Towards a Framework for Comparing Automatic Term Recognition Methods

Petr Knoth*, Marek Schmidt†, Pavel Smrž† and Zdeněk Zdráhal*

* *The Open University, Knowledge Media Institute*
{p.knoth, z.zdrahal}@open.ac.uk
† *Brno University of Technology, Faculty of Information Technology*
{ischmidt, smrz}@fit.vutbr.cz

Abstract. Automatic Term Recognition focuses on the extraction of words and multi-word expressions that are significant for a given domain. There is a considerable interest in using ATR for automatic metadata generation, creation of thesauri and terminological glossaries, keyword extraction, ontology building, etc. In this paper, we build upon the work done at the University of Sheffield, where a library with a few algorithms for ATR was recently developed. We enrich this library with new ATR algorithms and tools for evaluation. Our aim is to perform an experimental study comparing the base ATR methods as well as their combinations under various conditions. The results of the study indicate that better precision can be usually reached by combining ATR methods using foreground and ATR methods using background knowledge. The created platform is freely available and prepared for extensions by other researchers.[1]

1 Introduction

The roots of Automatic Term Recognition (ATR) date back to late 80s when the need for automatic extraction of terminological units from specialized texts became acute in various fields [3]. The amount of unstructured data in the electronic form has grown rapidly from that time. This encouraged further researchers to employ ATR for the tasks of automatic creation of thesauri, keyword extraction, glossary or index generation, tag suggestion, etc. Recently, ATR systems has gained popularity in the Semantic Web community as the first step in the automatic building of ontologies [4]. The results of ATR have also been successfully applied in information retrieval, machine translation and many other domains [12, 17, 6].

There are ATR systems available via a web-based user interface [18] or as a Web service [2, 20] today. Some of them try to exploit the additional information provided by the annotation of particular formats. For example, Term Extraction SEO Tool [18] focuses on HTML documents and applies certain weight to particular HTML elements to determine what could be the most descriptive or targeted terms. Unfortunately, the access policy of the online tools often disallow researchers to experiment with the implemented methods and discourages advanced processing of the output to refine the results.

The background of the work reported in this paper is given by two European projects – KiWi (Knowledge in a WiKi) and Eurogene (Pan-European Learning Service in the Field of Genetics). The aim of the KiWi project is to design and develop an advanced knowledge management system based on the semantic wiki technology and extend it by information extraction, personalisation, and reasoning. The objective of Eurogene is to establish a European reference portal that will support development and reuse of multimedia educational content in genetics. The project takes advantage of the emerging Semantic Web technologies supported by tools for text analysis, collaborative annotation of content, machine translation, advanced multilingual search and navigation.

The general tasks that will benefit from the ATR methods are shared across the projects:

[1] `http://code.google.com/p/jajatr/`

- *keyword extraction* – ATR will assist the user in enriching the content with metadata. This will enable advanced searching facilities.
- *ontology enrichment* – ATR will identify new concepts from the uploaded content. The concepts can be included into the ontology in order to keep the conceptualization up-to-date.

Having this context in mind, we were to choose and apply the state-of-the-art ATR algorithms that are most appropriate for our purposes. However, our comprehensive survey [11] revealed that, despite its popularity, the field still lacks proper comparative studies. Only a few methods have been evaluated and compared in terms of their precision. The rest of the developed tools is assessed just by an observation, often concluding that "it provides reasonably good results".

As in many other domains, it is reasonable to expect that there will be no "best" ATR method which would outperform others on all data sets and in all circumstances. To compare various ATR algorithms in realistic conditions, one therefore needs not only a referential implementation of a given set of ATR methods and necessary pre-processing tools (ideally available as an open source), but also annotated data to evaluate on.

It is also important to note that the evaluation criteria themselves depend on the task in hand. For example, the concept of keyword annotation of documents changed with the development of information retrieval in the last decades. Nowadays, annotators often see keywords as additional contextual information that can help non-standard terminology searches rather than repeating the terms used in the document title or abstract. Thus, the comparative studies of the ATR techniques need also evaluation tools that implement task-specific measures related to the annotated data in question.

This paper presents our effort to build an ATR evaluation framework reflecting the abovementioned parameters. Rather than develop it from scratch, we decided to reuse an ATR library that was recently developed at the University of Sheffield and is available as an open source [21]. Our contributions done on the top of the original work can be summarized in the following items:

- implementation of 3 ATR statistical methods (TF, RIDF and LR as described later in the text);
- development of an automatic evaluation tool;
- refactoring of the library (we had to fix quite a few bugs and added the possibility to choose a particular corpus as a background).

The rest of the paper is organized as follows: The next section outlines the theoretical foundations of the implemented methods. Section 3 presents an example of the experimental evaluation on the GENIA and Eurogene corpora. We conclude the paper with the discussion on the necessary steps to go beyond the current state-of-the-art in ATR.

2 Statistical ATR Methods

A typical approach of the advanced ATR methods consists of two phases:

- Linguistic phase employs a linguistic filter, based on part-of-speech (POS) tags, to extract a set of candidate terms. Term variant recognition techniques can be applied to associate different realizations of one term with its root form.
- Statistical phase uses a statistical method to assign a weight to each candidate term.

Linguistic methods use the linguistic knowledge on term formation to find terms in a text. They are generally language-dependent. The framework currently uses OpenNLP software package with POS tagger and a noun phrase chunker. Noun phrase chuncks are considered term candidates.

A concept may have many different surface realizations. For example 'human clones' and 'clones of human' could be considered as term variants. Identification of the term variants can have a positive impact on the results of ATR methods [15]. Several types of term variations are usually distinguished – orthographic, morphological, structural, acronyms, abbreviations, lexical synonyms, etc.

To measure the 'strength' of a candidate term, two characteristics are usually distinguished – *termhood* and *unithood*:

- Termhood is a measure of the degree by which a linguistic unit is related to the domain-specific concept. Termhood methods are based on the frequency of occurrence [10].
- Unithood is relevant for complex terms which consist of more linguistic units (words). It measures the collocation strength of the units. The basic idea of determining unithood consists in measuring significance of the words occurring together. Standard statistical techniques such as mutual information, t-test or log-likelihood are generally put to use [21, 7].

ATR methods can be also divided according to the use of background knowledge, i.e. a corpus in a general domain. Table 1 shows the classification of statistical measures that will be discussed in this section. Later, we will also discuss hybrid approaches that try to combine these measures.

	Termhood	**Unithood**
Only domain knowledge	TF, TFIDF, RIDF, DC	C-Value, LC
Background knowledge	Weirdness, LR, DR	

Table 1. Classification of statistical methods

The following paragraphs briefly introduce particular ATR methods implemented in our framework that took part in the experimental evaluation reported in the next section.

Term Frequency (TF) is the count of all occurrences of the candidate term in a corpus. Frequent terms are supposed to be more important. This simple method is used in systems to rank term candidates generated by linguistic pre-processing [6]. We compute term frequency Tf_i as a normalized frequency of term i in the document collection:

$$Tf(i) = \frac{f(i)}{\sum_k f(k)},$$

where f_i is the number of words i in the collection.

Term Frequency – Inverse Document Frequency (TFIDF) is a weighting score used often in information retrieval, where it corresponds to the fact that the most significant words for a document tend to occur frequently in that document, despite their possibly rare occurrence in the whole collection. Inverse document frequency $Idf(i)$ measures the general importance of term i in the collection of documents D by counting the number of documents which contain term i:

$$\begin{aligned} Idf(i) &= log \frac{|D|}{|\{d_j : t_i \in d_j\}|} \\ TfIdf(i) &= Tf(i).Idf(i) \end{aligned}$$

Note that in the context of ATR we can prefer to compute a single ranked list of terms rather than a list of terms for each file in the domain-specific collection. Therefore, we can compute *Tfidf*(i) as $Tf(i).Idf(i)$ considering $Tf(i)$ as the term frequency of word i in the domain collection. Roughly speaking, calculating the term frequency as there would be only one document in the domain-specific collection. The *Tfidf*(i) weighting score measures the termhood with respect to the documents in a collection. In ATR, it is often used as a baseline [21] or as one of several features to determine the termhood [14].

Residual IDF (RIDF) is an alternative to IDF, which looks for terms whose document frequency is larger than chance. More precisely, RIDF is defined as the difference between logs of actual document frequency and document frequency predicted by Poisson distribution [13].

$$RIDF(i) = Idf(i) - \log(1 - p(0; \lambda(i))),$$

where p is the Poisson distribution with parameter $\lambda(i) = \frac{f(i)}{D}$ (the average number of occurrences of word w_i per document). $f(i)$ is the number of words i in the collection. $1 - p(0; \lambda(i))$ is the Poisson probability of a document with at least one occurrence of i.

Weirdness measure is based on the idea that distribution of terms in a specialized corpus (domain) and in a general corpus (background) significantly differ [1]. This is expressed by the following formula:

$$Weirdness(i) = \frac{\frac{f_s(i)}{n_s}}{\frac{f_g(i)}{n_g}},$$

where $f_s(i)$ and $f_g(i)$ are the frequencies of word i in the specialized and the general corpus respectively, n_s and n_g are total numbers of words in the respective corpora. The original Weirdness was defined for one-word terms only, so we compute a geometric average of weirdnesses of each word in the term.

Likelihood Ratio (LR) [13] is one of the methods we have newly implemented in the framework. The motivation is the same as in the case of weirdness. As opposed to weirdness, however, a statistical test is employed to measure the significance of difference between word frequencies in the domain and those in the background corpus. The first hypothesis is that the probability of observing a given word in the background is equal to the probability of observing it in our domain. The second hypothesis is that the probability of seeing a given word in the domain is significantly higher than seeing it in the background. We assume binomial distribution for word frequencies.

$$p = \frac{f_s + f_g}{n_s + n_g} \qquad p_s = \frac{f_s}{n_s} \qquad p_g = \frac{f_g}{n_g}$$

$$LR = \log L(f_s, n_s, p) + \log L(f_g, n_g, p) - \log L(f_s, n_s, p_s) - \log L(f_g, n_g, p_g)$$

$$L(k, n, x) = x^k (1 - x)^{n-k}$$

Although Likelihood Ratio has been recently used in the related field of text summarization [9], there is no quantitative evaluation of the method in the context of ATR to the best of our knowledge.

C-Value Method is a unithood method which has been used for term recognition in the medical domain, which typically contains a large number of complex terms [8].
The formula to compute it is based on three principles – extracting the most frequent terms, penalizing the nested terms that occur as a substring of a longer candidate term, and considering the length of the candidates (the number of the words they consist of):

$$\textit{C-value}(a) = \begin{cases} log_2|a| \cdot f(a) & \text{if } a \text{ is not nested} \\ log_2|a| \cdot (f(a) - \frac{1}{|T_a|} \sum_{b \in T_a} f(b)) & \text{otherwise} \end{cases}$$

where a and b are the candidate terms, f denotes the frequency and T_a is the set of candidate terms which contain a.

Glossex Method [12] is based on two heuristics. The first measure evaluates the degree of domain specificity (TD) which is equal to our definition of weirdness.
The second measure investigates the idea of term cohesion. Let $|t| = n$ be the number of words forming term t. The term cohesion can be then expressed as:

$$TC_{D_i}(t) = \frac{n.tf_{t,D_i}.\log tf_{t,D_i}}{\sum_{j=0}^{n} tf_{w_j,D_i}},$$

where w_j is a j^{th} word in term t.
The two measures are combined using two user adjustable coefficients α and β.

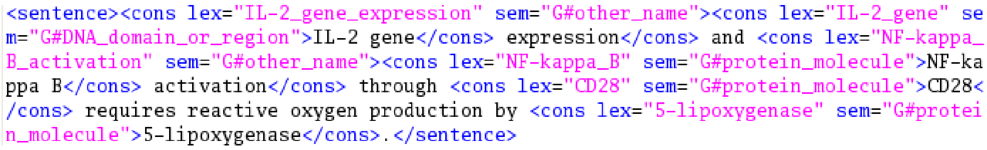
$$GlossEx(t) = \alpha.TD(t) + \beta.TC(t)$$

Combining Statistical Methods It is often advantageous to combine several above-mentioned methods. For example, a mixture of entropy and log-likelihood ratio as measures of unithood and tf.idf characterizing the termhood has been explored in [16]. Simple thresholds on each feature defined the weak classifiers, which were successfully combined by a kind of boosting algorithm. Similar combination of measures is discussed in [19] in the context of term extraction from medical documents in Spanish.

3 Evaluation

As an example of the use of our evaluation framework, we present results of the experiments on two large annotated data sets – the GENIA and Eurogene corpora in this section.

3.1 Experiments on the GENIA Corpus



```
<sentence><cons lex="IL-2_gene_expression" sem="G#other_name"><cons lex="IL-2_gene" se
m="G#DNA_domain_or_region">IL-2 gene</cons> expression</cons> and <cons lex="NF-kappa_
B_activation" sem="G#other_name"><cons lex="NF-kappa_B" sem="G#protein_molecule">NF-ka
ppa B</cons> activation</cons> through <cons lex="CD28" sem="G#protein_molecule">CD28<
/cons> requires reactive oxygen production by <cons lex="5-lipoxygenase" sem="G#protei
n_molecule">5-lipoxygenase</cons>.</sentence>
```

Fig. 1. Example of a GENIA annotation file

GENIA corpus is a collection of biomedical documents that were compiled and annotated within the scope of the GENIA project [5]. The goal of the project was to develop text mining systems for the domain of molecular biology. The annotation process aimed at manual annotation terms in almost 2,000 MedLine abstracts.
Let us discuss the origin of two variants of the evaluation data set extracted from the GENIA corpus. Figure 1 shows an example of an annotated sentence from the corpus. It can be seen that both terms – *IL-2 gene expression* as well as the nested *IL-2 gene* – are considered valid. This approach can be beneficial for some tasks such as ontology building where the nested part of the term can often be interpreted as a hypernym of the complex term. On the other hand, the nested terms are not desirable in other situations as they can inflate the terminological glossaries and refer to general concepts rather than domain-specific ones. Considering the potential dichotomy, we prepared two versions of the "gold standard" list of GENIA terms. The first one contains all the annotated terms (including the nested ones), the second takes only the longest part as a term in the case of nesting.

In the linguistic pre-processing phase, we have extracted 32,521 candidate terms. This set was ranked by the statistical methods. We report the precision of the methods at 3 points (cuts): after first 20 highly ranked terms, after first 200 and after 2000 terms. Although the first may seem to be a very small sample for the evaluation, it is a relevant benchmark when considering ATR for keyword extraction or tag suggestion.

Tables 2 and 3 report the precision achieved by ATR methods during the experiment. The precision is defined as

$$Precision = \frac{\sum_{i=0}^{|Recognized|} |t_i \in Reference|}{|Recognized|} \qquad (1)$$

where *Recognized* is a set of $|Recognized|$ highly ranked terms extracted by the system and $[t_i \in Reference]$ is equal 1 if term t_i is a member of the $Reference$ set (is listed in a reference file of correct terms). Otherwise it is 0. The reference set has been extracted from the Genia corpus and an evaluation tool was developed to easily measure the results.

As in many other fields, ATR can benefit from combinations of the base methods employing various voting strategies. We have experimented with many different combinations and proved the potential boost in precision. Tables 2 and 3 present the results of the base methods as well as the most promising combinations evaluated on the GENIA corpus with the English Gigaword Corpus as the background (for computing weirdness and other measures).

Number of terms	TF	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. Weirdness-TFIDF	Vot. LR - TFIDF	Vot. all
20	0,90	0,90	0,75	0,95	0,70	0,90	0,95	**1,00**	0,90	**1,00**
200	0,76	0,80	0,80	0,85	0,78	0,83	0,87	**0,96**	0,84	0,91
2000	0,70	0,71	0,70	0,63	0,64	0,62	0,67	**0,79**	0,67	0,73

Table 2. Precision on GENIA Corpus (nested terms)

Number of terms	TF	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. Weirdness-TFIDF	Vot. LR - TFIDF	Vot. all
20	0,90	0,90	0,75	0,95	0,65	0,90	0,90	**1,00**	0,90	**1,00**
200	0,75	0,79	0,76	0,84	0,59	0,83	0,85	**0,94**	0,83	0,82
2000	0,67	**0,68**	0,63	0,52	0,47	0,61	0,59	0,67	0,58	0,60

Table 3. Precision on GENIA Corpus (without nested terms)

Considering only the base methods (not their combinations), the C-Value method and LR achieved very good results. This fact is surprising especially with respect to the success of the LR measure that is basically neglected by the ATR community. Another notable point is that the results achieved by TF, which is the simplest method, are not significantly worse than the results of TFIDF and that the method sometimes even outperformed RIDF.

The best performer showed to be the combination of Weirdness and TFIDF, which provided excellent results in both – nested and not-nested settings. The method combining all non-voting methods scored well, but still not as good as voted Weirdness-TFIDF.

As the size of the gold standard for the setting without nested terms is lower than that for the nested terms, it is natural that the values of the precision also decrease. However, the drop in precision is rather small for most of the methods on the first 200 terms. We suppose that the radically different pattern of weirdness in this respect has much to do with the characteristics of the background corpus. Nevertheless, this hypothesis needs to be verified in future work.

In order to inspect the impact of the background corpus size, we run our experiments in two other settings:

1. replacing the English Gigaword Corpus by the British National Corpus (BNC) which is by about one order of magnitude smaller than English Gigaword;
2. without any background data (labelled Null in the following table).

The results of this experiment are reported in Table 4. Only methods that use background are listed, other methods would produce the same results as reported in Table 2. All the experiments were performed in the nested settings. The best results for each corpus and method are highlighted.

Number of terms	LR	Weirdness	Glossex	Vot. Weirdness-TFIDF	Vot. LR - TFIDF	Vot. all
English Gigaword						
20	**0,95**	0,70	0,90	**1,00**	**0,95**	**1,00**
200	**0,89**	**0,78**	0,83	**0,96**	0,88	0,90
2000	0,65	0,64	0,62	**0,79**	0,69	**0,75**
BNC						
20	**0,95**	**0,80**	**0,95**	**1,00**	**0,95**	**1,00**
200	0,87	0,69	0,84	0,95	**0,89**	**0,92**
2000	0,62	0,63	0,61	0,80	0,68	0,72
MedLine						
20	**0,95**	0,75	0,75	0,95	**0,95**	**1,00**
200	**0,89**	0,71	0,67	0,89	0,88	**0,92**
2000	0,53	0,57	**0,65**	0,75	0,65	0,73
Null						
20	0,85	0,70	0,90	0,95	0,90	**1,00**
200	0,75	0,61	0,66	0,85	0,78	0,85
2000	**0,70**	0,49	0,50	0,66	**0,70**	0,66
English Gigaword + BNC						
20	**0,95**	0,75	0,90	**1,00**	**0,95**	**1,00**
200	**0,89**	0,77	**0,85**	**0,96**	0,88	0,91
2000	0,65	**0,67**	0,63	**0,79**	0,69	**0,75**

Table 4. Impact of different sizes of the background corpus

The results show that there is not a significant difference in using English Gigaword and BNC corpora. Even using both one cannot expect significant improvements in precision. However, using no background knowledge significantly deteriorates the performance. Naturally, voting mechanisms are more robust since the fall of one method can be compensated by the other one.

3.2 Evaluation on the Eurogene Corpus

The ATR methods have been also tested one the resources developed within the Eurogene project. So far, we have collected 210 presentations used mainly for teaching genetics at the university level. First, we converted the presentations into plain text. The size of the whole corpus is approximately 4 MB (600,000 words). The terms are not annotated in the texts so we asked domain experts to evaluate the results of the compared ATR methods.

During the linguistic phase, 34,617 candidate terms were extracted. They were ranked and sorted using each particular method. Then, we asked two experts from different branches of genetics to inspect first 100 terms produced by each method. Their task was to decide which terms are characteristic for the genetic domain.

The task may seem simple, but the domain experts found it ill-defined. The lack of a precise definition of "the characteristic domain term" showed to be the major problem. Some terms, such as *p-value*, are terms of a specific branch of genetics (here, statistical hypothesis testing). These terms were considered differently by statistical geneticist and by clinical or molecular geneticist. Also, there were discussions on the terms found to be too general that were, finally, not accepted as proper terms (for example, *genetics*). The evaluators also found it difficult to be consistent across large set of results, In order to increase their consistency they had to evaluate the same results more than once.

Number of terms	TFIDF	RIDF	LR	Weirdness	Glossex	C-Value	Vot. Weirdness-TFIDF	Vot. LR - TFIDF
100	0. 70	0.63	0.60	0.79	0.75	0.66	**0.98**	0.49

Table 5. Precision on Eurogene corpus

The results of the experiment are reported in Table 5. As in the case of the GENIA corpus, we found that the method combining Weirdness with TFIDF provided the best precision. Other methods usually scored significantly lower. As these results were not expected, we asked the domain experts to assess the extracted terms from the qualitative point of view as well.

They found that the results of the Weirdness algorithm capture the important domain characteristics. At the same time, there were a few essential flaws in the output. Typical errors contained a name of an organization or a name of the author. This happens due to the absence of these terms in the general corpora and their high frequency in the domain-specific content. The high frequency of authors' names was due to the name re-occurring in the footer of each slide of their presentations. Such presentation style naturally results in generating noise for the statistical methods.

The TFIDF algorithm produced a list of terms which were probably characteristic for certain documents within the Eurogene corpus, but were often too general to be considered as domain-specific terms. We expect that this is caused by the fact that the TFIDF calculation does not involve any background knowledge.

The list of terms extracted using the combination of both the methods differs from those given by Weirdness and TFIDF separately. The extracted terms mainly consist of names of genes, substances and specific genetic terms. The combination produced significantly higher precision than the components.

4 Conclusions and Future Directions

The ATR evaluation framework discussed in this paper proved to be extremely useful for fast hypothesis formulation and testing. We have implemented new statistical ATR methods and compared their performance on the two included corpora. Many experiments have been also run with different combinations of statistical measures. The best results on both corpora were achieved by combination of Weirdness and TFIDF measures, which produced substantially better results than other methods.

The results were also inspected from the qualitative point of view. This leads to the conclusion that methods combining domain specific knowledge with background knowledge are generally more robust than methods using only one of these sources.

The results of our experiments on the Genia corpus are fully reproducible since all the source codes, the data and the software for evaluation can be downloaded. We would like to encourage

other researchers to contribute to the framework. It is especially important to add new evaluation data sets on which ATR techniques can be tested. The community could keep the set of statistical algorithms up-to-date as new approaches will arise. A web-based user interface can also be implemented in order to allow non-programmers to try and evaluate the system.
From the research point of view, we agree with [21] that many of the items identified as terms fall into the category that Information Extraction (IE) traditionally extracts from texts. For example, names of genes, diseases, substances, methods, etc. The employment of the IE techniques including both – traditional machine learning and weakly-learning techniques (active learning, co-learning, or expansion) could significantly improve the precision. ATR and IE techniques can also co-operate. For example, the extraction of names of people and organizations is a typical task of IE. The result could be used to filter the list of candidate terms and thus to solve the problems mentioned in Section 3.2.

5 Acknowledgement

This work has been partially supported by the Eurogene (The First Pan-European Learning Service in the Field of Genetics, Contract no. ECP-2006-EDU-410018) and KiWi (Knowledge in a WiKi, 211932) projects. We would also like to thank Dr Domenico Coviello of the European Society of Human Genetics and Dr Dawn Teare of the Sheffield University for their contribution to the evaluation of the genetic results.

References

1. Ahmad, K., Gillam, L., and Tostevin, L. University of Surrey participation in TREC 8: Weirdness indexing for logical document extrapolation and retrieval (WILDER).
2. Anadianou, S. TerMine. http://www.nactem.ac.uk/software/termine/.
3. Castellví, M. T., Bagot, R. E., and Palatresi, J. V. Automatic term detection: A review of current systems. In *Recent Advances in Computational Terminology*, D. Bourigault, C. Jacquemin, and M.-C. L'Homme, Eds. John Benjamins, Amsterdam/Philadelphia, 2001, pp. 53–88.
4. Cimiano, P. *Ontology Learning and Population from Text: Algorithms, Evaluation and Applications.* Springer-Verlag New York, Inc., Secaucus, NJ, USA, 2006.
5. Collier, N., Park, H. S., Ogata, N., Tateishi, Y., Nobata, C., Ohta, T., Sekimizu, T., Imai, H., Ibushi, K., and ichi Tsujii, J. The genia project: corpus-based knowledge acquisition and information extraction from genome research papers. In *In Ninth Conference of the European Chapter of the Association for Computational Linguistics (EACL-99* (1999), pp. 271–272. http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA.
6. Dagan, I., and Church, K. Termight: Identifying and translating technical terminology. In *Proceedings of the fourth conference on applied natural language processing* (San Francisco, CA, USA, 1994), Morgan Kaufmann Publishers Inc., pp. 34–40.
7. Daille, B., Éric Gaussier, and Langé, J.-M. Towards automatic extraction of monolingual and bilingual terminology. In *Proceedings of the 15th conference on Computational linguistics* (Morristown, NJ, USA, 1994), Association for Computational Linguistics, pp. 515–521.
8. Frantzi, K., Ananiadou, S., and Mima, H. Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries V3*, 2 (2000), 115–130.
9. Gupta, S., Nenkova, A., and Jurafsky, D. Measuring importance and query relevance in topic-focused multi-document summarization. In *ACL* (2007), The Association for Computer Linguistics.

10. Kageura, K., and Umino, B. Methods of automatic term recognition: A review. *Terminology 3*, 2 (1996), 259–289.
11. Knoth, P., Schmidt, M., and Smrž, P. KiWi project – Information Extraction – State of the Art, 2008.
12. Kozakov, L., Park, Y., Fin, T., Drissi, Y., Doganata, Y., and Cofino, T. Glossary extraction and utilization in the information search and delivery system for ibm technical support. *IBM Syst. J. 43*, 3 (2004), 546–563.
13. Manning, C. D., and Schütze, H. *Foundations of Statistical Natural Language Processing.* The MIT Press, June 1999.
14. Medelyan, O., and Witten, I. H. Thesaurus based automatic keyphrase indexing. In *JCDL '06: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital libraries* (New York, NY, USA, 2006), ACM, pp. 296–297.
15. Nenadié, G., Ananiadou, S., and McNaught, J. Enhancing automatic term recognition through recognition of variation. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics* (Morristown, NJ, USA, 2004), Association for Computational Linguistics, p. 604.
16. Patry, A., and Langlais, P. Corpus-based terminology extraction. http://www.iro.umontreal.ca/˜felipe/Papers/paper-tke-2005.pdf.
17. Peas, A., Verdejo, F., and Gonzalo, J. Corpus-based terminology extraction applied to information access, 2001.
18. Sclano, F., and Velardi, P. TermExtractor: A web application to learn the shared terminology of emergent web communities. In *3rd International Conference on Interoperability for Enterprise Software and Applications (I-ESA 2007)* (2007).
19. Vivaldi, J., Màrquez, L., and Rodríguez, H. Improving term extraction by system combination using boosting. *Lecture Notes in Computer Science 2167* (2001), 515–521.
20. Yahoo! Content analysis web services: Term Extraction. http://developer.yahoo.com/search/content/V1/termExtraction.html.
21. Ziqi Zhang, Jose Iria, C. B., and Ciravegna, F. A comparative evaluation of term recognition algorithms. In *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)* (Marrakech, Morocco, may 2008), E. L. R. A. (ELRA), Ed.