A Domain Adaptive Approach to Automatic Acquisition of Domain Relevant Terms and their Relations with Bootstrapping

Feiyu Xu*, Daniela Kurz¤, Jakub Piskorski*, Sven Schmeier¤

* DFKI – German Research Center for Artificial Intelligence
Stuhlsatzenhausweg 3, 66 123 Saarbrücken, Germany
{feiyu, piskosrk}@dfki.de

¤ XtraMind GmbH
Stuhlsatzenhausweg 3, 66 123 Saarbrücken, Germany
{kurz, schmeier}@xtramind.com

Abstract

In this paper, we present an unsupervised hybrid text-mining approach to automatic acquisition of domain relevant terms and their relations. We deploy the TFIDF-based term classification method to acquire domain relevant single-word terms. Further, we apply two strategies in order to learn lexico-syntatic patterns which indicate paradigmatic and domain relevant syntagmatic relations between the extracted terms. The first one uses an existing ontology as initial knowledge for learning lexico-syntactic patterns, while the second is based on different collocation acquisition methods to deal with the free-word order languages like German. This domain-adaptive method yields good results even when trained on relatively small training corpora. It can be applied to different real-world applications, which need domain-relevant ontology, for example, information extraction, information retrieval or text classification.

1. Introduction

Recent trends in information technology such as Text Mining (TM) provide dramatic improvement in the conversion of the overflow of raw textual data into structured knowledge for solving more complex real-world knowledge discovery tasks. Text mining concerns the discovery of useful and previously unknown information from unstructured free text (Feldmann, 1999) and it is strongly related to data mining (DM), natural language processing (NLP), machine learning (ML), information extraction (IE) and information retrieval (IR).

Mining terms and their relations from real-world free texts is attracting increasing attention, for example, the domain adaptation capability of IE systems relies on automatic acquisition of domain ontology and lexico-syntactic patterns for template filling (Riloff & Jones, 1999; Yangarber et. al., 2000). Recently, an ever-growing interest in automatic term and term collocation extraction methods in NLP (Church & Hanks, 1989; Smadja, 1994; Daille, 1996; Evert & Krenn, 2001), knowledge discovery (Hearst, 1992) and IR (Salton, 1991) has been observed. Landau-Finkelstein & Morin (1999) benefit from these approaches in IE.

In this paper, we present a hybrid approach to automatic acquisition of domain ontology. Compared to other supervised or weakly supervised approaches that use a handful initial "seed words" or "seed lexicon syntactic patterns" (Hearst, 1992; Hearst, 1998; Riloff, 1999; Yangarber et. al., 2000), the input of the presented method consists solely of a collection of classified documents. Our method is based on the integration of shallow parsing results, existing general ontology and statistical measures. It turned out that very good results may be achieved independent of the size of the training corpus. In particular, we can handle free word-order languages like German using specific term collocation techniques. We make use of TFIDF-based term classification method to identify the domain relevant single-word terms. In contrast to general ontologies, the presented approach allows for extracting not only strict paradigmatic relations but also near synonymy relations (Inpken & Hirst, 2001) crucial for solving real-world IE tasks.

For the linguistic annotation (stemming, morphological decomposition, pos-tagging, named-entity and phrase recognition) of the corpus, we use SPPC (Piskorski & Neumann, 2000). For accessing the semantic relations in GermaNet (Hamp & Feldweg, 1997), we integrated an ontology inference machine (Siegel et. al., 2001). For evaluation of our approach, three domains were chosen from German press texts from DPA (1999 and 2000): management succession, stock market and drug domain.

The remainder of this paper is organized as follows. A brief introduction of the Word/GermaNet inference machine is given in section 2. In section 3, we shortly describe our shallow processing system SPPC. A detailed description of our approach is presented in section 4. Finally, we summarize and outline the future work in section 5.

2. The Ontology Inference Machine

The lexical-semantic information encoded in online ontologies like WordNet (Miller et. al., 1993), GermaNet (Hamp & Feldweg, 1997) and EuroWordNets (Vossen ,

1998) provides valuable knowledge base which can be used in various natural language applications: IE, ontology acquisition and intelligent IR. The ontology inference machine was developed to enable search for relations in the WordNet and GermaNet (Siegel et. al., 2001). In our approach, we make use of GermaNet as our general ontology to learn lexico-syntactic patterns which indicate hyponymy and synonymy relations.

2.1. GermaNet

Compared to the huge amount of online English linguistic resources, there are not as many large-scale German lexica like GermaNet which have properly modelled lexical syntactic and semantic information. Therefore, GermaNet appears to us as a valuable resource to extend our lexicon.

GermaNet is a lexical semantic net for German, developed at the university of Tübingen. It is based mainly on the WordNet framework, containing 10.652 nouns, 6.904 verbs and 1.657 adjectives. One big advantage of GermaNet is that the semantic classification of the words is very fine grained. Like in WordNet, a semantic concept (so-called *synset*) is represented by a group of words. There are 19.213 synsets in GermaNet and in addition 24.920 synonyms in synsets. The synsets are connected through their lexical and conceptual relations. The basic lexical relations are *synonymy*, *antonymy* and *pertains to*, while the conceptual relations are *hyponymy* ('is-a'), *meronymy* ('has-a'), *entailment* and *cause*. The hyponymy relation information constitutes a hierarchical semantic structure of GermaNet. Compared to WordNet, verbs in GermaNet are annotated additionally with *selectional restrictions*, which are important for the deep natural language processing.

2.2. Inference Tool

GermaNet itself provides a simple search interface that allows searches for the relations assigned to a word. However, this search interface is still too restricted to be directly usable for different applications. The ontology inference machine provides three different functions:

- Retrieval of relations assigned to a word
- Retrieval of relations between two words
- Flexible navigation in the GermaNet graph starting from a certain node with search depth and search relationship as arguments

The first search function is actually a reimplementation of the search interface existing in GermaNet. An example of a query is 'find all synonyms of the German word *Bank*'. For the first sense *bench*, we find the word *Sitzmöbel* (engl. sitting furniture) as its synonym. For the sense corresponding to *financial institution*, its synomyms are *Geldinstitut* (engl. money institution) and *wirtschaftliche Institution* (engl. financial institution).

The second type of functions is to search for and test the relations between two words. This search type provides important information like 'is-a' and 'has-a' relation between words, which supports the coreference resolution between terms in the information extraction application. Let us give a simple example. We would like to know the relationship between the word *Internet-Service-Provider* and the word *Firma* (engl. company). Our search tool tells us that the *Internet-Service-Provider* is a hyponym of the word *Firma*. It indicates that the first word is a subconcept of the second one.

3. Shallow NLP Processing

SPPC (Shallow Processing Production Center) is an advanced domain independent extraction and navigation core system for processing German free-text documents (Piskorski & Neumann, 2000). It consists of a set of shallow processing components including, among others, fine-grained tokenization, lexical analysis including online compound decomposition, part-of-speech filtering, named-entity recognition, sentence boundary detection, chunk and subclause recognition. SPPC is capable of processing vast amount of textual data robustly and efficiently[1] since all subcomponents of the system were realized by means of cascaded optimized weighted finite-state devices. Due to the sophisticated linguistic knowledge, the system achieves good linguistic coverage on all levels of processing. The following components of SPPC were used for the linguistic preprocessing of the input data:

Tokenizer maps sequences of consecutive characters into word-like units, usually called tokens and classifies them according to fine-grained token class definitions (e.g., two digit number, First capital word, mixed word, candidate for abbreviation, number-word compositum). Overall there are currently over fifty default token classes and it proved that such variety simplifies processing on higher stages (e.g., definition of named-entity recognition patterns).

Lexical Processor processes each token identified as a potential word form, and tries to associate it with its corresponding lexical information. Further, it performs online compound recognition[2] (e.g., *Forschungsausgaben* engl. research expenses) and resolves compound coordination (e.g., *Produktionsumstellungen oder –erweiterungen* engl. production reorganization and expansion), which occur frequently in our test corpora. The sole resource used for retrieving lexical information is a full-form lexicon currently containing 750 000 entries.

Part-of-Speech Filtering performs word-based part-of-speech disambiguation based on three type of filtering rules: (a) case-sensitive rules, (b) contextual filtering rules (Brill, 1992) and (c) rules for filtering out rare readings.

Named-entity Finder identifies proper names (organizations, persons, locations), temporal expressions (time, date) and quantities (monetary values, percentages, numbers). This is primarily done by using simple pattern-

[1] circa 30000 words per second in standard PC environment

[2] Since compounding is very productive process in German, proper recognition of compounds is crucial task. SPPC achieves lexical coverage of 95% on unseen text and the accuracy of the compound recognition based is nearly 100%.

matching techniques since they can be easily identified because of the specific context they appear in (e.g., company designator). Additionally, a dynamic lexicon is used for proper identification of abbreviated variants of previously recognized named entities (e.g., company name appearing without designators) and acronyms. In this way, this component performs partial coreference resolution. SPPC achieves very good coverage in named entity recognition in the financial domain[3], which is an essential factor for performing successfully our mining task.

Chunk recognizer extracts text fragments which constitute nominal and prepositional phrases and verb clusters. The recognition of verb groups is only partial since in German a verb group may be split into a left and right part so that other phrases are spliced into the splitting point. Furthermore, fine-grained classification of recognized verb clusters is provided.

4. Mining Terms and their Relations

In this section we describe the core approach for detection of relevant domain terms and for learning relations, which hold among them. Our extraction engine comprises of three main components:

A. TFIDF-based single-word term classifier
B. Lexico-syntactic pattern Finder
  B.1 Learns the patterns based on the set of the known relations (initialized with GermaNet or WordNet)
  B.2 Learns the patterns based on term collocation methods
C. Relation Extractor which uses found lexico-syntactic patterns

The architecture of the system is depicted in figure 1.

Our bootstrapping algorithm works as follows:

Input: classified documents enriched with linguistic information computed by SPPC

Step1: extract single-word terms using (A)
Step2: learn multi-word terms and identify the lexico-syntactic patterns using (B.2)
Step3: learn patterns using (B.1)
Step4: extract related terms via the application of learned lexico-syntactic patterns to the corpus using (C)
Step5: go to step 3 with extracted new term relations

4.1. Mining Relevant Terms

Before mining term relations, the first step is to discover domain relevant terms. For fulfilling this task we apply specific TF-IDF measure (Salton, 1992), called KF-IDF, which is suitable when working with categorised documents.

The KF-IDF is defined as follows:

$$\boldsymbol{KFIDF\ (w, cat)\ ?\ docs(w, cat)\ ?\ LOG(\frac{n\ ?\ |cats|}{cats(word)}\ ?\ 1)}$$

docs(w, cat) = number of documents in the category cat containing the word w
n = smoothing factor
cats(word) = the number of categories in which the word occurs

According to this formula the KFIDF measure for a word grows logarithmic inversely proportional to the number of categories it occurs in. In other words, a term is regarded as relevant if it occurs more frequently than other words in a certain category, but occasionally elsewhere. In our approach, only adjectives, nouns and verbs are considered as potential term candidates.

We have conducted several experiments on using this measure for mining terms in document collections taken from DPA (Deutsche Presse Agentur) in three domains: management succession, stock market and crime-drug domain. An interesting phenomenon was observed. The distribution of the relevant terms concerning the part-of-speech information is domain dependent. In some domains, the most relevant terms are nouns, for example, the crime drug domain and the stock market domain, while in some domains like management succession, the relevant terms are verbs. For example, 1) illustrates the top ten noun terms in the drug domain. Prominent drug sorts and their related terms could have been detected correctly.

1)
Haschisch 79.13055
Droge 55.192017
Marihuana 55.151592
Rauschgift 53.61485
Kilogramm 52.038185
Marktwert 51.142445
Heroin 48.095898
Kokain 44.153614
Schwarzmarktwert 40.913956
Konsument 32.390213
Ecstasy-Tabletten 28.774744

2) shows an example of the top ten noun terms extracted from the stock-market document collection, where the terms below reflect the elements in the stock market.

[3] precision of almost 96% and recall of 85 %

2)

Aktienboerse	*237.05634*
Veraenderung	*143.48146*
Gewinner	*142.09517*
Verlierer	*142.09517*
Hochtief	*88.72284*
Tief	*88.72284*
Kugelfischer	*80.405075*
Carbon	*70.70101*
Aktie	*53.796547*
Kurs	*49.768997*

In contrast to the above two domains, the management succession was determined mainly by the verbs which indicate the change of employment in the company managements.

berufen	*38.45143*
waehlen	*35.155594*
uebernehmen	*32.95837*
bestellen	*28.56392*
verlassen	*20.873634*
wechseln	*19.77502*
ausscheiden	*17.577797*
nachfolgen	*15.380572*
zuruecktreten	*12.084735*
antreten	*8.788898*

4.2. Learning relations with lexico-syntactic patterns

Inspired by Hearst (1992) and Landau-Finkelstein and Morin (1999), we learn lexico-syntactic patterns indicating paradigmatic relations. Instead of using initial seeds of patterns, we employ the existing semantics relations provided GermaNet (Hamp & Feldweg, 1997)

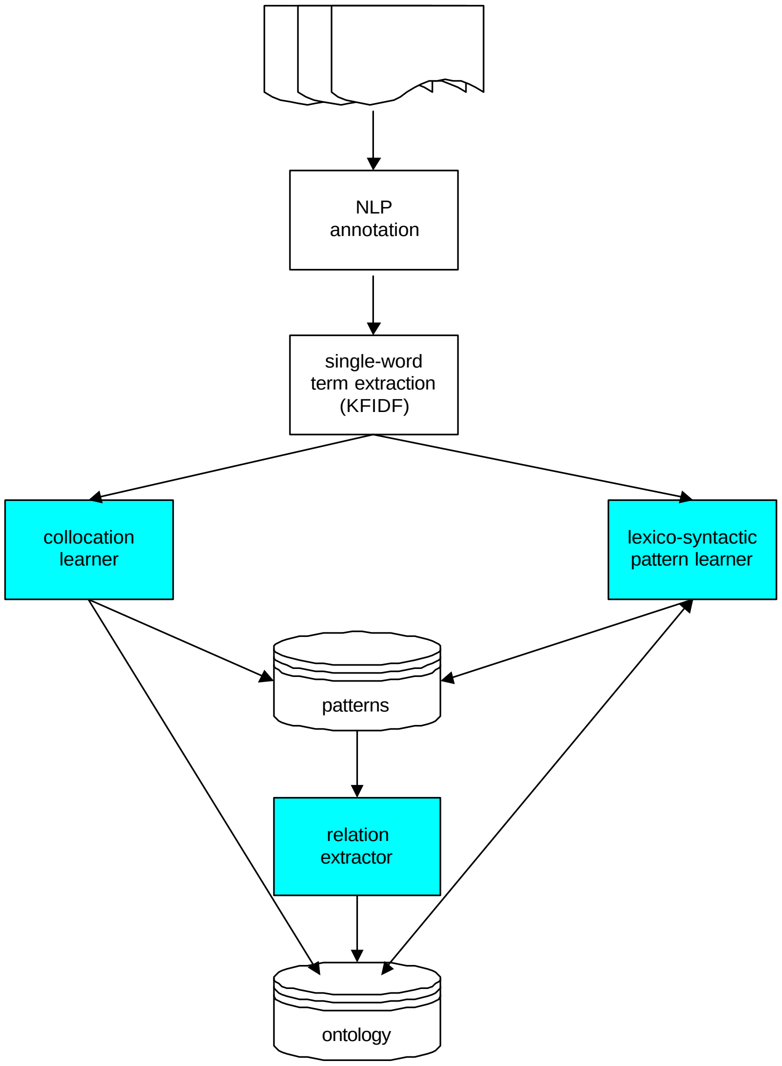
NLP annotation
single-word term extraction (KFIDF)
collocation learner
lexico-syntactic pattern learner
patterns
relation extractor
ontology

Figure 1: System Architecture

to assign synonymy, hyponymy and meronymy relations between the terms extracted from the corpus. Secondly, we extract the text fragments containing these semantic relations. Subsequently, we use the algorithm presented in (Landau-Finkelstein & Morin, 1999) for clustering similar patterns. Finally, two groups of patterns are identified: domain independent patterns and domain specific patterns. Domain specific patterns define reliable domain specific relations.

3)
Drogen *wie* LIST_of_NPS
Drogen *sowie* LIST_of_NPS

The above patterns indicate that each single NP in list of noun phrases (LIST_of_NPS) is a hyponym of *Drogen* (engl. drug), for example, *Drogen wie Cannabis-Produkten*, where the *Cannabis-Produkt* is a hyponym of *Drogen*. The general lexico-syntactic patterns could be defined as follows:

4)
NP, NP, …, NP, NP und andere N
NP bzw. NP
NP sowie NP
NP wie NP

After the term relation extractor has applied the newly extracted lexico-syntactic patterns, we obtain a list of related terms, which have potentially hyponymy relations among them, for example *5)*, the three NPs in the LIST_of_NPS are hyponyms of the term *smuggling countries*. These hyponymy relations are very domain specific.

5) Schmuggelländer wie [Niederlande, Türkei und Ungarn]$_{LIST\_of\_NPS}$

In many cases, we have observed that many term groups do not have strict hyponym or synonym relations, for example,

6) Kokain sowie Haschisch, LSD und Syntheseprodukt Schlafstoerung und Verfolgungswahn

Most of them are near synonyms (Hirst, 1995; Inkpen & Hirst, 2000). Near synonyms are words that are almost synonyms, playing the same semantic role in a domain. They share mostly a same supper concept. In order to identify their common supper concept, we use GermaNet to search for their shared hypernyms. Afterwards, we assign the found hypernyms to the rest terms which are not encoded in the GermaNet. The advantage of this method is that we can assign the new terms into the domain hierarchy and at the same time we have disambiguated the senses of the terms in this domain. For example 7), *Kokain* and *Haschisch* share the same supper concept *Droge* in GermaNet, therefore, we assign *Droge* as the supper concept of *LSD* and *Syntheseprodukt*. On the one hand we have obtained new drug sorts and on the other hand we have identified the senses of *LSD* and *Syntheseprodukt* in the drug domain, because *LSD* and *Syntheseprodukt* may have another senses in other domains. Many real-word applications, in particular, IE, typically require relatedness rather then just similarity, for example, we have also found the following related terms in the drug crime domain

7)
a. *Polizei, Zoll, Landeskriminalamt*
b. *Schlaflosigkeit, Halluzinationen, Verfolgungswahn*
c. *Polizei, Drogenhilfe, Sozialarbeiter*

These clusters of terms correspond to special semantic concepts in the drug domain, 7a) the concept "government institutions against drug traffic", 7b) to the concept "side effects of drug use" and 7c) to the concept "helper organizations for drug addicts".

4.3. Learning term collocations

The objective of term collocation in our approach is on the one hand the identification of multi-word terms and on the other hand learning lexico-syntactic patterns for extraction of semantic relations. In contrast to (Smadja, 1994), we also consider semantically related words as (Church & Hanks, 1989) do, in addition to so-called true collocations. (Church & Hanks, 1989) provide a statistical measure to capture phenomena ranging from semantic relations between *banker* and *trust* to lexico-syntactic co-occurrences like *succeed a person*. (Daille, 1996) claimed that a purely frequency-based measures deliver good results for technical domains. However, the drawback of frequency oriented approach is that bad candidates can not be excluded. Therefore, they preferred Log-Likelihood Measure, which takes into account the pair frequency but accepts very little noise for high values.

Due to the free word-order characteristic of German, it is not sufficient to take into account only bigrams, trigrams etc. as applied in the above approaches. Thus, we considered all possible term pairs in a sentence ignoring the linear order. We used following association measures: Mutual Information (Church & Hanks, 1989), Log-Likelihood Measures (Daille, 1996), and T-test (Manning & Schütze, 1999). Let us give a short explanation of the different measures:

Assocation measures

Mutual Information is defined as follows:

$$\mathrm{I(x,y)} = \frac{\log_2 P(x,y)}{P(x)P(y)}$$

where *P(x,y)* denotes the joint probability and *P(x)* and *P(y)* denote the probability of *x* and *y* separately.

This association measure assumes that the occurrence of one word predicts the occurrence of another one. If there is an interesting relationship between *x* and *y*, the mutual information is expected to increase. We observed as mentioned in (Manning & Schütze, 1999) that mutual information is not practical when dealing with sparse data.

The definition of **Log-Likelihood** is given bellow:

LogLike(x, y) =

$$\begin{aligned} &a\log a ? b\log b ? c\log c ? d\log d\\ &? (a ? b)\log(a ? b) ? (a ? c)\log(a ? c)\\ &? (b ? d)\log(b ? d) ? (c ? d)\log(c ? d)\\ &? (a ? b ? c ? d)\log(a ? b ? c ? d)\end{aligned}$$

with *a, b, c* and *d* being elements of the contingency table of words *x* and *y* occurring with each other or not, e.g. *a* stands for the frequency of pairs involving both *x* and *y* etc. This measure tells us how much more likely the occurrence of one pair is than the occurrence of another one.

T-test is defined as:

$$T = \frac{x ? ?}{\sqrt{\frac{s^2}{N}}}$$

where *x* denotes the sample mean, $?$ the mean of the distribution, s^2 the sample variance, *N* the sample size. This test tells how probable or improbable it is that a certain constellation occurs. The null hypothesis assumes that the occurrence of the two terms is independent. The T-test value tells us, if this hypothesis can be rejected or not.

Results

We focused on the extraction of noun-noun, verb-noun and adj-noun combinations. By looking at the precision values of the statistical measures, we can confirm the results from other studies (Krenn & Evert, 2001) suggesting that LogLike delivers the best precision values for low-frequency data. Moreover, they could show that the ranking of the association measure depends on the kind of collocation to be identified: the T-test delivers better results for preposition-noun-verb combinations, whereas the Log-Likelihood measure leads to significantly better results for Adjective-Noun combinations.

Since we worked on corpora of extremely small size, it can be expected that LogLike works best. It turned out that our method performs reasonably well. We evaluated four corpora of different size and different domains: drugs, stock market, running amok and management succession. The smallest corpus contains 6361 tokens, the biggest one contains 84747. The main observation we could make are the following:

1) There is a correlation between corpus size and precision. The bigger the corpus the more collocations could be correctly identified.

   Table 1 shows the precison values for the 200 highest-ranked words in a corpus applying Log-Likelihood for computing Noun-Verb collocations.

2) For both combinations Noun-Noun collocations and Noun-Verb collocations LogLike compared to Mutual Information and T-Test delivers the best results. A comparison between LogLike and Mutual Information for Noun-Verb collocations is shown in Table 1.

3) We could not observe a dominance of a certain collocation type depending on a certain domain. In each domain Noun-Verb collocation were most prominent and delivered best results. In the drugs domain we obtained a precision of 56% for Noun-Verb collocations. The precision for Noun-Noun collocations is only 41%.

Size of corpus	LogLike (Noun-Verb)	Mutual Information (Noun-Verb)
6361 tokens	52%	34%
29143 tokens	56%	42%
59134 tokens	63%	36%
84747 tokens	61%	49%

Table 1: Precision values for corpora of different size

The extracted collocations can expand the set of already learned patterns for bootstrapping, for example, the noun-noun combination in *7)* helps to find more hyponyms of 'Droge'.

8) Kilogramm <NP_drug>

Further, they indicate semantic relations for learning new lexico-syntactic patterns.

?? ***Hyponymy****: Arzneimittel, Medizinprodukte*
?? ***Hyponymy****: Reparatur, Wartung*

Additionally, they are often multi word terms:

9)

a. Frankfurter, Flughafen
b. Industrie, Handelskammer
c. Volksrepublik, China

Further, the verb-noun combinations can be used to enhance existing subcategorization lexicons and may also constitute candidates for template filling rules.

10)

a. sitzen, Untersuchungshaft
b. treten, Ruhestand
c. Leitung übernehmen

5. Conclusion

In this paper we have presented an unsupervised and domain adaptive approach to automatic extraction of domain relevant terms and relations among them. The KF-IDF based term extraction has proved to be very promising for the extraction of single word terms. We have combined two methods to acquire the patterns for identifying related terms: (a) using ontology (GermaNet), (b) using different statistical measures. The latter one proves to be suitable for handling the free-word order languages like German In the near future work, we will attempt to use clustering methods for discovering new relations.

6. References

Brill, B. (1992). A Simple Rule-Based Part-of-Speech Tagger. *In Proceedings of the Third Conference on Applied Computational Linguistics (ACL)*, Trento, Italy, 1992.

Church, K. W. & Hanks, P. (1989). Word association norms, mutual information and lexicography. *In Proceedings of the 27th Annual Conference of the Association of Computational Linguistics, pages 76-82, 1989*.

Daille, B. (1996). Study and Implementation of Combined Techniques of Automatic Extraction of Terminology. *In J.L. Klavans and P. Resnik, editors, The Balancing Act: Combining Symbolic and Statistical Approaches to Language, pages 49-66*, MIT Press, Cambridge, MA.

Evert, S. & Krenn, B. (2001). Methods for the Qualitative Evaluation of Lexical Association Measures. *In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics* Toulouse, France.

Finkelstein-Landau, M. & Morin, E. (1999). Extracting Semantic Relationships between Terms: Supervised vs. Unsupervised Methods. *In Proceedings of International Workshop on Ontological Engineering on the Global Information Infrastructure, pages 71-80,* Dagstuhl Castle, Germany, May 1999.

Hamp, B. & Feldweg, H. (1997). GermaNet - a Lexical-Semantic Net for German. *In Proceedings of ACL workshop Automatic Information Extraction and Building of Lexical Semantic Resources for NLP Applications*, Madrid, 1997.

Hearst, M. A.(1992). Automatic Acquisition of Hyponyms from Large Text Corpora. *In Proceedings of the Fourteenth International Conference on Computational Linguistics*, Nantes, France, July 1992.

Inkpen, D. Z. & Graeme (2001). Building a Lexical Knowledge-Base of Near-Synonym Differences. *In Proceedings of Workshop on WordNet and Other Lexical Resources (NAACL 2001), Pittsburgh, pages 47-52, June 2001*.

Manning, C. & Schütze, H. (1999). Foundations of Statistical Natural Language Processing. *MIT Press, Cambridge, MA.*

Piskorski, J. & Neumann, G. (2000). An Intelligent Text Extraction and Navigation System. *In Proceedings of 6th International Conference on Computer-Assisted Information Retrieval (RIAO-2000),* Paris, 2000

Riloff, E. & Jones, R. (1999). Learning Dictionaries for Information Extraction by Multi-Level Bootstrapping. *In Proceedings of the Sixteenth National Conference on Artificial Intelligence (AAAI-99)* , 1999, pp. 474-479.

Salton, G. (1991). Developments in Automatic Text Retrieval. *Science, Vol 253, pages 974-979, 1991.*

Smadja, G. (1994). Retrieving Collocations from Text: Xtract. *Computational Linguistics 19(1): 143-177 (1994).*

Yangarber R., Grishman R., Tapanainen P. and Huttunen , S. (2000). Automatic Acquisition of Domain Knowledge for Information Extraction. *In Proceedings of COLING 2000: The 18th International Conference on Computational Linguistics*, August 2000, Saarbrücken, Germany.