# Hot Topics and Schisms in NLP: Community and Trend Analysis with Saffron on ACL and LREC Proceedings

**Georgeta Bordea, Paul Buitelaar, Barry Coughlan**
INSIGHT, National University of Ireland, Galway
georgeta.bordea, paul.buitelaar, barry.coughlan@insight-centre.org

**Abstract**

In this paper we present a comparative analysis of two series of conferences in the field of Computational Linguistics, the LREC conference and the ACL conference. Conference proceedings were analysed using Saffron by performing term extraction and topical hierarchy construction with the goal of analysing topic trends and research communities. The system aims to provide insight into a research community and to guide publication and participation strategies, especially of novice researchers.

**Keywords:** trend analysis,term extraction,community analysis

## 1. Introduction

The Natural Language Processing (NLP) research community is one of the oldest in Computer Science, starting with the first conferences on Computational Linguistics in the 60s. Over time, many research trends and sub-communities within the NLP community developed, changing every few years with topics appearing and disappearing. It is instructive to analyse these developments in order to map out promising trends and community developments in hindsight. Consider for instance, "Statistical Machine Translation" (SMT), which is currently one of the most successful and widely studied topics of research in NLP. An analysis of the occurrence of SMT in the ACL Anthology[1] can be seen in the Figure 1.

The nature of conferences is that they bring together a research community. Therefore, different conferences will display a difference in emphasis and distribution of mostly studied research topics, depending on the community or sub-community they represent. Consider for instance a comparison on the most studied research topics in the ACL family of conferences on Computational Linguistics (as represented by the ACL Anthology) vs. the LREC conference on Language Resources and Evaluation. An analysis of this is shown in Figure 2, which lists the top-most 30 research topics occurring in both conference proceedings. "Natural Language Processing" is the most prominent research topic in both conferences, which is to be expected. However, the lists also show the continuing emphasis in LREC on resources and spoken language vs. for instance the strong occurrence of grammar engineering and parsing in older ACL work.

Research topics define by their nature also a community of researchers working on them, as can be analyzed by identifying the experts around a certain topic. Figure 3 below shows the community of top-most experts on the research topic of "language resources" as derived from LREC proceedings through text analysis.

Communities are obviously concerned with several related topics, which can be visualized by clustering them as shown in Figure 4 for topics extracted from LREC proceedings. For instance note the significant topic clusters around "language resources" (middle right) and "semantic information" (bottom left), which traditionally have been core topics of the LREC conference. Nodes in the graph are extracted research topics and arcs represent generalisation relations between them, which allows us to identify clusters of closely related topics that reflect research communities. This method was previously applied also to study research communities in the Web Science domain (Hooper et al., 2013). The methods for constructing and visualizing the graph are described in more detail in the next section.

## 2. Topic and Community Analysis with Saffron

A conference can be seen as a collection of people (researchers), terms (research topics) and documents (conference papers and proceedings). The community and trend analysis we report in this paper has been developed in the context of Saffron [2], a system that provides insights in a research community or organisation by analysing main topics of investigation (terms) and the individuals associated with them (people) through text mining on their writings (documents). Currently, Saffron analyses mainly Computer Science areas (NLP, IR, Semantic Web, Web Science), but there is an ongoing effort to extend this to other research domains. Saffron is developed primarily as an Expert Finder system for the exploration and discovery of experts and expertise within a community or organisation.

### 2.1. Term Extraction

At the core of Saffron is a term extraction algorithm that is used to identify research topics in a document collection of conference proceedings. Term extraction plays an important role in a wide range of applications including information retrieval citelingpeng2005improving, 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

[1] ACL Anthology corpus: `http://aclweb.org/anthology/`

[2] Saffron: `http://saffron.deri.ie/`

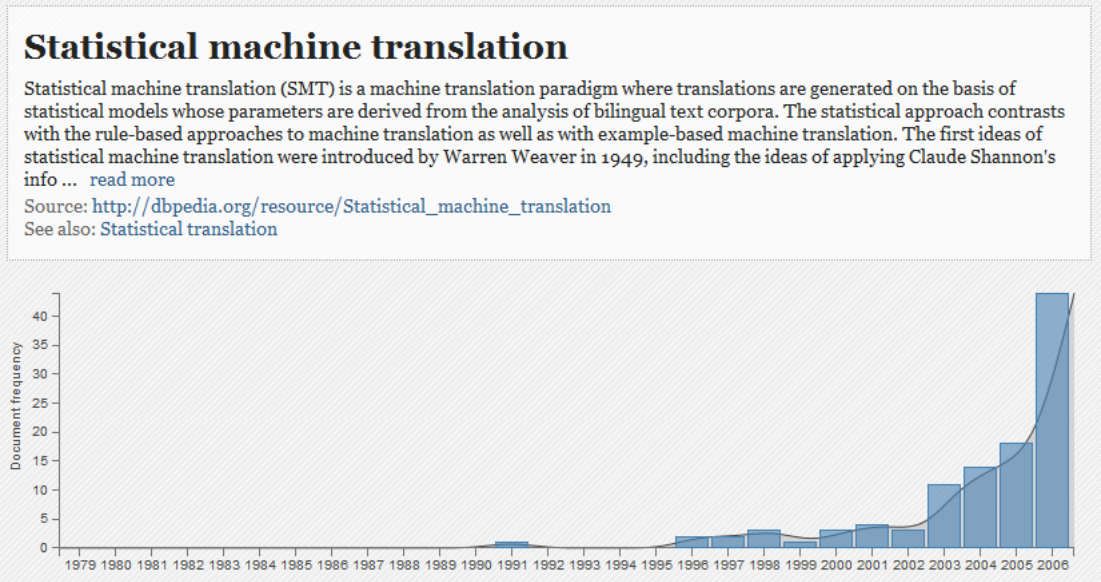


Figure 1: Trend analysis on ”Statistical Machine Translation“ in the ACL Anthology

| | | | |
|---|---|---|---|
| 1 | Natural language processing | 1 | Natural Language Processing |
| 2 | Natural language | 2 | Language Resources |
| 3 | Language model | 3 | Natural Language |
| 4 | Training data | 4 | Human Language Technology |
| 5 | Machine translation | 5 | Machine translation |
| 6 | Machine learning | 6 | Language Technology |
| 7 | Information retrieval | 7 | Language model |
| 8 | Feature structures | 8 | Information retrieval |
| 9 | Applied Natural Language Processin... | 9 | Language Processing |
| 10 | Computer Science | 10 | Word sense disambiguation |
| 11 | Statistical machine translation | 11 | Annotation tool |
| 12 | Language processing | 12 | Speech recognition |
| 13 | Context-free grammar | 13 | Machine learning |
| 14 | hidden Markov model | 14 | Support Vector Machines |
| 15 | Support Vector Machines | 15 | Knowledge base |

Figure 2: Top 15 topics in the ACL Anthology (left) and LREC proceedings (right)


Figure 3: Experts identified in LREC proceedings for the topic ”language resources“

term extraction there are few methods that are able to identify general terms or intermediate level terms. Intermediate level terms are specific to a domain but are broad enough to be usable for analytics tasks such as the one described here. 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 identi-

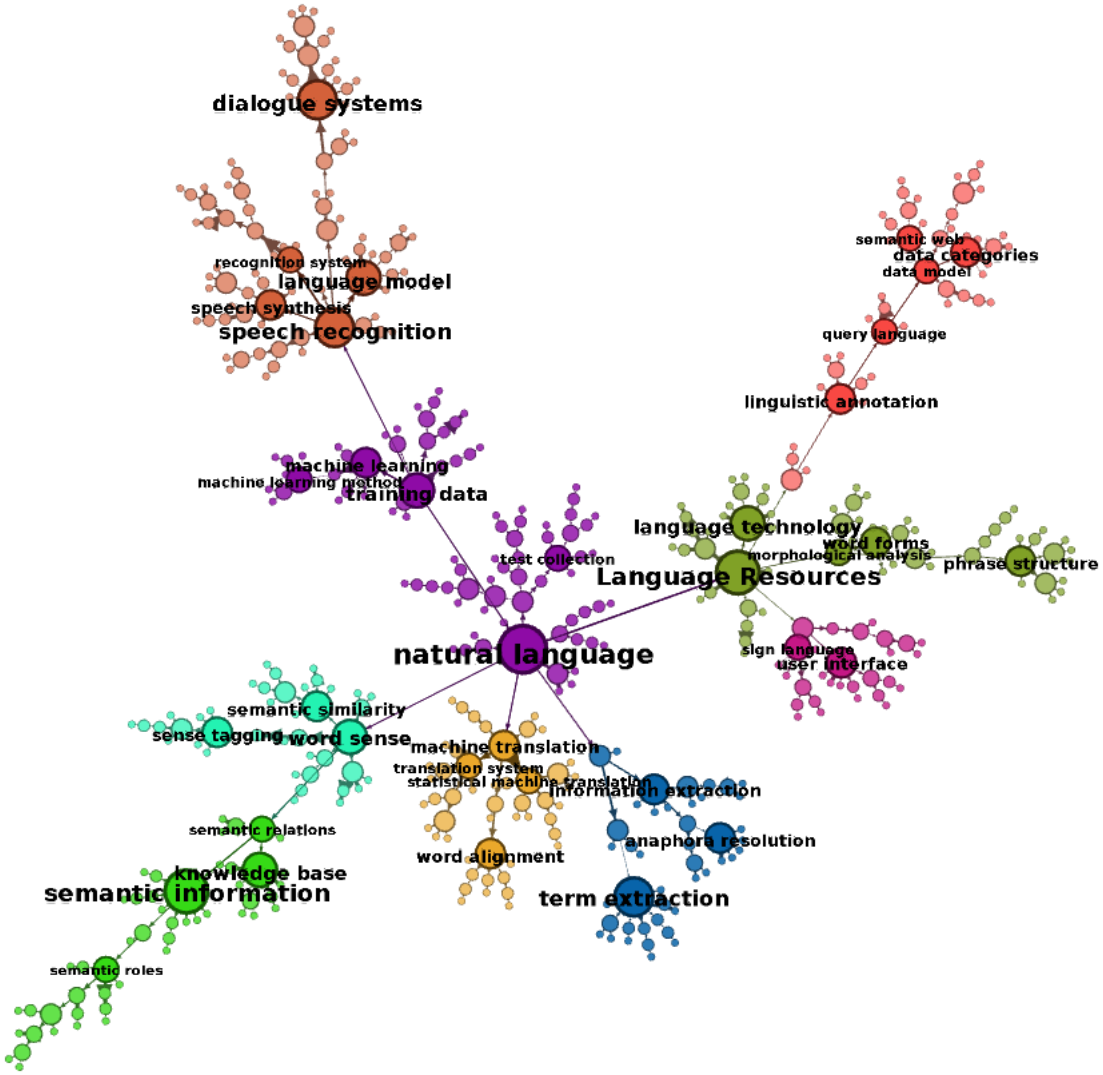


Figure 4: Topical hierarchy of LREC topics

fying 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. The Saffron method for term extraction by use of a domain model is described in more detail in (Bordea et al., 2013).

## 2.2. Topical Hierarchy Construction

Saffron takes into consideration the relations between terms for expert search, by automatically constructing a topical hierarchy of a domain, similar to the one displayed in Figure 4. This structure can be used to measure expertise at different levels of granularity, through inexact matches of expertise. Take for example the ”speech recognition“ subtree, which can be seen in the top-left part of Figure 4. This topical hierarchy identifies the terms ”speech synthesis“ and ”dialogue systems“ as subtopics of ”speech recognition“, providing valuable information for measuring expertise in this field as we will see in Section 2.3..

Topical hierarchies are constructed starting from a list of extracted terms as follows. First, the strength of the relationship between two research terms is measured by counting the number of documents where the two terms are mention together, normalised by the number of documents where each term appears independently. Then, edges are added in a graph where nodes are research terms for all the pairs that appear together in at least three documents. Saffron uses a global generality measure to direct the edges from generic concepts to more specific ones. This step results in a highly dense and noisy directed graph that is further trimmed using an optimal branching algorithm. An optimal branching is a rooted tree where every node but the root has in-degree 1, and that has a maximum overall weight. This yields a tree structure where the root is the most generic term and the leaves are the most specific terms.

## 2.3. Expert Finding

Expert finding is the task of identifying the most knowledgeable person for a given topic. In this task, several competent people have to be ranked based on their relative expertise on a topic. Documents written by a person can be

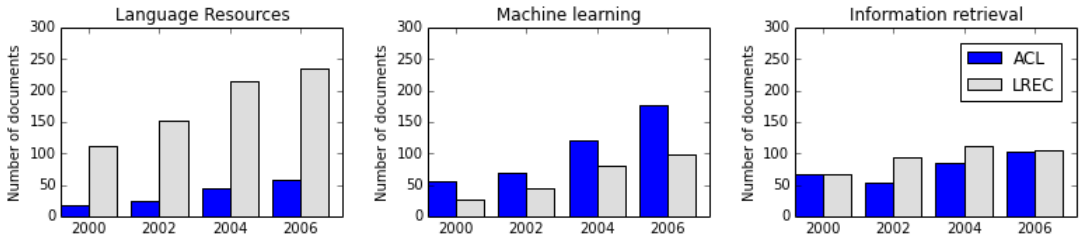


Figure 5: Examples of widely mentioned topics based on number of documents

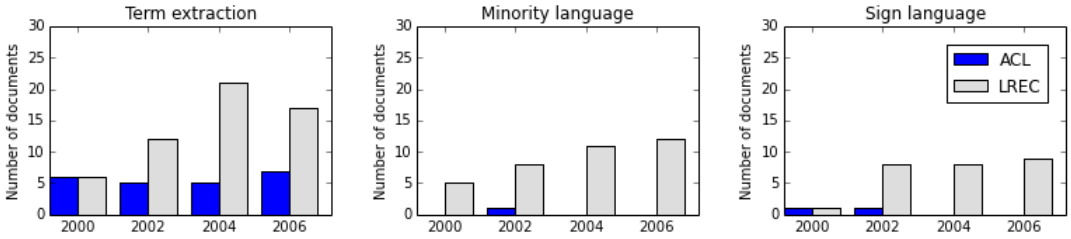


Figure 6: Examples of sparsely mentioned topics based on number of documents

used as an indirect evidence of expertise, assuming that an expert often mentions his areas of interest. Saffron considers various measures of expertise to rank individuals including the relevance of a term for a person, their experience in a domain, as well as their area coverage (i.e., knowledge of sub-topics in a domain).

First, we consider the standard measure of relevance TF-IDF to measure the relevance of a given term for a person. Each person is represented by an aggregated document that is constructed by concatenating all the documents authored by a person. Therefore, the relevance score $R(i,t)$ that measures the interest of an individual $i$ for a given topic $t$ is defined as:

$$R(i,t) = tfidf(t,i) \tag{1}$$

Expertise is closely related to the notion of experience, assuming that the more a person works on a topic, the more knowledgeable they are. This performance indicator is similar to the frequency indicator mentioned in (Paquette, 2007). We estimate the experience of a person based on the number of documents that they wrote about a topic. It is only those documents for which a term is extracted as a top ranked keyphrase that are considered. Let $D_{i,t}$ be the set of documents authored by the individual $i$, that have the term $t$ as a keyphrase. Then, the experience score $E(i,t)$ is defined as:

$$E(i,t) = |D_{i,t}| \tag{2}$$

where $|D_{i,t}|$ is the cardinality, or the total number of documents, in the set of documents $D_{i,t}$.

Both the relevance score and the experience score rely on query occurrences alone, but the relations between topics, as identified in a topical hierarchy, can provide valuable information for further improving expert finding results. A topical hierarchy, such as the one constructed in Section 2.2., can provide valuable information for improving expert finding results. When the subtopics of a term are known, we can evaluate the expertise of a person based on their knowledge of specialised fields.

A previous study showed that experts have increased knowledge at more specific category levels than novices (Tanaka and Taylor, 1991). We introduce a novel measure for expertise called Area Coverage that measures whether an expert has in depth knowledge of a term. Let $Desc(t)$ be the set of descendants of a node $t$, then the Area Coverage score $C(i,t)$ is defined as:

$$C(i,t) = \frac{|\{t' \in Desc(t) : t \in p(i)\}|}{|Desc(t)|} \tag{3}$$

where $p(i)$ is the profile of an individual $i$ constructed using the method presented in the following section. In other words, Area Coverage is defined as the proportion of descendants of a query that appear in the profile of a person. Area coverage is larger than zero only for topics that have more than one descendant, therefore this measure does not contribute to finding experts for specialised topics that appear as leaves in a topical hierarchy.

Finally, the score $REC(i,t)$ used to rank people for expert finding is defined as follows:

$$REC(i,t) = R(i,t) \cdot E(i,t) \cdot C(i,t) \tag{4}$$

This score combines different performance indicators, measuring the expertise of a person based on the relevance of a term, the number of documents about the given topic, as well as his depth of knowledge of the field, also called Area Coverage.

## 2.4. Expert Profiling

We define a topical profile of a candidate as a vector of terms along with scores that measure the expertise of that

candidate. The expert profile $p$ of an individual $i$ is defined as:

$$p(i) = \{S(i,t_1), S(i,t_2), ..., S(i,t_n)\} \quad (5)$$

where $t_1$, $t_2$,...,$t_n$ are the expertise topics extracted from a domain-specific corpus.

A first step in constructing expertise profiles is to identify terms that are appropriate descriptors of expertise. A large number of terms can be extracted for each document, but only the top ranked ones are considered for expert profiling. These are assigned to documents by combining the overall termhood rank of a candidate term and the relevance for each document, as described in the previous paragraph. Once a list of terms is identified, we proceed to the second step of expert profiling, the assignment of scores to each term for a given expert. We rely on the notion of relevance, effectively used for document retrieval, to associated terms with researchers. A researchers interests and expertise are inferred based on their publications. Each term mentioned in one of these publications is assigned to their expertise profile using an adaptation of the standard information retrieval measure TF-IDF. The set of documents authored by a researcher is aggregated in a virtual document, allowing us to compute the relevance of a term over this virtual document.

A term is added to the expert profile of a person using the following scoring function:

$$S(i,t) = termhood(t) \cdot tfidf(t,i) \quad (6)$$

Where $S(i,t)$ represents the score for an expertise topic $t$ and an individual $i$, $termhood(t)$ represents the rank computed in Section 2.1. for the topic $t$ and $tfidf(t,i)$ stands for the TF-IDF measure for the topic $t$ on the aggregated document of an individual $i$. In this way, we construct profiles with terms that are representative for the domain as well as highly relevant for a given individual.

## 3. A Comparative Analysis of ACL and LREC

We used the Saffron system described above for a comparative analysis of the leading NLP conferences, "ACL" (including ACL, ANLP, COLING, EACL, HLT) and "LREC". We restricted the analysis to the years 2000 to 2006 as the ACL Anthology data is restricted to this date range. On the other hand, given the biannual nature of LREC we analyzed both data sets only for the years that LREC took place. Our analysis is concerned with the identification of research topics that are more or less prominent in these conferences. For instance, as we discussed already before, the topic of "language resources" is very prominent in LREC and less so in ACL. However, as can be seen from the first graph on the left in Figure 5 this topic is becoming more prominent in ACL over time as well. The graphs were constructed by selecting two sets of topics that are either widely (Figure 5) or sparsely (Figure 6) mentioned in both data sets. The figures show the number of documents, which mention the particular topic for each year. Note that the scale for widely mentioned topics is much larger than that of the graphs on sparsely mentioned topics.

As may be expected, topics such as "language resources", "minority language" and "sign language" are well represented in LREC, whereas a topic such as "machine learning" is represented more strongly in ACL. In fact, there seems to be no research reported on "minority language" and "sign language" at all at ACL conferences for the years 2004 to 2006. Other topics such as "information retrieval" are represented equally well in ACL and LREC. Interestingly, a topic such as "term extraction", which should be of equal relevance to both communities, is nevertheless more clearly represented at LREC.

Our analysis seems to indicate that the two communities have a complementary research agenda, with ACL focusing on algorithmic approaches to NLP tasks using methods from machine learning, information retrieval etc. whereas LREC has a focus on resource development to be used in combination with such approaches.

## 4. Evaluation

Several data sets for evaluating expert search systems are publicly available (Bailey et al., 2007; Balog et al., 2007; Soboroff et al., 2007), providing gold standard assignments of expertise that are gathered through self-assessment or by asking the opinion of co-workers. These evaluation datasets have multiple limitations, as self-assessed expert profiles are subjective and incomplete, and the opinions of colleagues are biased towards their social and geographical network. To address these challenges, a more recent dataset (Bordea et al., 2012) exploits the information about program committees from different workshops in Computer Science. In our experiments we make use of a subset of this dataset which covers 340 workshops in Computational Linguistics. In average there are almost 25 program committee members associated with each workshop. These experts are associated with 4,660 unique topics manually extracted from each call for papers.

Evaluation measures initially proposed for document retrieval can be used to evaluate the expert finding and the expert profiling tasks. These tasks are evaluated based on the quality of ranked lists of topics and experts, respectively, which is not different from evaluating a ranked list of documents. The most basic evaluation measures used in information retrieval are precision and recall. In our experiments we make use of the following measures of effectiveness:

**Precision at N (P@N)** This is the precision computed when $N$ results are retrieved, which is usually used to report early precision at top 5, 10, or 20 results.

**Average Precision (AP)** Precision is calculated for every retrieved relevant result and then averaged across all the results.

**Reciprocal Rank (RR)** This is the reciprocal of the first retrieved relevant document, which is defined as 0 when the output does not contain any relevant documents.

To get a more stable measurement of performance, these measures are commonly averaged over the number of

queries. In our experiments, we report the values for the Mean Average Precision (MAP), and Mean Reciprocal Rank (MRR). In this setting, recall is less important than achieving a high precision for the top ranked results. It is more important to recommend true experts than to find all experts in a field.

The approach proposed in this paper are evaluated against two information retrieval methods for expert finding. Both methods model documents and expertise topics as bags of words and take a generative probabilistic approach (Balog et al., 2009).

| Measure | LM1 | LM2 | Saffron |
|---|---|---|---|
| MAP | 0.0071 | 0.0056 | **0.0340** |
| MRR | 0.0631 | 0.0562 | **0.2754** |
| P@5 | 0.0202 | 0.0173 | **0.1347** |

Table 1: Expert finding results for the language modelling approach (LM) and Saffron

The results of our experiments are shown in Table 1. The language modelling approaches fail to identify experts because a much larger number of topics is available in our dataset than previously considered (Balog et al., 2009). The Saffron approach, which makes use of a topical hierarchy, consistently achieves higher results based on all the considered evaluation measures.

## 5. Conclusion

The analysis methods and tools provided by our approach enable us to do a comparative study of topic occurrence in the two data sets, as shown above. However, combining this with the community and topical hierarchy analysis discussed above as well, we will be able to draw even broader conclusions about the community and research agenda development. Finally, we would argue that such studies will provide more insight into both NLP communities and will help to guide publication and participation strategies, especially of novice researchers in the field.

It can be argued that it is not only the number of documents that indicates expertise, but the quality of those documents as well. For example, in a peer-review setting, the impact of a publication measured using citation counts is often used as an indicator of publication quality. Similarly, page rank can be used as a quality indicator for web pages, the number of comments for blogs, the number of retweets for tweets, the number of followers for users. But each of these indicators is specific to content type and have to be investigated separately depending on the domain, therefore we leave the integration of document quality measures for future work.

## 6. Acknowledgements

This work has been funded in part by a research grant from Science Foundation Ireland (SFI) under Grant Number SFI/12/RC/2289 (INSIGHT).

## 7. References

Bailey, P., Craswell, N., de Vries, A. P., and Soboroff, I. (2007). Overview of the trec 2007 enterprise track draft. In *TREC 2007 Working notes*.

Balog, K., Bogers, T., Azzopardi, L., de Rijke, M., and van den Bosch, A. (2007). Broad expertise retrieval in sparse data environments. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, SIGIR '07, pages 551–558, New York, NY, USA. ACM.

Balog, K., Azzopardi, L., and de Rijke, M. (2009). A language modeling framework for expert finding. *Information Processing & Management*, 45(1):1–19.

Basili, R., Moschitti, A., and Pazienza, M. T. (2002). Empirical investigation of fast text classification over linguistic features. In *ECAI*, pages 485–489.

Bordea, G., Bogers, T., and Buitelaar, P. (2012). Benchmarking domain-specific expert search using workshop program committees. In *Workshop on Computational Scientometrics: Theory and Applications, at CIKM*.

Bordea, G., Buitelaar, P., and Polajnar, T. (2013). Domain-independent term extraction through domain modelling. In *the 10th International Conference on Terminology and Artificial Intelligence (TIA 2013), Paris, France*. 10th International Conference on Terminology and Artificial Intelligence.

Hooper, C. J., Bordea, G., and Buitelaar, P. (2013). Web science and the two (hundred) cultures: representation of disciplines publishing in web science. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 162–171. ACM.

Kietz, J.-U., Volz, R., and Maedche, A. (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*, pages 167–175. Association for Computational Linguistics.

Lopez, P. and Romary, L. (2010). Humb: Automatic key term extraction from scientific articles in grobid. In *Proceedings of the 5th international workshop on semantic evaluation*, pages 248–251. Association for Computational Linguistics.

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

Paquette, G. (2007). An ontology and a software framework for competency modeling and management. *Educational Technology & Society*, 10(3):1–21.

Soboroff, I., de Vries, A. P., and Craswell, N. (2007). Overview of the trec 2006 enterprise track. In *The fifteenth Text REtrieval Conference Proceedings (TREC 2006)*.

Tanaka, J. W. and Taylor, M. (1991). Object categories and expertise: Is the basic level in the eye of the beholder? *Cognitive Psychology*, 23(3):457–482, July.

Yangarber, R., Grishman, R., Tapanainen, P., and Huttunen, S. (2000). Automatic acquisition of domain knowledge for information extraction. In *Proceedings of the 18th conference on Computational linguistics-Volume 2*, pages 940–946. Association for Computational Linguistics.