Domain-independent term extraction through domain modelling

Georgeta Bordea, Paul Buitelaar, Tamara Polajnar

Domain-independent term extraction through domain modelling

Georgeta Bordea
UNLP, DERI
National University
of Ireland, Galway
name.surname
at deri.org

Paul Buitelaar
UNLP, DERI
National University
of Ireland, Galway
name.surname
at deri.org

Tamara Polajnar
Computer Laboratory
University of Cambridge
name.surname
at cl.cam.ac.uk

Abstract

Extracting general or intermediate level terms is a relevant problem that has not received much attention in literature. Current approaches for term extraction rely on contrastive corpora to identify domain-specific terms, which makes them better suited for specialised terms, that are rarely used outside of the domain. In this work, we propose an alternative measure of domain specificity based on term coherence with an automatically constructed domain model. Although previous systems make use of domain-independent features, their performance varies across domains, while our approach displays a more stable behaviour, with results comparable to, or better than, state-of-the-art methods.

Term extraction plays an important role in a wide range of applications including information retrieval (Yang et al., 2005), keyphrase extraction (Lopez and Romary, 2010), information extraction (Yangarber et al., 2000), domain ontology construction (Kietz et al., 2000), text classification (Basili et al., 2002), and knowledge mining (Mima et al., 2006). In many of these applications the specificity level of a term is a relevant characteristic, but despite the large body of work in term extraction there are few methods that are able to identify general terms or intermediate level terms. Take for example the following structure from the AGROVOC vocabulary[1]: *resources* → *natural resources* → *mineral resources* → *lignite*, where *resources* is an upper level term, *natural resources* and *mineral resources* are intermediate level terms, and *lignite* is a leaf. Intermediate level terms are specific to a domain but are broad enough to be usable for summarisation and classification. Methods that make use of contrastive corpora to select domain specific terms favour the leaves of the hierarchy, and are less sensitive to generic terms that can be used in other domains.

Instead, we construct a domain model by identifying upper level terms from a domain corpus. This domain model is further used to measure the coherence of a candidate term within a domain. The underlying assumption is that top level terms (e.g., *resource*) can be used to extract intermediate level terms, in our example *natural resources* and *mineral resources*. Our method for constructing a domain model is evaluated directly through an expert survey as well as indirectly based on its contribution to intermediate level term extraction. While domain modelling is tested and exemplified with English, the ideas presented here are not language dependent and can be applied to other languages, but this is outside the scope of this work.

We start by giving an overview of related work in term extraction in Section 1. Then, an approach to construct a domain model based on domain coherence is proposed in Section 2, followed by a method to apply domain models for term extraction. The experimental part of the paper starts with a direct evaluation of a domain model through a user survey (Section 3). A first set of experiments is carried in a standard setting for term evaluation, while the second set of experiments is application-driven, using corpora annotated for keyphrase extraction, information extraction, and information retrieval. We conclude this paper in Section 4, giving a few directions for future work.

[1] AGROVOC: http://aims.fao.org/standards/agrovoc/about

1 Related work

Methods for term extraction that use corpus statistics alone are faced with the challenge of distinguishing general language expressions (e.g., *last week*) from terminological expressions. A solution to this problem is to use contrastive corpora (Huizhong, 1986). Several contrastive measures are proposed including domain relevance (Park et al., 2002), domain consensus (Velardi et al., 2001), and word impurity (Liu et al., 2005). In this work we propose an approach to compute domain specificity based on a domain model, that is less sensitive to leaf terms and is better suited for intermediate level terms.

The domain model proposed in this work is derived from the corpus itself, without the need for external corpora. An automatic method for identifying the upper level terms of a domain has applications beyond the task of term extraction. Although not named as such, upper level terms were previously used for text summarisation (Teufel and Moens, 2002). The authors manually identified a set of 37 nouns including *theory*, *method*, *prototype* and *algorithm*, without considering a principled approach to extract them. The work presented here is similar to (Barrière, 2007), but instead of re-ranking terms based on their similarity to each other we make use of domain model terms, reducing data sparsity issues.

In our experiments we employ two state of the art methods for term extraction, the NC-value approach (Frantzi et al., 2000) and TermExtractor[2] (Velardi et al., 2001). The former is a hybrid method that ranks terms using only corpus statistics, while the latter exploits contrastive corpora. NC-value is based on raw frequency counts and considers nested multi-word terms by penalising frequency counts of shorter embedded terms. Additionally, it incorporates context information in a re-ranking step using top ranked terms. Context words (nouns, verbs and adjectives) are identified based on their occurrence with top candidates. Our method is an extension of this approach that uses domain models instead of selecting context words based on frequency alone.

TermExtractor is a popular approach that combines different term extraction techniques including domain relevance, domain consensus and lexical cohesion. Domain Relevance (DR) compares the probability of a term t in a given domain D_i with the maximum probability of the term in other domains used for contrast D_j and is measured as:

$$DR_{D_i}(t) = \frac{P(t/D_i)}{max_j\left(P(t/D_j)\right)}, j \neq i \quad (1)$$

Domain Consensus (DC) identifies terms that have an even probability distribution across the corpus that represents a domain of interest, and is estimated through entropy as follows:

$$DC_{D_i}(t) = -\sum_{d \in D_i} P(t/d) \cdot log\left(P(t/d)\right) \quad (2)$$

where d is a document in the domain D_i. Finally, the degree of cohesion among the words w_j that compose the term t is computed through a measure called Lexical Cohesion (LC). Let $|t|$ be the length of t in number of words, and $f(t, D_i)$ the frequency of t in the domain D_i, then Lexical Cohesion is defined as:

$$LC_{D_i}(t) = \frac{|t| \cdot f(t, D_i) \cdot log\left(f(t, D_i)\right)}{\sum_{w_j} f(w_j, D_i)} \quad (3)$$

The weight TE used for ranking terms by TermExtractor is a linear combination of the three methods described above:

$$TE(t, D_i) = \alpha \cdot DR + \beta \cdot DC + \gamma \cdot LC \quad (4)$$

While general terms typically have a high domain consensus, the domain relevance measure boosts narrow terms that have limited usage outside of the domain. For example the term *system* is not identified as relevant for Computer Science because it is frequently used in general language and in other specific domains as biology. In this work we take a different approach to compute domain specificity that can be applied for general terms by using a domain coherence measure that does not use external corpora. Two general purpose corpora, the Open American National Corpus[3] and a corpus of books from Project Gutenberg[4], are used as contrastive corpora for our implementation of TermExtractor. The books selected from

[2]TermExtractor demo: http://lcl.uniroma1.it/sso/index.jsp?returnURL=\%2Ftermextractor\%2F

[3]Open American National Corpus: http://www.americannationalcorpus.org/OANC/

[4]Project Gutenberg: http://www.gutenberg.org/

Project Gutenberg include the bible, the complete works of William Shakespeare, James Joyce's *Ulysses* and Tolstoy's *War and Peace*. We consider only the default setting of TermExtractor assigning equal weights to each measure in Equation 4.

2 Constructing a domain model based on domain coherence

We begin this section by describing an approach for domain modelling based on domain coherence in Section 2.1. Then, we discuss a modification of the NC-value approach which makes it better suited for intermediate level terms (Section 2.2). We conclude this section by describing a novel method for term extraction using a domain model in Section 2.3.

2.1 Domain modelling

A domain model is represented as a vector of words which contribute to determine the domain of the whole corpus. Let Δ be the domain model, and w_1 to w_n a set of generic words, specific to the domain, then:

$$\Delta = \{w_1, ..., w_n\} \quad (5)$$

The number of words n can be empirically set according to a cutoff associated weight. Previous work on using domain information for word sense disambiguation (Magnini et al., 2002) has shown that only about 21% of the words in a text actually carry information about the prevalent domain of the whole text, and that nouns have the most significant contribution (79.4%). Several assumptions are made to identify words that are used to construct a domain model from a domain corpus:

1. **Distribution:** Generic words should appear in at least one quarter of the documents in the corpus;
2. **Length:** Only single-word candidates are considered, as longer terms are more specific;
3. **Content:** Only content-bearing words are of interest (i.e., nouns, verbs, adjectives);
4. **Semantic Relatedness:** A term is more general if it is semantically related to many specific terms.

The distribution assumption implies that rare terms are more specific, similar with the frequency-based measure previously used for measuring tag generality (Benz et al., 2011). This might not always be the case, for example a simple search with a search engine shows that *artefact* or *silverware* are more rarely used than the term *spoon*, although the first two concepts are more generic. However, in this work we are interested in extracting basic-level categories as theorised in psychology (Hajibayova, 2013). A basic-level category is the preferred level of naming, that is the taxonomical level at which categories are most cognitively efficient. A counter example can be found for the length assumption as well, as the longer term *inorganic matter* is more general than the single word *knife*, but in this case we would simply consider as a candidate the single word *matter* which is more generic than the compound term. Both length and frequency of occurrence are proposed as general criteria for identifying basic-level categories (Green, 2005).

The first three assumptions are used for candidate selection, while the fourth assumption is used to filter the candidates. A possible solution for building a domain model is to use a standard termhood measure for single-word terms. Most approaches for extracting single-word terms make use of contrastive corpora, ranking higher specific words that are rarely used outside of the domain. But our domain model is further used for term extraction, therefore it is important that we use generic words to insure a high recall.

We interpret coherence as semantic relatedness to quantify the coherence of a term in a domain. The measure used for semantic relatedness is Pointwise Mutual Information (PMI). First, we extract multi-word terms using a standard term extraction technique, then we use the top ranked terms to filter candidate words using the following scoring function for domain coherence:

$$s(\theta) = \sum_{\sigma \in \Omega} PMI(\theta, \sigma) = \sum_{\sigma \in \Omega} log\left(\frac{P(\theta, \sigma)}{P(\theta) \cdot P(\sigma)}\right) \quad (6)$$

where θ is the domain model candidate, σ is top ranked multi-word term, Ω is the set of top ranked multi-word terms and $P(\theta, \sigma)$ is the probability that the word θ appears in the context of the term σ. In our implementation, the set Ω contains the best terms extracted by our baseline term extraction method described in Section 2.2, but any other term extraction method can be applied in this step. A small sample from domain models extracted us-

Computer Science	Biomed	Food and Agriculture
development	mechanism	control
software	evidence	farm
framework	antibody	supply
information	molecule	food
system	system	forest

Table 1: Example words from domain models extracted for different domains

ing our domain coherence method for Computer Science, Food and Agriculture, and the Biomedical Domain, is shown in Table 1.

2.2 Baseline term extraction method

Our baseline approach for intermediate level term extraction is frequency-based, similar to the C-value method (Ananiadou, 1994), but we modify its ranking function. The main difference is the way we take into consideration embedded terms. In previous work, this information is used to decrease frequency counts, as shorter terms are counted both when they appear by themselves and when they are embedded in a longer term. We argue that the number of longer terms that embed a term can be used as a termhood measure. In our experiments, this measure only works for embedded multi-word terms, as single-word terms are too ambiguous. The baseline scoring method b is defined as:

$$b(\tau) = |\tau| \log f(\tau) + \alpha e_\tau \quad (7)$$

where τ is the candidate string, $|\tau|$ is the length of τ, f is its frequency in the corpus, and e_τ is the number of terms that embed the candidate string τ. The parameter α is used to linearly combine the embeddedness weight and is empirically set to 3.5 in our experiments.

2.3 Using domain coherence for term extraction

Although we proposed a method to build a domain model in Section 2.1, the question of how to use this domain model in a termhood measure remains unanswered. Again, the solution is to rely on the notion of domain coherence, which is defined in this case as the semantic relatedness between a candidate term and the domain model described above. The assumption is that a correct term should have a high semantic relatedness with representative words from the domain. This method favours more generic candidates than contrastive corpora approaches, therefore it is better suited for extracting intermediate level terms.

The same measure of semantic relatedness is used as for the domain model, the PMI measure. The domain coherence DC of a candidate string τ is defined as follows:

$$DC(\tau) = \sum_{\theta \in \Delta} PMI(\tau, \theta) \quad (8)$$

where θ is a word from the domain model, and Δ is the domain model constructed using Equation 6. Using generic terms to build the domain model is crucial for ensuring a high recall as these words are more frequently used across the corpus. In our implementation context is defined as a window of 5 words.

3 Experiments and Results

Evaluating term extraction results across domains is a challenge, because finding domain experts is difficult for more than one domain. An alternative is to reuse datasets annotated for applications where term extraction plays an important role, for example, keyphrase extraction or index term assignment. Three technical domain corpora are used in our experiments: *Krapivin*, a corpus of scientific publications in Computer Science (Krapivin et al., 2009); *GENIA*, a corpus of abstracts from the biomedical domain (Ohta et al., 2001); and *FAO*, a corpus of reports about Food and Agriculture (Medelyan and Witten, 2008) collected from the website of the Food and Agriculture Organization of the United Nations[5]. The *Krapivin* corpus provides author and reviewer assigned keyphrases for each publication. The *GENIA* corpus is exhaustively annotated with biomed terms, with about 35% of all noun phrases annotated as biomed terms. The *FAO* dataset provides index terms assigned to each document by professional indexers. It is not only the document size that varies considerably across these three corpora, but also the number of annotations assigned to each document as can be seen in Table 2.

We evaluate our measure for building a domain model in Computer Science, by identifying a list of general words with the help of a domain expert in Section 3.1. We envision two sets of experiments: a standard term extraction evaluation

[5]Food and Agriculture Organization of the United States: `http://www.fao.org`

Corpus	Documents	Tokens	Avg. Annotations
Krapivin	2304	$22 \cdot 10^6$	5
GENIA	1999	$0.5 \cdot 10^6$	37
FAO	780	$28 \cdot 10^6$	8

Table 2: Corpora statistics

where the top ranked terms are evaluated against the list of unique annotations provided in the evaluation datasets (Section 3.2.1), and a second set of experiments where each term extraction approach is used to assign candidates to documents in combination with a document relevance measure in Section 3.2.2.

3.1 Intrinsic evaluation of a domain model

A domain expert was asked to investigate nouns used in the ACM Computing Classification System[6]. The expert was provided with the list of nouns and their frequency in the taxonomy and was required to identify nouns that refer to generic concepts. A set of 80 nouns were selected in this manner including *system*, *information*, and *software*. Only one annotator was involved because of the complexity of the task, that implies the analysis and filtering of several hundred words. We estimate the inter-annotator agreement by analysing a subset of the selected words through a survey with 27 participants. A quarter of the selected words are combined with the same number of randomly selected rejected words and the resulting list is sorted alphabetically. The Fleiss kappa statistic for interrater agreement is 0.34, lying in the fair agreement range. 80% of the words from our gold standard domain model were selected by at least half of the participants.

We compare our method (DC) with two other benchmarks, the contrastive termhood measure used in TermExtractor, and the frequency-based method used by NC-value to select context words ($NCVweight$). Again, context is defined as a window of 5 words. A domain model has many similarities with probabilistic topic modelling, although it provides less structure. We compare our approach with a popular approach to topic modelling, Latent Dirichlet Allocation (LDA) (Blei et al., 2003). We experimented with different numbers of topics but we report only the best results

[6]ACM Computing Classification System: `http://www.acm.org/about/class/1998/`

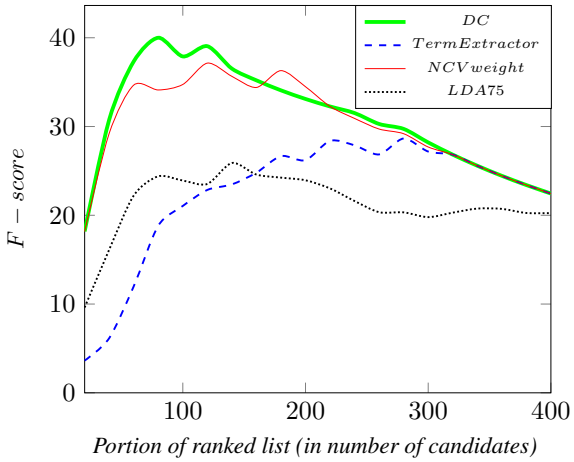

Figure 1: Methods for extracting a domain model

achieved for 75 topics ($LDA75$).

The results of this experiment are shown in Figure 1, in terms of F-score. Several conclusions can be drawn from this experiment. First, the methods that analyse the context of top ranked terms (i.e., our domain coherence measure, DC, and the weight used for context words in the NC-value, w_{NCV}) perform better than the contrastive measure used in TermExtractor, with statistically significant gains. Also, our domain coherence method outperforms the more simple frequency-based weight used in NC-value, although this result is not statistically significant. As expected, the words ranked high by TermExtractor are too specific for a generic domain model. The topic modelling approach identifies several words from the gold standard but much less than our approach and these are evenly distributed across latent topics. These conclusions will be further investigated across two other domains, using gold standard terms annotated for three different applications in Section 3.

3.2 Term extraction evaluation results

We implement and compare the baseline method presented in Section 2.2 and the method based on domain coherence described in Section 2.3, against the NC-value and TermExtractor methods, which are used as benchmarks. The same candidate selection method is used for all the evaluated approaches. Candidate terms are selected through syntactic analysis by defining a syntactic pattern for noun phrases. To assure the results are comparable, the same number of context words is used

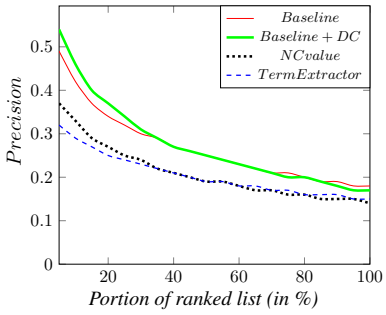

Figure 2: Precision for top 10k terms from the Krapivin corpus

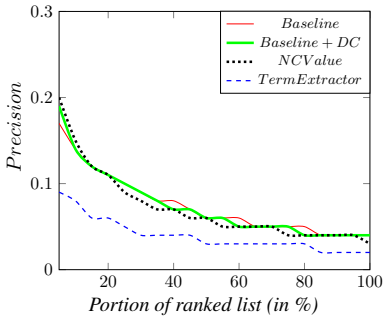

Figure 3: Precision for top 10k terms from the FAO corpus

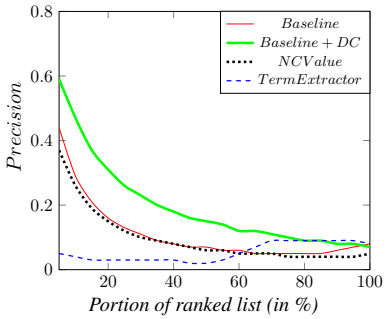

Figure 4: Precision for top 10k terms from the GENIA corpus

in our implementation of the NC-value approach as the size of the domain model. Two general purpose corpora, the Open American National Corpus[7] and a corpus of books from Project Gutenberg[8], are used as contrastive corpora for our implementation of TermExtractor. We considered only the default setting for TermExtractor, assigning equal weights to each measure.

3.2.1 Standard term extraction evaluation

While keyphrases and index terms suit well our purposes, as they are terms of an intermediate level of specificity, meant to summarise or classify documents, many of the terms annotated in GENIA are too specific. We discard the annotated terms that are mentioned in less than 1% of the documents from corpus, based on our distribution assumption. For each of the three datasets, the top ten thousand ranked terms were evaluated. We incrementally analysed portions of the ranked lists computed using the baseline approach (*Baseline*), the baseline approach linearly combined with the domain coherence measure (*Baseline+DC*), and the two benchmarks, *NC-value* and *TermExtractor*. The precision value for a portion of the list is scaled against the overall number of candidates considered. First, we observe that all methods perform better on the GENIA (Figure 4) and the Krapivin corpus (Figure 2), with the best methods achieving a maximum precision close to 60% at the top of the ranked list.

The Food and Agriculture use case is more challenging, as the best method achieves a precision of less than 20%, as can be seen in Figure 3. Also, the contrastive corpora measure employed in TermExtractor yields considerably worse results on all three domains, because the extracted terms are too specific. The baseline method, that rewards embedded terms, outperforms the NC-value method on the Computer Science domain, and in the biomedical domain, but it performs slightly worse on the Agriculture domain. The combination of our baseline method with the domain coherence measure (referred to as *Baseline + DC* in the legend) yields the most stable behaviour, outperforming all other measures across the three domains, considerably so in the biomedical domain (Figure 4) and at the top of the ranked list in Computer Science (Figure 2). In Computer Science, domain coherence significantly outperforms the best performing state-of-the-art method, NC-value (Figure 2). In Biomedicine, the improvement is statistically significant, with a gain of 106% at top 20% of the list (Figure 4).

3.2.2 Application-based evaluation

An important reason for developing termhood measures is that they are needed in specific applications, for example keyphrase extraction and index term extraction. Typically, a termhood measure is combined with different measures of document relevance in such applications, as the candidates are assigned at the document level. We make use of the standard information retrieval measure *TF-IDF* in combination with the considered term extraction scoring functions to assign terms to documents. The best results are obtained by using domain coherence as a post-processing step. In this experiment, the *PostRankDC* approach was computed by re-ranking the top 30 candidates selected using our baseline approach described in

[7]Open American National Corpus: `http://www.americannationalcorpus.org/OANC/`

[8]Project Gutenberg: `http://www.gutenberg.org/`

Top	F@5	F@10	F@15	F@20
Baseline	12.24	12.81	12.14	11.32
PostRankDC	**13.42**	**14.55**	**13.72**	**12.51**
NC-value	6.77	7.32	7.18	6.75
TermExtractor	1.41	1.77	1.95	1.97

Table 3: Keyphrase extraction evaluation on the Krapivin corpus

Top	F@5	F@10	F@15	F@20
Baseline	3.17	3.76	4.03	4.20
PostRankDC	**5**	5.8	5.62	5.29
NC-value	4.65	**5.88**	**6.09**	**5.94**
TermExtractor	0.2	0.31	0.34	0.35

Table 4: Index term evaluation on the FAO corpus

Top	F@5	F@10	F@15	F@20
Baseline	9.67	15.71	20.17	23.19
PostRankDC	**11.36**	17.63	21.52	23.55
NC-value	7.79	11.97	14.01	14.6
TermExtractor	10.77	**17.75**	**22.14**	**24.63**

Table 5: Term extraction at the document level on the GENIA corpus

Equation 7, based on their domain coherence.

The application-based evaluation proposed in this work allows us to evaluate both precision and recall, and consequently F-score can be used as an evaluation metric. The results for keyphrase extraction in Computer Science are presented in Table 3, while the results for index term extraction in the Agriculture domain are shown in Table 4. The results for document level term extraction from the Biomed corpus appear in Table 5. All three methods yield a higher performance on the GENIA corpus. The results on the Agriculture corpus are again the lowest, because a larger number of candidates has to be analysed.

Our $Baseline$ method outperforms the NC-value approach on the Krapivin corpus and on the GENIA corpus, but not on the FAO corpus. We can observe that the domain coherence approach ($PostRankDC$) improves over our baseline approach ($Baseline$) on all three domains. The improvement is statistically significant compared to the best state-of-the-art method in Computer Science, NC-value. NC-value outperforms TermExtractor in Computer Science and Agriculture, but TermExtractor performs better in Biomedicine. Although both NC-value and TermExtractor make use of domain-independent features for ranking, their performance varies across domains and applications. At the same time, combining our domain coherence approach (*PostRankDC*) with our baseline method in a post-ranking step displays a more stable behaviour, achieving the best performance on the Computer Science domain (Krapivin) and similar results with the results of the best method in Biomedicine (GENIA) and Agriculture (FAO).

4 Conclusions

In this study, we proposed an approach to identify intermediate level terms through domain modelling and a novel domain coherence measure, arguing that approaches that make use of contrastive corpora are only suitable for updating existing terminology resources with more specific terms and not for summarisation or classification tasks. The contributions described in this work are three-fold:

i) A method for extracting top level terms from a domain corpus

ii) A novel domain coherence metric based on semantic relatedness with a domain model

iii) A novel application-based evaluation for term extraction systems

Experiments discussed in this paper show that term extraction performance depends on the domain, although systems make use of domain-independent features. Our domain coherence approach based on a domain model performs well across domains, while the performance of the NC-value and TermExtractor benchmarks is more domain-dependent. The results lead to the conclusion that using a domain model is more appropriate than using statistical approaches based on contrastive corpora, for extracting intermediate level terms. Future work will include an unsupervised learning-to-rank approach for term extraction, that will allow a more principled integration of domain coherence measures with standard term extraction features. The method proposed here can be used as a specificity measure, and we currently investigate this in the context of constructing generalisation hierarchies of concepts.

Acknowledgments

This work has been funded in part by the European Union under Grant No. 258191 for the PROMISE

project, as well as by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289.

References

Sophia Ananiadou. 1994. A methodology for automatic term recognition. In *Proceedings of the 15th International Conference on Computational Linguistics (COLING '94)*, page 1034–1038, Kyoto, Japan.

Caroline Barrière. 2007. Une perspective interactive à l'extraction de termes. In *7ème Confèrence" Terminologie et intelligence artificielle"*, pages 95–104.

Roberto Basili, Alessandro Moschitti, and Maria Teresa Pazienza. 2002. Empirical investigation of fast text classification over linguistic features. In Frank van Harmelen, editor, *ECAI*, pages 485–489. IOS Press.

Dominik Benz, Christian Körner, Andreas Hotho, Gerd Stumme, and Markus Strohmaier. 2011. One tag to bind them all: Measuring term abstractness in social metadata. In Grigoris Antoniou, Marko Grobelnik, Elena Simperl, Bijan Parsia, Dimitris Plexousakis, Pieter Leenheer, and Jeff Pan, editors, *The Semanic Web: Research and Applications*, volume 6644 of *Lecture Notes in Computer Science*, pages 360–374. Springer Berlin Heidelberg.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3:993–1022, March.

Katerina Frantzi, Sophia Ananiadou, and Hideki Mima. 2000. Automatic recognition of multi-word terms : the C-value / NC-value method. *Journal on Digital Libraries, Natural language processing for digital libraries*, 3 (2):115–130.

Rebecca Green. 2005. Vocabulary alignment via basic level concepts. In *Final Report, 2003 OCLC/ALISE Library and Information Science Research Grant Project*, Dublin, OH: OCLC.

Lala Hajibayova. 2013. Basic-level categories: A review. *Journal of Information Science*.

Y Huizhong. 1986. A new technique for identifying scientific/technical terms and describing science texts. *Lit. Linguist. Comput.*, 1:93–103, April.

Jörg-Uwe Kietz, Raphael Volz, and Alexander Maedche. 2000. Extracting a domain-specific ontology from a corporate intranet. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*, ConLL '00, pages 167–175, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mikalai Krapivin, Aliaksandr Autayeu, and Maurizio Marchese. 2009. Large dataset for keyphrases extraction. In *Technical Report DISI-09-055, DISI*. University of Trento, Italy.

Tao Liu, X Wang, Guan Yi, Zhi-Ming Xu, and Qiang Wang, 2005. *Domain-Specific Term Extraction and Its Application in Text Classification*, volume 1481, pages 1481–1484.

Patrice Lopez and Laurent Romary. 2010. HUMB : Automatic Key Term Extraction from Scientific Articles in GROBID. In *Proceedings of the ACL 2010 Workshop on Evaluation Exercises on Semantic Evaluation (SemEval 2010)*, number July, pages 248–251.

Bernardo Magnini, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8:359–373.

Olena Medelyan and Ian H. Witten. 2008. Domain independent automatic keyphrase indexing with small training sets. *J. Am. Soc. Information Science and Technology*.

Hideki Mima, Sophia Ananiadou, and Katsumori Matsushima. 2006. Terminology-based knowledge mining for new knowledge discovery. *ACM Trans. Asian Lang. Inf. Process.*, 5(1):74–88.

Tomoko Ohta, Yuka Tateisi, Jin-Dong Kim, Sang-Zoo Lee, and Jun'ichi Tsujii. 2001. Genia corpus: A semantically annotated corpus in molecular biology domain. In *Proceedings of the ninth International Conference on Intelligent Systems for Molecular Biology (ISMB 2001) poster session*, page 68, July.

Youngja Park, Roy J. Byrd, and Branimir Boguraev. 2002. Automatic glossary extraction: Beyond terminology identification. In *19th International Conference on Computational Linguistics - COLING 02*, Taipei, Taiwan, August-September. Howard International House and Academia Sinica.

Simone Teufel and Marc Moens. 2002. Summarizing scientific articles - experiments with relevance and rhetorical status. *Computational Linguistics*, 28:2002.

Paola Velardi, Michele Missikoff, and Roberto Basili. 2001. Identification of relevant terms to support the construction of domain ontologies. In *Proceedings of the ACL 2001 Workshop on Human Language Technology and Knowledge Management*, Toulouse, July.

Lingpeng Yang, Dong-Hong Ji, Guodong Zhou, and Nie Yu. 2005. Improving retrieval effectiveness by using key terms in top retrieved documents. In David E. Losada and Juan M. Fernández-Luna, editors, *ECIR*, volume 3408 of *Lecture Notes in Computer Science*, pages 169–184. Springer.

Roman Yangarber, Ralph Grishman, Pasi Tapanainen, and Silja Huttunen. 2000. Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th conference on Computational linguistics - Volume 2*, COLING '00, pages 940–946, Stroudsburg, PA, USA. Association for Computational Linguistics.