):143–159.

Guarino, N. (1992). Concepts, Attributes and Arbitrary Relations. *Data & Knowledge Engineering*, 8:249–261.

Guevara, E. (2010). A Regression Model of Adjective-Noun Compositionality in Distributional Semantics. In *Proceedings of the 2010 Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, Stroudsburg, PA. Association for Computational Linguistics.

Harrell, F. E. (2013). *rms. Regression Modeling Strategies*. R package version 4.0-0.

Harris, Z. (1954). Distributional Structure. *Word*, 10:146–162.

Hartung, M. and Frank, A. (2010a). A Semi-supervised Type-based Classification of Adjectives. Distinguishing Properties and Relations. In *Proceedings of the 7th International Conference on Language Resources and Evaluation,* Valletta, Malta, May, pages 1029–1036.

Hartung, M. and Frank, A. (2010b). A Structured Vector Space Model for Hidden Attribute Meaning in Adjective-Noun Phrases. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING),* Beijing, China, pages 430–438.

Hartung, M. and Frank, A. (2011a). Assessing Interpretable, Attribute-related Meaning Representations for Adjective-Noun Phrases in a Similarity Prediction Task. In *Proceedings of the EMNLP Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, Edinburgh, UK.

Hartung, M. and Frank, A. (2011b). Exploring Supervised LDA Models for Assigning Attributes to Adjective-Noun Phrases. In *Proceedings of EMNLP 2011*, Edinburgh, UK.

Hartung, M. and Frank, A. (2014). Distinguishing Properties and Relations in the Denotation of Adjectives. An Empirical Investigation. In Gamerschlag, T., Gerland, D., Osswald, R., and Petersen, W., editors, *Frames and Concept Types*, pages 179–197. Springer.

Hatzivassiloglou, V. and McKeown, K. R. (1993). Towards the Automatic Identification of Adjectival Scales: Clustering Adjectives According to Meaning. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics,* Columbus, Ohio, pages 172–182.

Hearst, M. A. (1992). Automatic Acquisition of Hyponyms from Large Text Corpora. In *Proceedings of the 14th Conference on Computational Linguistics (COLING '92)*, pages 539–545, Stroudsburg, PA, USA. Association for Computational Linguistics.

Heim, I. and Kratzer, A. (1998). *Semantics in Generative Grammar*. Blackwell, Oxford, UK.

Hindle, D. and Rooth, M. (1993). Structural Ambiguity and Lexical Relations. *Computational Linguistics*, 19:103–120.

Hu, X., Cai, Z., Wiemer-Hastings, P., Graesser, A. C., and McNamara, D. S. (2007). Strengths, Limitations and Extensions of LSA. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis. A Road to Meaning*, pages 401–426. Lawrence Erlbaum Associates, Hillsday, NJ.

Jonnalagadda, S., Cohen, T., Wu, S., and Gonzalez, G. (2012). Enhancing Clinical Concept Extraction with Distributional Semantics. *Journal of Biomedical Informatics*, 45:129–140.

Jänich, K. (1994). *Linear Algebra*. Springer, New York.

Kamp, H. (1975). Two Theories about Adjectives. In Keenan, E. L., editor, *Formal Semantics of Natural Language*, pages 123–155. Cambridge University Press.

Karlgren, J. and Sahlgren, M. (2001). From Words to Understanding. In Uresaka, Y., Kanerva, P., and Asoh, H., editors, *Foundation of Real-World Intelligence*, pages 294–308. CSLI Publications, Stanford, CA.

Katrenko, S. and Adriaans, P. (2008). Qualia Structures and their Impact on the Concrete Noun Categorization Task. In *Proceedings of the ESSLLI Workshop on Bridging the Gap between Semantic Theory and Computational Simulations*, pages 17–24.

Kilgarriff, A. (1997). Putting Frequencies in the Dictionary. *International Journal of Lexicography*, 10(2):135–155.

Kilgarriff, A. and Rosenzweig, J. (2000). Framework and Results for English SENSEVAL. *Computers and the Humanities*, 34(1/2):15–48.

Klinger, R. and Cimiano, P. (2013). Joint and Pipeline Probabilistic Models for Fine-Grained Sentiment Analysis. Extracting Aspects, Subjective Phrases and their Relations. In *Proceedings of the 2013 IEEE 13th International Conference on Data Mining Workshops (ICDMW)*, pages 937–944.

Kolda, T. G. and Bader, B. W. (2009). Tensor Decompositions and Applications. *SIAM REVIEW*, 51(3):455–500.

Kotlerman, L., Dagan, I., Szpektor, I., and Zhitomirsky-Geffet, M. (2010). Directional Distributional Similarity for Lexical Inference. *Natural Language Engineering*, 16(4):359–389.

Kremer, G., Hartung, M., Padó, S., and Riezler, S. (2012). Statistical Machine Translation Support Improves Human Adjective Translation. *Translation: Computation, Corpora, Cognition. Special Issue on the Crossroads Between Contrastive Linguistics, Translation Studies, and Machine Translation*, 2(1):103–126.

Krengel, U. (2003). *Wahrscheinlichkeitstheorie und Statistik*. Vieweg, Wiesbaden.

Kucera, H. and Francis, W. N. (1967). Computational Analysis of Present-Day American English. Technical report, Department of Linguistics, Brown University, Providence, RI.

Lacoste-Julien, S., Sha, F., and Jordan, M. I. (2009). DiscLDA. Discriminative Learning for Dimensionality Reduction and Classification. In Koller, D., Schuurmans, D., Bengio, Y., and Bottou, L., editors, *Advances in Neural Information Processing Systems 21*, pages 897–904. Curran Associates, Inc.

Landauer, T., Foltz, P., and Laham, D. (1998). Introduction to Latent Semantic Analysis. *Discourse Processes*, 25:259–284.

Lapata, M. (2001). *The Acquisition and Modeling of Lexical Knowledge. A Corpus-based Investigation of Systematic Polysemy*. Ph.D. Dissertation, University of Edinburgh.

Leacock, C. and Chodorow, M. (1998). Combining Local Context and WordNet Similarity for Word Sense Identification. In Fellbaum, C., editor, *WordNet. An Electronic Lexical Database*, pages 265–284. MIT Press.

Lee, T., Wang, Z., Wang, H., and Hwang, S.-w. (2013). Attribute Extraction and Scoring. A Probabilistic Approach. In *Proceeding of the 29th IEEE International Conference on Data Engineering (ICDE 2013)*, pages 194–205, Brisbane, Australia.

Lenat, D. (1995). Cyc. A Large-Scale Investment in Knowledge Infrastructure. *Communications of the ACM*, 38:33–38.

Lenci, A. (2010). The Life Cycle of Knowledge. In Huang, C.-R., Calzolari, N., Gangemi, A., Lenci, A., Oltramari, A., and Prevot, L., editors, *Ontologies and the Lexicon. A Natural Language Processing Perspective*, pages 241–257. Cambridge University Press.

Lenci, A. and Benotto, G. (2012). Identifying Hypernyms in Distributional Semantic Spaces. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 75–79, Montréal, Canada. Association for Computational Linguistics.

Levin, B. (1993). *English Verb Classes and Alternations. A Preliminary Investigation*. University of Chicago Press, Chicago, Illinois.

Levinson, S. C. (1983). *Pragmatics*. Cambridge University Press, Cambridge, UK.

Lewis, M. and Steedman, M. (2013). Combined Distributional and Logical Semantics. *Transactions of the Association for Computational Linguistics*, 1:179–192.

Lin, D. (1998). Automatic Retrieval and Clustering of Similar Words. In *Proceedings of the 17th International Conference on Computational Linguistics (COLING-ACL'98)*, pages 768–774, Morristown, NJ. Association for Computational Linguistics.

Lin, D. and Pantel, P. (2001). Discovery of Inference Rules for Question Answering. *Natural Language Engineering*, 7(4):343–360.

Lin, D., Zhao, S., Qin, L., and Zhou, M. (2003). Identifying Synonyms among Distributionally Similar Words. In *Proceedings of IJCAI-03*, pages 1492–1493.

Liu, B. (2015). *Sentiment Analysis. Mining Opinions, Sentiments, and Emotions*. Cambridge University Press.

Lutz, J. G. (1983). A Method for Constructing Data which Illustrate Three Types of Suppressor Variables. *Educational and Psychological Measurement*, 43:373–377.

Löbner, S. (2013). *Understanding Semantics*. Routledge, 2nd edition.

MacKay, D. J. C. (2003). *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press.

Manandhar, S., Klapaftis, I. P., Dligach, D., and Pradhan, S. S. (2010). SemEval-2010 Task 14: Word Sense Induction and Disambiguation. In *Proceedings of the 5th International Workshop on Semantic Evaluation (SemEval-2010)*, pages 63–68.

Manning, C. D., Raghavan, P., and Schütze, H. (2008). *Introduction to Information Retrieval*. Cambridge University Press.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing*. MIT Press.

Marcus, M. P., Marcinkiewicz, M. A., and Santorini, B. (1993). Building a Large Annotated Corpus of English. The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Martin, D. I. and Berry, M. W. (2007). Mathematical Foundations of Latent Semantic Analysis. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis. A Road to Meaning*, pages 35–55. Lawrence Erlbaum Associates, Hillsday, NJ.

Matuszek, C., Cabral, J., Witbrock, M., and De Oliveira, J. (2006). An Introduction to the Syntax and Content of Cyc. In *Proceedings of the 2006 AAAI Spring Symposium on Formalizing and Compiling Background Knowledge and Its Applications to Knowledge Representation and Question Answering*, pages 44–49.

McAuley, J., Leskovec, J., and Jurafsky, D. (2012). Learning Attitudes and Attributes from Multi-Aspect Reviews. In *Proceedings of the 2012 IEEE 12th International Conference on Data Mining*, pages 1020–1025, Washington, DC, USA. IEEE Computer Society.

McCallum, A. K. (2002). MALLET. A Machine Learning for Language Toolkit. Technical report. http://mallet.cs.umass.edu.

McIntosh, T. and Curran, J. R. (2009). Reducing Semantic Drift with Bagging and Distributional Similarity. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 396–404, Suntec, Singapore. Association for Computational Linguistics.

McNamara, D. S., Cai, Z., and Louwerse, M. M. (2007). Optimizing LSA Measures of Cohesion. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis. A Road to Meaning*, pages 379–400. Lawrence Erlbaum Associates, Hillsday, NJ.

McNemar, Q. (1947). Note on the Sampling Error of the Difference between Correlated Proportions or Percentages. *Psychometrika*, 12(2):153–157.

McRae, K., Cree, G. S., Seidenberg, M. S., and McNorgan, C. (2005). Semantic Feature Production Norms for a Large Set of Living and Nonliving Things. *Behavior Research Methods*, 37:547–559.

Merlo, P. and Ferrer, E. E. (2006). The Notion of Argument in Prepositional Phrase Attachment. *Computational Linguistics*, 32:341–378.

Mervis, C. B., Catlin, J., and Rosch, E. H. (1976). Relationships among Goodness-of-Example, Category Norms and Word Frequency. *Bulletin of the Psychonomic Society*, 7:283–284.

Michelbacher, L., Evert, S., and Schütze, H. (2011). Asymmetry in Corpus-derived and Human Word Associations. *Corpus Linguistics and Linguistic Theory*, 7(2):245–276.

Miller, A. (2002). *Subset Selection in Regression*. Chapman & Hall, London, UK.

Miller, G. A., Leacock, C., Tengi, R., and Bunker, R. T. (1993). A Semantic Concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, pages 303–308, Stroudsburg, PA, USA. Association for Computational Linguistics.

Miller, K. J. (1998). Modifiers in WordNet. In Fellbaum, C., editor, *WordNet. An Electronic Lexical Database*, pages 47–67. MIT Press.

Miller, T., Biemann, C., Zesch, T., and Gurevych, I. (2012). Using Distributional Similarity for Lexical Expansion in Knowledge-based Word Sense Disambiguation. In *Proceedings of COLING*, pages 1781–1796.

Mintz, M., Bills, S., Snow, R., and Jurafsky, D. (2009). Distant Supervision for Relation Extraction Without Labeled Data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, ACL '09, pages 1003–1011, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mirkin, S., Dagan, I., and Geffet, M. (2006). Integrating Pattern-based and Distributional Similarity Methods for Lexical Entailment Acquisition. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics (COLING/ACL)*, pages 579–586, Sydney, Australia.

Mitchell, J. and Lapata, M. (2009). Language Models Based on Semantic Composition. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pages 430–439, Singapore.

Mitchell, J. and Lapata, M. (2010). Composition in Distributional Models of Semantics. *Cognitive Science*, 34:1388–1429.

Mitchell, T. M. (1997). *Machine Learning*. McGraw-Hill.

Miyao, Y. and Tsujii, J. (2009). Supervised Learning of a Probabilistic Lexicon of Verb Semantic Classes. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pages 1328–1337.

Mohammad, S., Dorr, B., and Hirst, G. (2008). Computing Word-Pair Antonymy. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 982–991, Honolulu, Hawaii. Association for Computational Linguistics.

Montague, R. (1970). Universal Grammar. *Theoria*, 36:373–398.

Montague, R. (1974). English as a Formal Language. In Thomason, R., editor, *Formal Philosophy*, pages 247–270.

Moro, A., Raganato, A., and Navigli, R. (2014). Entity Linking meets Word Sense Disambiguation. A Unified Approach. *Transactions of the Association of Computational Linguistics*, 2:231–244.

Navigli, R. (2009). Word Sense Disambiguation. A Survey. *ACM Computing Surveys*, 41(2):1–69.

Niles, I. (2003). Mapping WordNet to the SUMO Ontology. In *Proceedings of the IEEE International Knowledge Engineering Conference*, pages 23–26.

Niles, I. and Pease, A. (2001). Toward a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, pages 2–9.

Niles, I. and Pease, A. (2003). Linking Lexicons and Ontologies. Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of the IEEE International Conference on Information and Knowledge Engineering*, pages 412–416.

Nivre, J., Hall, J., Nilsson, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., and Marsi, E. (2007). MaltParser. A Language-independent System for Data-driven Dependency Parsing. *Natural Language Engineering*, 13:95–135.

Niwa, Y. and Nitta, Y. (1994). Co-occurrence Vectors from Corpora vs. Distance Vectors from Dictionaries. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING'94)*, pages 304–309.

Ó Séaghdha, D. (2010). Latent Variable Models of Selectional Preference. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 435–444, Uppsala, Sweden. Association for Computational Linguistics.

Padó, S. (2006). *User's Guide to `sigf`. Significance Testing by Approximate Randomisation*.

Padó, S. and Lapata, M. (2003). Constructing Semantic Space Models from Parsed Corpora. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 128–135, Sapporo, Japan. Association for Computational Linguistics.

Padó, S. and Lapata, M. (2007). Dependency-based Construction of Semantic Space Models. *Computational Linguistics*, 33:161–199.

Padó, S., Šnajder, J., and Zeller, B. (2013). Derivational Smoothing for Syntactic Distributional Semantics. In *Proceedings of ACL 2013*, pages 731–735, Sofia, Bulgaria.

Pandey, S. and Elliott, W. (2010). Suppressor Variables in Social Work Research. Ways to Identify in Multiple Regression Models. *Journal of the Society for Social Work and Research*, 1:28–40.

Pantel, P. and Lin, D. (2002). Discovering Word Senses from Text. In *Proceedings of ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 613–619.

Pantel, P. and Pennacchiotti, M. (2006). Espresso. Leveraging Generic Patterns for Automatically Harvesting Semantic Relations. In *Proceedings of COLING/ACL-06*, pages 113–120.

Pantel, P. and Pennacchiotti, M. (2008). Automatically Harvesting and Ontologizing Semantic Relations. In Buitelaar, P. and Cimiano, P., editors, *Ontology Learning and Population. Bridging the Gap between Text and Knowledge*. IOS Press.

Pantel, P. and Ravichandran, D. (2004). Automatically Labeling Semantic Classes. In *Proceedings of HLT-NAACL 2004*, pages 321–328, Boston, Massachusetts, USA. Association for Computational Linguistics.

Pasca, M. (2011). Attribute Extraction from Synthetic Web Search Queries. In *Proceeding of the 5th International Joint Conference on Natural Language Processing*, pages 401–409, Chiang Mai, Thailand.

Pearson, K. (1896). Mathematical Contributions to the Theory of Evolution. Regression, Heredity and Panmixia. *Philosophical Transactions of the Royal Society A*, 187:253–318.

Pease, A., Niles, I., and Li, J. (2002). The Suggested Upper Merged Ontology. A Large Ontology for the Semantic Web and its Applications. In *Working Notes of the AAAI-2002 Workshop on Ontologies and the Semantic Web*.

Pedersen, T., Patwardhan, S., and Michelizzi, J. (2004). WordNet::Similarity. Measuring the Relatedness of Concepts. In *Proceedings of the 5th Annual Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 38–41, Boston, MA.

Poesio, M. and Almuhareb, A. (2005). Identifying Concept Attributes Using a Classifier. In *Proceedings of the ACL Workshop on Frontiers in Corpus Annotation*, pages 76–83.

Pontiki, M., Galanis, D., Pavlopoulos, I., Papageorgiou, H., Androutsopoulos, I., and Manandhar, S. (2014). SemEval 2014 Task 4. Aspect Based Sentiment Analysis. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014) at COLING 2014*, pages 27–35, Dublin, Ireland.

Ponzetto, S. P. and Navigli, R. (2010). Knowledge-rich Word Sense Disambiguation Rivaling Supervised Systems. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics (ACL'10)*, pages 1522–1531.

Ponzetto, S. P. and Strube, M. (2006). Exploiting Semantic Role Labeling, WordNet and Wikipedia for Coreference Resolution. In *Proceedings of the Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics (HLT-NAACL 2006)*, pages 192–199.

Portner, P. (2005). *What is Meaning? Fundamentals of Formal Semantics*. Blackwell.

Prescher, D., Riezler, S., and Rooth, M. (2000). Using a Probabilistic Class-Based Lexicon for Lexical Ambiguity Resolution. In *Proceedings of the 18th COLING*, pages 649–655.

Probst, K., Ghani, R., Krema, M., Fano, A., and Liu, Y. (2007). Semi-supervised Learning of Attribute-Value Pairs from Product Descriptions. In *Proceedings of the International Joint Conference in Artificial Intelligence (IJCAI-07)*, pages 2838–2843.

Pulman, S. G. (2005). Lexical Decomposition. For and Against. In Tait, J. I., editor, *Charting a New Course. Natural Language Processing and Information Retrieval*, volume 16 of *The Kluwer International Series on Information Retrieval*, pages 155–173. Springer Netherlands.

Purandare, A. and Pedersen, T. (2004). Word Sense Discrimination by Clustering Contexts in Vector and Similarity Spaces. In *Proceedings of the Conference on Computational Natural Language Learning (CoNLL)*, pages 41–48.

Pustejovsky, J. (1995). *The Generative Lexicon*. MIT Press.

R Core Team (2013). *R. A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Rahman, A. and Ng, V. (2011). Coreference Resolution with World Knowledge. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (ACL HLT 2011)*, pages 814–824.

Ramage, D., Hall, D., Nallapati, R., and Manning, C. D. (2009). Labeled LDA. A Supervised Topic Model for Credit Attribution in Multi-labeled Corpora. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, Singapore, August 2009, pages 248–256.

Raskin, V. and Nirenburg, S. (1998). An Applied Ontological Semantic Microtheory of Adjective Meaning for Natural Language Processing. *Machine Translation*, 13:135–227.

Resnik, P. (1995). Using Information Content to Evaluate Semantic Similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 448–453, Montreal, Canada.

Riloff, E. and Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. In *Proceedings of the 16th National Conference on Artificial Intelligence (AAAI-99)*, pages 474–479.

Ritter, A., Mausam, and Etzioni, O. (2010). A Latent Dirichlet Allocation Method for Selectional Preferences. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 424–434, Uppsala, Sweden. Association for Computational Linguistics.

Rooth, M., Riezler, S., Prescher, D., Carroll, G., and Beil, F. (1999). Inducing a Semantically Annotated Lexicon via EM-based Clustering. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics*, ACL '99, pages 104–111, Stroudsburg, PA, USA. Association for Computational Linguistics.

Rosch, E. H. (1973). Natural Categories. *Cognitive Psychology*, 4:328–350.

Rosch, E. H. and Mervis, C. B. (1975). Family Resemblances. Studies in the internal structure of categories. *Cognitive Psychology*, 7:573–605.

Rothenhäusler, K. and Schütze, H. (2009). Unsupervised Classification with Dependency Based Word Spaces. In *Proceedings of the EACL Workshop on Geometrical Models of Natural Language Semantics (GEMS)*, pages 17–24.

Rozenfeld, B. and Feldman, R. (2006). High-Performance Unsupervised Relation Extraction from Large Corpora. In *Proceedings of the 6th International Conference on Data Mining (ICDM)*, pages 1032–1037.

Saba, W. S. (2007). Language, Logic and Ontology. Uncovering the Structure of Commonsense Knowledge. *International Journal of Human-Computer Studies*, 65(7):610–623.

Sahlgren, M. (2006). *The Word-Space Model. Using Distributional Analysis to Represent Syntagmatic and Paradigmatic Relations between Words in High-Dimensional Vector Spaces*. Ph.D. Dissertation, Department of Linguistics, Stockholm University.

Sahlgren, M. (2008). The Distributional Hypothesis. *Italian Journal of Linguistics*, 20:33–53.

Salton, G., Wang, A., and Yang, C. (1975). A Vector-Space Model for Information Retrieval. *Journal of the American Society for Information Science*, 18:613–620.

Scheffczyk, J., Pease, A., and Ellsworth, M. (2006). Linking FrameNet to the Suggested Upper Merged Ontology. In *Proceedings of the Fourth International Conference on Formal Ontology in Information Systems (FOIS 2006)*, pages 289–300, Amsterdam, The Netherlands. IOS Press.

Schulte im Walde, S. (2010). Comparing Computational Approaches to Selectional Preferences. Second-Order Co-Occurrence vs. Latent Semantic Clusters. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 1381–1388, Valletta, Malta.

Schulte im Walde, S., Melinger, A., Roth, M., and Weber, A. (2008). An Empirical Characterisation of Response Types in German Association Norms. *Research on Language and Computation*, 6(2):205–238.

Schütze, H. (1998). Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–124.

Schütze, H. and Pedersen, T. (1997). A Cooccurrence-based Thesaurus and two Applications to Information Retrieval. *Information Processing and Management*, 33:307–318.

Sekine, S. (2008). Extended Named Entity Ontology with Attribute Information. In *Proceedings of the 6th International Conference on Language Resources and Evaluation*, Marrakech, Morocco, pages 52–57.

Sekine, S. and Nobata, C. (2004). Definition, Dictionaries and Tagger for Extended Named Entity Hierarchy. In *Proceedings of the 4th International Conference on Language Resources and Evaluation*, pages 1977–1980, Lisbon, Portugal.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 27:379–423.

Sheinman, V., Fellbaum, C., Julien, I., Schulam, P., and Tokunaga, T. (2013). Large, Huge or Gigantic? Identifying and Encoding Intensity Relations among Adjectives in WordNet. *Language Resources and Evaluation*, pages 1–20.

Sheinman, V., Tokunaga, T., Julien, I., Schulam, P., and Fellbaum, C. (2012). Refining WordNet Adjective Dumbbells Using Intensity Relations. In *Proceedings of the 6th International Global Wordnet Conference*, pages 330–337.

Silberer, C. and Lapata, M. (2012). Grounded Models of Semantic Representation. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1423–1433, Jeju Island, Korea.

Socher, R., Perelygin, A., Wu, J., Chuang, J., Manning, C. D., Ng, A., and Potts, C. (2013). Recursive Deep Models for Semantic Compositionality Over a Sentiment Treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1631–1642, Seattle, Washington, USA. Association for Computational Linguistics.

Sowa, J. F. (2000). *Knowledge Representation. Logical, Philosophical, and Computational Foundations*. Brooks Cole.

Spearman, C. (1904). The Proof and Measurement of Association between Two Things. *The American Journal of Psychology*, 15:72–101.

Steyvers, M. and Griffiths, T. (2007). Probabilistic Topic Models. In Landauer, T., McNamara, D., Dennis, S., and Kintsch, W., editors, *Latent Semantic Analysis. A Road to Meaning*, pages 424–440. Lawrence Erlbaum Associates, Hillsday, NJ.

Tandon, N., de Melo, G., Suchanek, F., and Weikum, G. (2014). WebChild. Harvesting and Organizing Commonsense Knowledge from the Web. In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining*, pages 523–532, New York, NY, USA. ACM.

Thater, S., Fürstenau, H., and Pinkal, M. (2010). Contextualizing Semantic Representations Using Syntactically Enriched Vector Models. In *Proceedings of ACL 2010*, pages 948–957, Uppsala, Sweden.

Thater, S., Fürstenau, H., and Pinkal, M. (2011). Word Meaning in Context. A Simple and Effective Vector Model. In *Proceedings of IJCNLP 2011*, pages 1134–1143, Chiang Mai, Thailand.

Thill, S., Padó, S., and Ziemke, T. (2014). On the Importance of a Rich Embodiment in the Grounding of Concepts. Perspectives from Embodied Cognitive Science and Computational Linguistics. *Topics in Cognitive Science*, 6:545–558.

Tokunaga, K., Kazama, J., and Torisawa, K. (2005). Automatic Discovery of Attribute Words from Web Documents. In Dale, R., Wong, K.-F., Su, J., and Kwong, O. Y., editors, *Proceedings of the 2nd International Joint Conference on Natural Language Processing (IJCNLP-05)*, volume 3651 of *Lecture Notes in Computer Science*, pages 106–118. Springer.

Tsoumakas, G. and Katakis, I. (2007). Multi-label Classification. An Overview. *International Journal of Data Warehousing and Mining*, 3:1–13.

Turney, P. D. (2008). The Latent Relation Mapping Engine. Algorithm and Experiments. *Journal of Artificial Intelligence Research*, 33:615–655.

Turney, P. D., Neuman, Y., Assaf, D., and Cohen, Y. (2011). Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the 2012 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 680–690, Edinburgh, UK.

Turney, P. D. and Pantel, P. (2010). From Frequency to Meaning. Vector Space Models of Semantics. *Journal of Artificial Intelligence Research*, 37:141–188.

Unger, C., Bühmann, L., Lehmann, J., Ngonga Ngomo, A.-C., Gerber, D., and Cimiano, P. (2012). Template-based Question Answering over RDF Data. In *Proceedings of the 21st International Conference on World Wide Web (WWW)*, pages 639–648. ACM.

Vecchi, E. M., Baroni, M., and Zamparelli, R. (2011). (Linear) Maps of the Impossible. Capturing Semantic Anomalies in Distributional Space. In *Proceedings of the ACL Workshop on Distributional Semantics and Compositionality*, pages 1–9. Association for Computational Linguistics.

Walter, S., Unger, C., Cimiano, P., and Lanser, B. (2014). Automatic Acquisition of Adjective Lexicalizations of Restriction Classes. In *Proceedings of NLP&DBpedia Workshop at ISWC 2014*, Riva del Garda, Italy.

Wang, Q. I., Schuurmans, D., and Lin, D. (2005). Strictly Lexical Dependency Parsing. In *Proceedings of the Ninth International Workshop on Parsing Technology*, pages 152–159, Stroudsburg, PA, USA. Association for Computational Linguistics.

Wang, W., Besançon, R., Ferret, O., and Grau, B. (2011). Filtering and Clustering Relations for Unsupervised Information Extraction in Open Domain. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, CIKM '11, pages 1405–1414, New York, NY, USA. ACM.

Weeds, J., Weir, D., and McCarthy, D. (2004). Characterising Measures of Lexical Distributional Similarity. In *Proceedings of the 20th International Conference on Computational Linguistics (COLING'04)*, pages 1015–1021. Association for Computational Linguistics.

Widdows, D. (2004). *Geometry and Meaning*. CSLI Publications, Stanford, CA.

Widdows, D. (2008). Semantic Vector Products. Some Initial Investigations. In *Proceedings of the 2nd Conference on Quantum Interaction*, Oxford, UK.

Wierzbicka, A. (1996). *Semantics. Primes and Universals*. Oxford University Press.

Witten, I. H. and Frank, E. (2005). *Data Mining. Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, California, 2nd edition.

Wittgenstein, L. (2001). *Philosophische Untersuchungen. Kritisch-genetische Edition*. Wissenschaftliche Buchgesellschaft.

Wu, F., Hoffmann, R., and Weld, D. S. (2008). Information Extraction from Wikipedia. Moving Down the Long Tail. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 731–739.

Yarlett, D. G. (2008). *Similarity-based Generalization in Language*. PhD thesis, Stanford University, CA.

Yeh, A. (2000). More Accurate Tests for the Statistical Significance of Result Differences. pages 947–953.

Yianilos, P. N. (1993). Data Structures and Algorithms for Nearest Neighbor Search in General Metric Spaces. In *Proceedings of the Fourth Annual ACM/SIGACT-SIAM Symposium on Discrete Algorithms (SODA)*.

Yoshinaga, N. and Torisawa, K. (2007). Open-Domain Attribute-Value Acquisition from Semi-Structured Texts. In *Proceedings of the ISWC Workshop "From Text to Knowledge: The Lexicon/Ontology Interface" (OntoLex07)*, pages 55–66, Busan, South Korea.

Zhang, J., Salwen, J., Glass, M., and Gliozzo, A. (2014). Word Semantic Representations using Bayesian Probabilistic Tensor Factorization. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1522–1531, Doha, Qatar. Association for Computational Linguistics.

Zhitomirsky-Geffet, M. and Dagan, I. (2009). Bootstrapping Distributional Feature Vector Quality. *Computational Linguistics*, 35(4):435–461.