The Termolator: Terminology Recognition based on Chunking, Statistical and Search-based Scores[1]

Adam Meyers[1], Yifan He[2], Zachary Glass[3] and Olga Babko-Malaya[4]

[1] *meyers@cs.nyu.edu*
New York University, Dept. of Computer Science, 715 Broadway, NY, NY (USA)

[2] *yhe@cs.nyu.edu*
New York University, Dept. of Computer Science, 715 Broadway, NY, NY (USA)

[3]*zglass@alumni.princeton.edu*
New York University, Dept. of Computer Science, 715 Broadway, NY, NY (USA)

[4]*olga.babko-malaya@baesystems.com*
BAE Systems, 6 New England Executive Park, Burlington, MA, (USA)

Abstract

The Termolator is a high-performing terminology extraction system, which will soon be available as open source software. The Termolator combines several different approaches to get superior coverage and accuracy. The system identifies potential instances of terminology using a chunking procedure, similar to noun group chunking, but favoring chunks that contain out-of-vocabulary words, nominalizations, technical adjectives, and other specialized word classes. The system ranks such term chunks according to several metrics including: (a) a set of metrics that favors term chunks that are relatively more frequent in a "foreground" corpus about a single topic than they are in a "background" or multi-topic corpus and (b) a relevance score which measures how often terms appear in articles and patents in a Yahoo web search. We analyse the contributions made by each of these metrics and show that all modules contribute to the system's performance, both in terms of the number and quality of terms identified.

Workshop Topic

Terminology Extraction

Introduction

Automatic terminology extraction systems aim to collect word sequences to be used as Information Retrieval key words, terms to be included in domain-specific glossaries or ontologies. Terms are also potential arguments of information extraction relations or entities to be tracked for technology forecasting applications. This paper describes ***the Termolator***, a terminology extraction system which will soon be released as open source software. The Termolator selects the terms (scientific noun sequences) that are characteristic of a particular technical area. The system identifies all instances of terms in sets of files using a sequential pattern matching process called chunking. It is similar to the noun group chunkers used in many natural language processing systems, but adds additional constraints so that the noun group chunks must contain words belonging to specialized vocabulary classes including: out-of-vocabulary words, nominalizations, technical adjectives, and others. To find chunks that are characteristic of a topic, the system compares the frequencies of particular terms in 2 sets of documents: the foreground corpus (documents about a single topic) and the background corpus (documents about a mixture of topics). It uses several statistical measures to make this determination including Document Relevance Document Consensus or DRDC (Navigli and Velardi**,** 2004), Term Frequency-Inverse Document Frequency (TFIDF) and Kullback-Leibler

[1] Approved for public release; unlimited distribution

Divergence or KLD (Cover and Thomas, 1991; Hisamitsu et al., 1999). For each foreground set of documents, the system produces a list of terms, which is initially ordered based on the distributional means just described. Two other types of scores are factored in to the system's ranking: a well-formedness score based on linguistic constraints, and a relevance score, based on how often a Yahoo (https://search.yahoo.com) web-search results for that term point to patents or articles. The final ranking is used to extract the top terms. We have found that given about 5000 foreground documents and 5000 background documents, we can generate about 5000 terms that are approximately 85% accurate. The system has been tested and scored on US patents and Web of Science abstracts. We have also performed preliminary tests on English journal articles (PubMed Central corpus, http://www.ncbi.nlm.nih.gov/pmc/). We have implemented some of the components of a Chinese version of the system and have plans to continue development.

System Description

System Overview

Our system consists of three stages: terminological chunking and abbreviation, distributional ranking and filtering. The first stage identifies instances of potential terms in text. The second stage orders the terms according to their relative distribution in the foreground and background corpora. The final stage reorders the top N terms from the second stage based on a well-formedness metric and a relevance metric. The assumption behind the ranking is that the higher ranked terms are preferred over lower ranked ones in three respects: 1) higher ranked terms are less likely to be errors, noun sequences that are not really instances of terminology because they are ill-formed as noun groups or represent phrases that are part of the general vocabulary, rather than specialized vocabulary; 2) higher ranking terms tend to be more characteristic of a particular field of interest than lower ranking terms; and 3) higher ranking terms tend to have greater relevance than the low ranking ones, i.e., specialists and others are currently more interested in the concepts represented by the high ranking terms.

Stage 1: Terminological Chunking and Abbreviation

In Meyers, et. al. (2014a), we describe the component of our system designed for identifying terms in sentences, independent of their distribution in sets of documents. Like Justeson and Katz (1995), we assume that most instances of terminology are noun groups, head nouns and pre-modifiers other than determiners. Consequently, we currently exclude non-noun instances of terminology (verbs like *calcify* or *coactivate*; adjectives like *covalent* or *model theoretic* and adverbs like *deterministically* or *stochastically*). Unlike previous approaches, we consider only a subset of noun groups as we adapt a more stringent set of chunking rules than used for standard noun group detection. We also identify an additional set of terms by means of rules for identifying abbreviations.

We incorporate into our chunking rules requirements that constituents contain nominalizations, out of vocabulary words, technical adjectives and other classes of a more fine-grained nature than typical parts of speech used in noun chunking. Nominalizations, such as *amplification* and *radiation* are identified and classified using NOMLEX_PLUS (Macleod, et. Al. 1998 and Meyers, et. al. 2004), contributing to the ranking of the terms *optical amplification medium fiber* and *optical radiation*). Out of vocabulary words (e.g., *photoconductor* and *collimate*) are words not found in COMLEX Syntax (Macleod, et. Al. 1997) or classified as names (thus selecting terms like *electrophotographic photoconductor*

and *optical collimate*). Technical adjectives are adjectives found in COMLEX or classified by a POS tagger that end in *-ic, -cal*, or *–ous*, but are not part of a manually selected outlist (e.g., *public, jealous*).[2] The chunking component is modelled as a finite state machine using a fine-grained set of parts of speech (FPOS) to determine transitions between Beginning, Ending, Internal and Other states in the style of Ramshaw and Marcus (1995). The FPOS include nominalizations, technical adjectives and Out of Vocabulary (OOV) words as defined above, as well as several other categories such as nationalities (the adjectival form of a country, state, city or continent, e.g., *European, Indian*, and *Peruvian*); adjectives or nouns with the first letter capitalized, person names, and roman numerals. The technical noun chunks are sequences of these categories, omitting preceding determiners, normal adjectives and other words that were not likely to be parts of instances of terminology.[3]

We extract instances of abbreviations and full forms, using pattern matching similar to Schwartz and Hearst (2003) in contexts where a full form/abbreviation pair are separated by an open parentheses, e.g., *Hypertext Markup Language (HTML).* In the simplest case, the abbreviation consists of the initials for each word of the full form, but variations in which words are skipped, multiple letters match a single word, etc. are incorporated as well. Keyword-based heuristics and gazetteers are used to differentiate non-terminology abbreviation cases from terminology ones, e.g., New York University, Acbel Polytech Inc., can be ruled out because the words *Inc., University* indicate organizations, and *British Columbia* is ruled out due to a gazetteer.

Both the terminology chunker and the abbreviation system identify terms in sentences in each document. These instances are collected and output to be used for stage 2.

We also use the stage one output independently of the rest of the Termolator, to find instances of terms that are arguments of the Information Extraction relations discussed in Meyers, et. al. (2014b). Some example relations from the PubMed corpus follow:

1. found *in the **IκB protein**, an **inhibitor of NF-κB***
    - Relation: **Exemplify**, Arg_1: ***IκB protein***, Arg_2*:* ***inhibitor of NF-κB***
    - Interpretation: Arg_1 is an instance of Arg_2
2. *a **necrotrophic effector system** that is an exciting contrast to the **biotrophic effector models** that have been intensively studied*
    - Relation: **Contrast**, Arg_1: ***necrotrophic effector system***, Arg_2: ***biotrophic effector models***
    - Interpretation: Arg_1 and Arg_2 are in contrast with each other
3. ***Bayesian networks** hold a considerable advantage over **pairwise association tests***
    - Relation**: Better than**, Arg_1: ***Bayesian networks,*** Arg_2: ***pairwise association tests***
    - Interpretation: Arg_1 is better than Arg_2 (in some respect)
4. ***housekeeping gene 36B4*** (***acidic ribosomal phosphoprotein P0***)
    - Relation: **Alias**, Arg_1: ***housekeeping gene 36B4***, Arg_2: ***acidic ribosomal phosphoprotein P0***

[2] There are 1445 adjectives in COMLEX with these endings in COMLEX and it is possible to quickly go through these by eye in a few hours. All but 237 of these adjectives were deemed to be technical.

[3] This set of constraints is based on informal observations of the composition of valid terms in corpora. We validate this set of constraints by showing that results that are constrained this way have higher scores than results that are not so constrained, as discussed below in the Evaluation section.

- Interpretation: Arg_1 and Arg_2 are alternative names for the same concept, but neither is a shortened form (acronym or abbreviation).

Stage 2: Distributional Ranking

While stage 1 identifies term instances or tokens, stage 2 groups together these tokens into general types, i.e., it determines which noun sequences would belong in a terminology dictionary for a particular field. Furthermore, this classification is relative to a particular field or topic, represented by contrasting sets of documents. This methodology is based on many previous systems for identifying terminology (Damerau 1993, Drouin 2003, Navigli and Velardi 2004, etc.) which aim to find nouns or noun sequences (n-grams or noun groups) that are the most characteristic of a topic. Towards this goal, noun sequences are ranked according to their characteristic-ness of one topic, where a noun sequence N_1 is more characteristic to a topic T than a noun sequence N_2, if N_1 scores higher than N_2 using a metric that rewards a term for occurring more frequently in some target set of documents about a single topic than it does in a set of documents about a wide variety of topics. The output of systems of this type have been used as Information Retrieval key words (Jacquemin and Bourigault 2003) or terms to be defined in thesauri or glossaries for a particular field (Velardi, et. al. 2001). We plan to use terms derived this way as part of a technology forecasting system (Daim et al., 2006, Babko-Malya, et. al. 2015).

We rank our terms using a combination of three metrics: (1) the standard Term Frequency Inverse Document Frequency (TFIDF); (2) Document Relevance Document Consensus (DRDC) metric **(Navigli and Velardi,** 2004); and (3) Kullback-Leibler Divergence (KLD) metric (Cover and Thomas, 1991; Hisamitsu et al., 1999). The TFIDF metric selects terms specific to a domain by favoring terms that occur more frequently in the relevant (foreground) documents than they do in the background. The formula is:

$$TFIDF(t) = \frac{freqRDG(t)}{freqTotalDoc(t)} * \log(\frac{numTotalDocs}{numDocContains(t)})$$

In the DRDC metric, two factors are considered: (i) document relevance (DR), which measures the specificity of a terminological candidate with respect to the target domain via comparative analysis of a general domain; and (ii) document consensus (DC), which measures the distributed use of a terminological candidate in the target domain. The formula for DRDC is:

$$DRDC(t) = \frac{freqRDG(t)}{freqTotalDoc(t)} * \sum_{d \in RDG} \frac{freq(t,d)}{freqRDG(t)} * \log(\frac{freqRDG(t)}{freq(t,d)})$$

where freqRDG means the frequency of a specific term t (for example, "cth2 mrna") in a Related Document Group (RDG), documents relevant to the same topic. FreqTotalDoc of t means the frequency in all documents (RDG+nonRDG), freq(t,d) means the frequency of t in document d. The KLD metric measures the difference between two probability metrics: the probability that a term will appear in the foreground corpus vs the background corpus. The formula is:

$$KLD(t) = \log(freqRDG(t)) - \log(freqTotalDoc(t))) * freqRDG(t)$$

These three metrics are combined together with equal weights, ranking both the terms produced in stage 1 and substrings of those terms, producing an ordered list.

Stage 3: Well-formedness Score and Relevance Score

The previous stages produce a ranked list of terms, the ranking derived from the distributional score, which we encode as ***D***, a percentile score between 0 and 1. This score can be reranked by creating other scores between 0 and 1 and multiplying all the scores together. Weights can be applied as exponents on each of the scores, resulting in one aggregate score that we use for reranking the terms. However, we currently assume all weights to equal 1. We assume 2 additional scores: *W*, a well-formedness score and ***R***, a relevance score. The aggregate score which we use for reranking purposes is simply: ***D*W*R.*** Like stage 1, the stage 2 components (***W*** and ***R***) can be used separately from the other portions of Termlator, to score or rank terms entered by a user, e.g., terms produced by other terminology extraction systems.[4]

Well-formedness Score

Our well-formedness (W) score is based on several linguistic rules and subjective evaluations about violations of those rules. Although many of these linguistic rules are built into the chunking rules in stage 1, stage 2 includes highly frequent substrings of stage 1 terms in the output. Also the abbreviation rules may introduce terms that would not have been licensed by the chunking rules. So when applied to our own terms ***W*** usually has a value of 1 for terms produced in stage 1, 0 for substrings that are not well-formed and very rarely intermediate values. The W score filters out erroneous substrings, since a ***W*** score of 0 multiplies with all other scores to produce an aggregate score of 0. As mentioned, the stage 3 filters can be applied to term lists not produced by The Termolator, e.g., terms based on N-grams, rather than noun-groups. For this application, the W score is likely to have a larger effect than it does for terms produced by the Termoloator.

We assume that applications of the following rules are reason to give a candidate term a perfect score (1.0):

- **ABBREVIATION_OR_TERM_THAT_IS_ABBREVIATED** – This rule matches terms that are either abbreviations or a full length term that has been abbreviated, e.g., *html, hypertext markup language, OCR, optical character recognition, ...*
- **Out_of_Vocabulary_Word** – This rule matches terms consisting of single words (and their plurals) that are not found in our dictionaries, e.g., *radionuclide, photoconductor, …*
- **Hyphenated Word + OOV Noun** – This applies if a word contains one or more hyphen and the part of the word following the last hyphen would matches the conditions described in the previous bullet, e.g., *mono-axial, lens-pixel, …*

These rules yield a score of **0.7:**

- **Common_Noun_Nominalization** – This means that the term is a single word, identified as a nominalization using dictionary lookup, e.g., *demagnetization, overexposure,*
- **Hyphenated Word + Nominalization** – This applies if a word contains one or more hyphen and the part of the word following the last hyphen would match the conditions described in the previous bullet, e.g., *de-escalation, cross-fertilization*

[4] We have used these components to evaluate sets of terms that were not produced by the Termolator. Our subjective analysis is that they can be used effectively in this way to rate or rerank such terms, but a formal evaluation is outside the scope of this paper.

This rule gives a score of **0.3**:

- **Normal_Common_Noun_or_Number** – This means that the term consists of a single word that is either a number, a common noun, a name or a combination of numbers and letters (e.g., *ripcord, H1D2*).

The following rules have scores that vary, depending on the type of words found in the phrase:

- **Normal_NP** – This means that the term consists of a word sequence that is part of a noun group according to our chunker, described above. The score can be as high as **1.0** if the term would be recognised as such by our stage 1 chunker (e.g., *electrophotographic photoconductor*). A noun group containing one "unnecessary" element such as a preceding adjective, would have a score of **0.5 (*acceptable organic solvent*)**. Other noun groups or noun phrases would have scores of **0.2 (*wheel drive capacity*)**.
- **2_Part_NP** – This means that the term consists of 2 noun groups according to our chunker, possibly separated by a preposition. Currently 2_Part_NPs containing prepositions receive scores of **0.45 (*voltage of the output buffer*)**, and those without receive scores of **0 (*service provider issues remittance*).**

There are several other rules which have scores of 0 associated with them including:

- **Single_Word_Non_Noun** – This means that the word is identified as a non-noun, either by dictionary lookup or by simple morphological rules, e.g., we assume that an out of vocabulary word ending in *-ly* is an adverb, e.g., *downwardly, optical, tightening*
- **Bad_character** -- This means that the term contains at least one character that is not either: a) a letter; b) a number; c) a space; d) a hyphen; e) a period; or f) an apostrophe, e.g., box™, *sum_l, slope Δa*
- **Contains_conjunction** – This rule matches sequences including coordinate conjunctions (*and, or, but, nor*), e.g., *or reproducing, asic or other integrated*
- **Too many verbs** – This means that the sequence contains multiple verbs, e.g., *insulating film corresponding, emitting diodes disposed*
- **Verbal or Sentential Structure** – This means that some chunking rules found a verbal constituent other than an adjective-like pre-modifier (*broken record*), e.g., *developer containing, photoelectric converting*
- **Unexpected_POS_sequence** – This applies to multi-word terms that do not fit any of the profiles above, e.g., *of the developing roll, beam area of the charged*

Relevance Score

The relevance score is derived by searching for the term using Yahoo's search engine (powered by Microsoft Bing) and applying some heuristics to the search result. This score is intended to measure the "relevance" of a term to technical literature. The Relevance Score $\boldsymbol{R} = HT^2$ where the two factors $\boldsymbol{H}$ and T are defined as follows and the weight on $\boldsymbol{T}$ was determined experimentally:

- $\boldsymbol{H}$ = the total number of hits for an exact match. The log 10 of this number (up to a maximum of 10) is normalized between 0 and 1.
- $\boldsymbol{T}$ = the percentage of the top 10 hits that are either articles or patents

The following information from a Yahoo search are used to compute this score: (1) the total number of hits; (2) a check to see if this result is based on the search or if a similar search was substituted, i.e., if the result includes the phrase ***including results for*** or the phrase ***showing***

results for, then we know that our search was not matched at all and we should downgrade the value of H to a very small number.[5]; and (3) the top 10 search results as represented by URLs, titles and summaries. For each result, we search the URL, title and summary for key words which indicate that this hit is probably an article or a patent (*patent, article, sciencedirect, proceedings, journal, dissertation, thesis, abstract*). ***T*** is equal to the number of these search results that match, divided by 10. In practice, this heuristic seems to capture the intuition that a good term is likely to be the topic of current scientific articles or patents, i.e., that the term is relevant.

Runtime is a limiting factor for the Relevance scores because it takes about .75 seconds to search for each term. This means that producing Relevance scores for 30K terms takes about 6 hours, whereas the rest of the system for producing terms takes minutes.

Evaluation

We ran the complete system with 5000 patents about optical systems and components as the foreground (US patent codes 250, 349, 356, 359, 362, 385, 398 and 399) and 5000 diverse patents as background. We collected a total of 219K terms, ranked by the stage 2 system. We selected the top 30K of these terms and ran the stage 3 processes on these 30K terms. We ranked these top terms 3 different ways, each time selecting a different top 5000 terms for evaluation. We selected the top 5000 terms after ranking these 30K terms in the following ways: (a) according to stage 2 (Distributional Score); (b) according to the Relevance Score (c) according to the Combined Score (D*R*W). As W primarily was used to remove ill-formed examples, it was not well-suited for this test as a separate factor. For each list of 5000 terms, we sampled 100 terms, took 20 random terms from each 20% interval, manually inspected the output, and rated each term as correct or incorrect. 71% of the terms ranked according to D only were correct; 82% of the terms ranked according to R were correct and 86% of the terms ranked according to the Combined Score were correct. While we believe that it is signicant that the combined score produced the best result, it is unclear whether the fact that R alone did better than the stage 2 ranking because the R score was applied to the 30K terms out of 219K terms with the highest D scores. While in principle, we could run R on all 219 K terms, time constraints make it impractical to do this, in general, for all output of our system.[6]

Coverage of a term extractor is difficult to measure for terms without having a human being do the task, e.g., reading all 5000 articles and writing out the list of terms.[7] Informally however, we have observed a significant increase in term output since we adopted the chunking model described above, compared to a previous version of the system that used a

[5] Our current strategy is to treat instances of fewer than 10 hits the same as if the term did match, but to set H as if there were 500 hits.

[6] We evaluated the correctness of terms ourselves. We previously did some experiments in which graduate biology students evaluated our biology terms. We discontinued this practice primarily because we could not afford to have experts in all of the domains for which we had terms. In addition, the domain expertise was rarely accompanied by linguistic expertise. So the process of training domain experts to make consistent determinations about what does and does not constitute a linguistic unit was difficult. In contrast, using one set of annotators resulted in more consistent evaluation. Most unknown terms could be looked up and identified with high accuracy.

[7] There are no established sets of manually encoded data to test the system with. Note that the SemEval keyword extraction task (Kim, et. al. 2010) while overlapping with terminology extraction, does not capture the task we are doing here. In particular, we are not attempting to find a small number of keywords for a small number of articles, but rather large sets of terms that cover fields of study. We believe that constructing such a shared task manually would be prohibitive.

standard noun chunker. In other words, we are able to take a larger number of top ranked terms than before without a major decline in accuracy. One of the tasks for future work is to develop a good metric for measuring this.

Examples

Table 1 provides some sample potential terms along with scores ***D, W, R*** and the aggregate score. The table is arranged in descending order by the aggregate score. These terms are excerpts from the best of the three rankings described in the previous section, i.e., the terms ordered by the total score. In the right-most column is an indication of whether or not these are valid terms, as per the judgement of one of the authors. The incorrect examples include: (a) ***irradiation time t***, which is really a variable (a particular irradiation time), not a productively used noun group that should be part of a glossary or a key word; (b) ***evolution***, a common word that is part of the general language and should no longer be relegated to a list of specialized vocabulary; and (c) ***crystal adjacent***, a word sequence that does not form a natural constituent – it is part of longer phrases like ***a one-dimensional photonic crystal adjacent to the magneto-optical metal film.*** In this sequence the word ***crystal***, is modified by a long adjectival modifier beginning with the word ***adjacent*** and it would be an error to consider this pair of words a single constituent.

Table 1: System Output with aggregate scores, component scores and correctness judgements

Rank	Term	D	W	R	Total	Correct
41	stimulable phosphor	.866	1	.174	.151	Yes
104	ion beam profile	.889	1	.117	.126	Yes
346	x-ray receiver	.906	1	.099	.089	Yes
533	wavelength-variable	.838	1	.091	.076	Yes
556	irradiation time t	.460	1	.163	.075	No
1275	quadrupole lens	.460	1	.113	.052	Yes
1502	evolution	.439	1	.109	.048	No
1581	proximity correction	.451	1	.103	.046	Yes
1613	dfb laser	.943	1	.049	.046	Yes
1685	asymmetric stress	.493	1	.067	.033	Yes
3834	panoramagram	.483	1	.056	.027	Yes
4203	crystal adjacent	.316	1	.080	.025	No
4244	single-mode optical fiber	.875	1	.029	.025	Yes
4467	total reflection plane	.988	1	.024	.024	Yes
4879	photosensitive epoxy resin	.286	1	.079	.022	Yes

The Chinese System

Our current Chinese Termolator implements several components parallel to the English system and we intend to implement additional components in future work. The Chinese Termoloator uses an in-house CTB[8] word segmenter and part-of-speech tagger and a rule based noun group chunker, but without additional rules requiring technical words. Stage 2 is similar to the English system in that we compare word distribution in a given domain with word distribution in a general background set and find topic words of the given domain.

[8] https://catalog.ldc.upenn.edu/LDC2013T21

One challenge for the Chinese system is that Chinese word boundaries are implicit, and are automatically induced by the word segmenter, which is prone to errors. We accordingly implemented an accessor-variety (AV) based filter (Feng et al., 2004), which calculates an accessor-variety score for each word based on the number of distinct words that appear before or after it. Character sequences with low AV scores are not independent enough, and usually should not be considered as valid Chinese words (Feng et al., 2004). We therefore filter out words whose accessor-variety scores are less than 3. We evaluated the precision of extracted terms on a set of speech processing patents: the precision was 85% for the top 20 terms and 78% for the top 50 terms. This evaluation was based on 1,100 terms extracted from 2,000 patents related to speech processing.

We developed a well-formedness-based automatic evaluation metric for Chinese terms, which follows the same spirit as the English well-formedness score. This metric penalizes noun phrases that contain non-Chinese characters, contain words that are not nouns or adjectives, contain too many single character words, or are longer than 3 characters. Since this error is exactly the sort of error that would be ruled out by the AV-based filter, we do not use it as part of our own terminology system. Rather, we use it when we are applying our filters to score term lists created externally, just as we are doing with parts of the English system.

We expect to implement a version of the Relevance Score that will work with Chinese language search engines in future work. As with the English, this will be a separable component of the system that can be applied to Chinese term lists created independently from our system.

Concluding Remarks

We have described a terminology system with state-of-the-art results for English that combines several different methods including linguistically motivated rules, a statistical distribution metric and a web-based relevance metric. We can derive at least 5000 highly accurate (86%) terms from 5000 documents about a topic. We have partially implemented this system for Chinese and are currently achieving high accuracy for Chinese as well. In future work, we intend to further develop the system for Chinese and improve the evaluation measures for English.

Acknowledgments

Supported by the Intelligence Advanced Research Projects Activity (IARPA) via Department of Interior National Business Center contract number D11PC20154. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. Disclaimer: The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of IARPA, DoI/NBC, or the U.S. Government.

References

Babko-Malaya, O., Seidel, A., Hunter, D., HandUber, J., Torrelli, M., and Barlos, F. (2015). Forecasting Technology Emergence from Metadata and Language of Scientific Publications and Patents. *15th International Conference on Scientometrics and Informetrics*.

Cover, T. and Thomas, J. A. (1991). *Elements of Information Theory.* Wiley-Interscience, New York.

Daim, T. U., Rueda, G., Martin, H., and Gerdsri, P. (2006). Forecasting emerging technologies: Use of bibliometrics and patent analysis. *Technological Forecasting and Social Change*, 73(8):981–1012.

Damerau, F. J. (1993). Generating and evaluating domain-oriented multiword terms from texts. *Information Processing and Management*, 29:433–447.

Drouin, P. (2003). Term Extraction Using Non-technical Corpora as a Point of Leverage. *Terminology*, 9: 99—115.

Feng, H., Chen, K., Deng, X., and Zheng, W. (2004). Accessor variety criteria for chinese word extraction. *Computational Linguistics*, 30:75–93.

Hisamitsu, T., Niwa, Y., Nishioka, S., Sakurai, H., Imaichi, O., Iwayama, M., and Takano, A. (1999). Term extraction using a new measure of term representativeness. *Proceedings of the First NTCIR Workshop on Research in Japanese Text Retrieval and Term Recognition*.

Jacquemin, C. and Bourigault, D. (2003). Term Extraction and Automatic Indexing. In Mitkov, R., editor, *Handbook of Computational Linguistics*. Oxford University Press, Oxford.

Justeson, J. S. and Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural Language Engineering*, 1(1):9–27.

Kim, S. N., Medelyan, O., Kan, M. Y., and Baldwin T. (2010). SemEval-2010 Task 5: Automatic Keyphrase Extraction from Scientific Articles. SemEval 2010, pages 21—26.

Macleod, C., Grishman, R., and Meyers, A. (1997). COMLEX Syntax. *Computers and the Humanities*, 31:459–481.

Macleod, C., Grishman, R., Meyers, A., Barrett, L., and Reeves, R. (1998). Nomlex: A lexicon of nominalizations. *Proceedings of Euralex98*.

Meyers, A., Glass, Z., Grieve-Smith, A., Y. He, S. L., and Grishman, R. (2014a). Jargon-Term Extraction by Chunking. *COLING Workshop on Synchronic and Diachronic Approaches to Analyzing Technical Language*.

Meyers, A., Lee, G., Grieve-Smith, A., He, Y., and Taber, H. (2014b). Annotating Relations in Scientific Articles. *LREC-2014*.

Meyers, A., Reeves, R., Macleod, C., Szekeley, R., Zielinska, V., and Young, B. (2004). The Cross-Breeding of Dictionaries. *Proceedings of LREC-2004*, Lisbon, Portugal.

Navigli, R. and Velardi, P. (2004). Learning Domain Ontologies from Document Warehouses and Dedicated Web Sites. *Computational Linguistics*, 30.

Ramshaw, L. A. and Marcus, M. P. (1995). Text Chunking using Transformation-Based Learning. *ACL Third Workshop on Very Large Corpora*, pages 82–94.

Schwartz, A. and Hearst, M. (2003). A simple algorithm for identifying abbreviation definitions in biomedical text. *Pacific Composium on Biocomputing*.

Velardi, P., Missikoff, M., and Basili, R. (2001). Identification of relevant terms to support the construction of domain ontologies. *Workshop on Human Language Technology and Knowledge Management*.