# A Comparative Evaluation of Term Recognition Algorithms

**Ziqi Zhang, José Iria, Christopher Brewster and Fabio Ciravegna**
Department of Computer Science,
University of Sheffield, Sheffield, S1 4DP
Initial.LastName@dcs.shef.ac.uk

### Abstract

Automatic Term recognition (ATR) is a fundamental processing step preceding more complex tasks such as semantic search and ontology learning. From a large number of methodologies available in the literature only a few are able to handle both single and multi-word terms. In this paper we present a comparison of five such algorithms and propose a combined approach using a voting mechanism. We evaluated the six approaches using two different corpora and show how the voting algorithm performs best on one corpus (a collection of texts from Wikipedia) and less well using the Genia corpus (a standard life science corpus). This indicates that choice and design of corpus has a major impact on the evaluation of term recognition algorithms. Our experiments also showed that single-word terms can be equally important and occupy a fairly large proportion in certain domains. As a result, algorithms that ignore single-word terms may cause problems to tasks built on top of ATR. Effective ATR systems also need to take into account both the unstructured text and the structured aspects and this means information extraction techniques need to be integrated into the term recognition process.

## 1. Introduction

Automatic Term Recognition (ATR) is an important research area that deals with the extraction of technical terms from domain-specific language corpora. ATR is often a processing step preceding more complex tasks, such as semantic search (Bhagdev *et al.* 2007) and especially ontology engineering (Park, Byrd, & Boguraev 2003; Brewster *et al.* 2007).

There have been many studies into ATR. In the majority of these studies (Ananiadou 1994; Bourigault 1992; Fahmi, Bouma, & van der Plas 2007; Frantzi & Ananiadou 1999; Wermter & Hahn 2005) linguistic processors (e.g. POS tagger, phrase chunker) are used to filter out stop words and restrict candidate terms to nouns or noun phrases, while in others any n-gram sequences are selected as candidate terms (Deane 2005). Statistical measures are then used to rank the candidate terms. These measures can be categorised into two kinds: measures of 'unithood' indicating the collocation strength of units that comprise a single term; and measures of 'termhood' indicating the association strength of a term to domain concepts. For measuring 'unithood' measures such as mutual information (Daille 1996), log likelihood (Cohen 1995), t-test (Fahmi, Bouma, & van der Plas 2007; Wermter & Hahn 2005), and the notion of 'modifiability' and its variants (Caraballo & Charniak 1999; Deane 2005; Wermter & Hahn 2005) are employed. In contrast, measures for 'termhood' are circumscribed to frequency-based approaches and the use of reference corpora: the classic TFIDF used in (Evans & Lefferts 1995; Medelyan & Witten 2006); the notion of 'weirdness' as introduced in (Ahmad, Gillam, & Tostevin 1999), which compares the term frequency in the corpus with its frequency in a reference corpus from a different domain; and measures such as 'domain pertinence' in (Sclano & Velardi 2007) and 'domain specificity' in (Kozakov *et al.* 2004; Park, Byrd, & Boguraev 2002), which extend and revise 'weirdness.' The trend in recent research is to use hybrid approaches, in which 'unithood' and 'termhood' are combined to produce an unified indicator, such as 'C-value'(Frantzi & Ananiadou 1999), and many others (Fahmi, Bouma, & van der Plas 2007; Kozakov *et al.* 2004; Park, Byrd, & Boguraev 2002; Sclano & Velardi 2007).

Despite the plethora of methods available, seldom is the full range of the problem dealt with by any one method. Firstly, most works rely on the simplifying assumption that the majority of terms consist of multi-word units (Caraballo & Charniak 1999; Deane 2005; Fahmi, Bouma, & van der Plas 2007; Wermter & Hahn 2005). However, while (Nakagawa & Mori 1998) claims that 85% of domain-specific terms are multi-word units, (Krauthammer & Nenadic 2004) claims that only a small percentage of gene names are multi-word units. Hence, for some domains such an assumption leads to very low recall, which, in turn, can hamper tasks built on top of ATR. Secondly, some approaches (Deane 2005; Frantzi & Ananiadou 1999; Wermter & Hahn 2005) apply frequency thresholds to reduce the algorithm's search space by filtering out low frequency candidate terms. This, however, does not take into account Zipf's law (that word frequencies follow highly skewed distributions and there are large number of rare events), again leading to reduced recall. Finally, experimental evaluations throughout the literature are disparate, differing in aspects such as corpus selection (e.g. domain, size), evaluation methodology (e.g. human judges, dictionary based, gold standard based), and scope (e.g. entire results, parts of results, top n best results). Consequently, the methods' performances are not directly comparable.

The purpose of the study presented in this paper is to select and implement a subset of the methods in the ATR literature and compare their performance under the same experimental setting. We then introduce a voting mechanism to combine the results from different methods into an integrated output to improve the result. The methods selected are exempt from the aforementioned limitations, namely the multi-word unit assumption and the frequency thresholding, which, from our experience in working on ontology learning from text, prevents the successful application of ATR to a wide range of domains and corpora. To our best knowledge, the methods presented by (Ahmad, Gillam, & Tostevin 1999; Frantzi & Ananiadou 1999; Kozakov *et al.* 2004; Park, Byrd, & Boguraev 2002; Sclano & Velardi 2007) are capable of recognising both single- and multi-word terms and do not apply frequency thresholds. Thus, we selected these methods, together with a corpus and an evaluation methodology so as to make it possible to directly compare them. The study is easily reproducible as all code and data were made publicly available[1].

The rest of the paper briefly overviews the methods, describes the experimental setting, and presents the results obtained, followed by a discussion of the results and conclusions.

## 2. Experimental Setting

### 2.1 Methods Selected

We implemented and compared five algorithms: TF-IDF (as a baseline), 'weirdness' (Ahmad, Gillam, & Tostevin 1999), 'C-value' (Frantzi & Ananiadou 1999), 'Glossex' (Kozakov *et al.* 2004; Park, Byrd, & Boguraev 2002) and 'TermExtractor' (Termex) (Sclano & Velardi 2007). These algorithms were selected because they were capable of recognising both single- and multi-word terms and did not throw away candidate terms on the basis of frequency only. TF-IDF and weirdness are 'termhood' only algorithms: the former makes use of term frequencies and document frequencies in the target corpus, while the latter compares term frequencies in both the target and a reference corpus. The remainder are hybrid approaches that combine 'termhood' and 'unit-hood' C-value measures 'termhood' by using term frequencies, and 'unithood' by examining frequencies of a term used as parts of longer terms. It has been widely used and tested in medical and biological domains and for multi-word term recognition. Glossex measures 'termhood' by comparing term frequency in the target corpus with its frequency in a reference corpus; then measures 'unithood' based on the overall term frequency normalised with respect to the frequencies of the component words. TermExtractor is similar to Glossex, the most notable difference being in its approach to 'termhood' where it additionally takes into account a novel measure called 'domain consensus' which captures domain concepts that exhibit high frequencies within small subset of the corpus (e.g., single document) but are completely absent in the remainder of the corpus. TermExtractor has been implemented as a web application and is publicly available[2].

### 2.2 The Voting Algorithm

Voting systems are often used as an improvement strategy in word sense disambiguation systems (Klein *et al.* 2002; Sinha & Mihalcea 2007; Su´arez & Palomar 2002), in which a voting algorithm is applied to the outputs from multiple classification algorithms to select the most appropriate word sense. In (Klein *et al.* 2002) two types of voting strategies are introduced: majority voting in which the sense output by most classifiers is selected; and weighted voting in which each classifier is assigned a weight, and the sense receiving the greatest total weighted vote is selected. In this paper we introduce a weighted voting strategy based on the rankings of a term produced by each term recognition algorithm. The new rank of a term t is measured by

$$rank = \sum_{i}^{k} \frac{1}{R(t_i)} w_i \qquad (1)$$

where $k$ is number of algorithms to be combined, $R(t_i)$ is the rank of term $t$ given by algorithm $i$, $w_i$ is the weight assigned to that algorithm. $w_i$ is measured by

$$w_i = \frac{P_i}{\sum_i^k P_i} \qquad (2)$$

where $P_i$ is the precision of algorithm $i$. Our study aims to find out whether such a voting algorithm improves the outputs from any individual term recognition algorithm.

### 2.3 Dataset and Pre-processing

Many ATR approaches are evaluated in medical or biological domains. We find that evaluation in other kinds of domains, notably less technical ones, have been lacking. An important research question that arises is whether the relative success of some methods is due to the fact that terms

[1] Available at http://nlp.shef.ac.uk/abraxas/termrecog/

[2] http://lcl2.uniromal.it/termextractor/demo.jsp

| Total number of terms evaluated = 100 | | | | | | |
|---|---|---|---|---|---|---|
| **Judge** | **TF-IDF** | **Weirdness** | **C-value** | **Glossex** | **Termex** | **Voted** |
| *1* | 0.67 | 0.8 | 0.59 | 0.81 | 0.93 | 0.97 |
| *2* | 0.79 | 0.85 | 0.69 | 0.83 | 0.95 | 0.97 |
| *3* | 0.77 | 0.77 | 0.68 | 0.83 | 0.95 | 0.97 |
| **Total number of terms evaluated = 300** | | | | | | |
| **Judge** | **TF-IDF** | **Weirdness** | **C-value** | **Glossex** | **Termex** | **Voted** |
| *1* | 0.57 | 0.81 | 0.55 | 0.86 | 0.9 | 0.95 |
| *2* | 0.59 | 0.8 | 0.56 | 0.88 | 0.86 | 0.96 |
| *3* | 0.64 | 0.77 | 0.61 | 0.88 | 0.9 | 0.97 |

Table 1: Precision of each algorithm evaluated by each judge - Wikipedia corpus

in those domains are more clearly characteristic and specific to that domain. For that reason, we chose two domains for our experiment, i.e., biological domain and a domain ‘closer’ to common knowledge in order to compare the methods.

We firstly selected the GENIA corpus[3] (Kim *et al.* 2003) which contains 2000 abstracts totaling 420,000 words selected from National Library of Medicine’s MEDLINE database for the biological domain; for the other domain we manually created a corpus of roughly 1.3 million words consisting of Wikipedia articles describing 1,052 different animals. This was created by extracting only the main textual content from the HTML pages and ignoring any formatting or navigational elements. These corpora were then POS tagged and the linguistic filters described in (Frantzi & Ananiadou 1999) were applied to extract nouns and noun phrases as candidate terms. The candidate list is then filtered by removing stop words. However we do not apply a frequency filter but rather submit the full list of candidates to each algorithm for ranking.

### 2.4 Evaluation Methodology

The way methods are evaluated in the ATR literature is diverse, ranging from human judges manually assessing a selected section of the output (Frantzi & Ananiadou 1999; Park, Byrd, & Boguraev 2002; Sclano & Velardi 2007), to unsupervised matching of parts of the output against dictionary resources (Deane 2005; Fahmi, Bouma, & van der Plas 2007; Wermter & Hahn 2005). Nevertheless, evaluations have two things in common: first, the majority only measure precision but not recall; second, they evaluate only a subset of the output. Accordingly, our experiments compared precision (and other precision-related metrics) of the top terms returned by the algorithms. For the Wikipedia corpus these are calculated from the judgment of native language speakers. We instructed three judges to manually inspect the top 300 candidate terms produced by each algorithm and mark those they believed to be terms one would expect to encounter when reading texts about animals; for GENIA corpus we extracted the annotations from the corpus as the gold standard terminologies for evaluation. Next we applied two evaluation metrics: ‘classic’ precision measuring the proportion of correct terms to all terms considered; and the Un-interpolated Average Precision (UAP) (Schone & Jurafsky 2001) which averages precision at the $i^{\text{th}}$ correct term out of the total K correct terms in the ranked output considered (cf. Eq. 3).

$$\frac{1}{k}\sum_{i=1}^{k} P_i \qquad (3)$$

In Eq. 3, $K$ is the set of $K$ correct terms in the output, and $P_i = i/H_i$ is precision at rank $i$, where $H_i$ is the number of hypothesized terms required to find the $i^{th}$ correct term.

## 3. Results

Table 1 and 2 illustrate the precision of each algorithm on each corpus. Table 3 illustrates precision obtained from the experiment on the Wikipedia corpus depending on how strictly inter-annotator agreement was required. We calculated precision of each algorithm under both a strict mode, in which a candidate is counted a correct term only if all judges agreed, and a lenient mode, in which any candidates marked as correct terms by any judges are counted.

Figure 1 and Figure 2 show the UAP results for each corpus.

| **Total number of terms evaluated = N** | | | | | | |
|---|---|---|---|---|---|---|
| **N** | **TF-IDF** | **Weird-ness** | **C-value** | **Glossex** | **Termex** | |
| *100* | 0.9 | 0.48 | 0.91 | 1 | 0.92 | 0.97 |
| *1k* | 0.82 | 0.55 | 0.91 | 0.82 | 0.75 | 0.86 |
| *5k* | 0.8 | 0.58 | 0.83 | 0.69 | 0.62 | 0.81 |
| *10k* | 0.75 | 0.58 | 0.68 | 0.66 | 0.61 | 0.73 |
| *20k* | 0.6 | 0.56 | 0.58 | 0.59 | 0.55 | 0.62 |

Table 2: Precision of each algorithm – Genia corpus

[3] Available at http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

| | *Strict* | *Lenient* |
|---|---|---|
| **TF-IDF** | 0.52 | 0.66 |
| **Weirdness** | 0.64 | 0.93 |
| **C-Value** | 0.49 | 0.66 |
| **Glossex** | 0.82 | 0.93 |
| **Termex** | 0.83 | 0.93 |
| **Voted** | 0.93 | 0.98 |

Table 3: Precision with respect to inter-annotator agreement Wikipedia corpus, top 300 terms

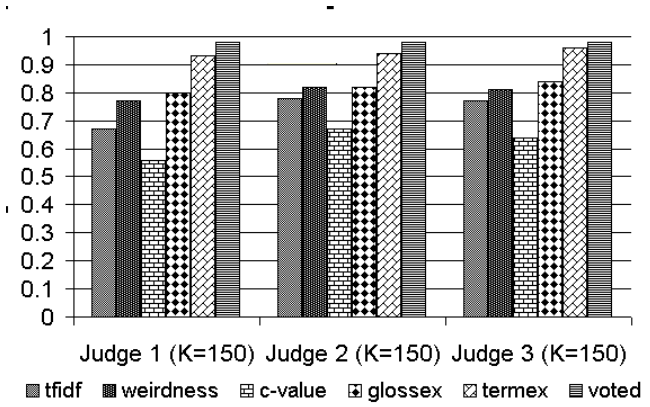

Figure 1: UAP metric for the Wikipedia corpus

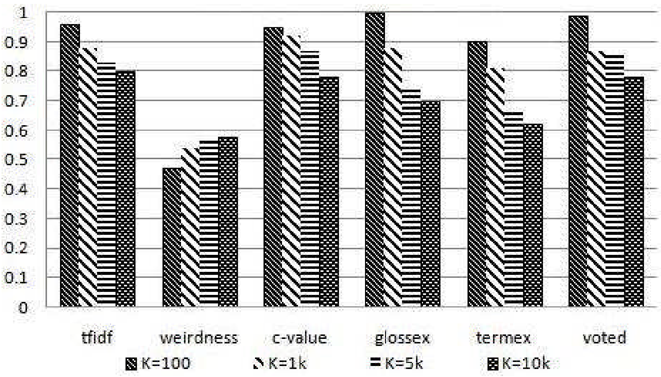

Figure 2: UAP metric for the Genia corpus

## 4. Discussion

In terms of overall precision, our experiments do not produce consistent results from two testing corpus, indicating that domains and quality of corpus do have an impact on the performance of ATR algorithms. Experiment on the Wikipedia corpus shows that Termex performs best among the five algorithms; however the algorithm does not perform so well on the GENIA corpus. Likewise, the C-value algorithm, which has been tested extensively and proved to be effective in biological and medical domains, outperforms the rest on the GENIA corpus but fails on the Wikipedia corpus. This is also due to its specificity to multi-word term recognition, although it is an algorithm applicable to single-word units. Manual inspection showed that for the Wikipedia corpus if considering only the multi-word units (46 out of 100) from the C-value output, the precision is 0.67 under strict mode, and 0.82 under lenient mode.

Using the UAP metric, Termex again outperforms the rest on the Wikipedia corpus, whereas on the GENIA corpus C-value is again the winning algorithm. A high value of UAP indicates that the algorithm produces an even distribution of precision at every correct term on the list, and that it manages to bring correct terms to the top.

Surprisingly we noticed that TF-IDF produces the second best output from the GENIA corpus. We believe this is due to the high quality of the corpus, as manual inspection shows that the 35% of noun phrases extracted from the corpus are annotated as correct terms.

Experiments also show that using a weighted voting system does not necessarily improve performance of an ATR algorithm. Test on the GENIA corpus indicates that the performance of the voting system does not outperform C-value. However, this may be again caused by the quality of corpus as well as of C-value's specificity to multi-word terms.

We also manually counted the proportions of single-word terms in the outputs from two testing corpus. For the GENIA corpus, only 11% of gold standard terms are single-word terms. For the Wikipedia corpus, the numbers of single-word terms in the top 100 candidates from each algorithm are 100, 59, 54, 78 and 99 for TFIDF, Weirdness, C-value, Glossex and Termex respectively. Clearly, ignoring single-word terms will reduce system recall significantly, while algorithms favouring multi-word terms also suffer from low precision when single-word terms are counted (C-value).

Careful examination of the Wikipedia corpus and the selected terms showed that a number of errors were due to fact that the original pages had a certain structure (for example details of the classification of a species in a table on the right hand side of the web page). The elimination of this structure meant that structured data was lost, and some errors of collocation were created largely due to errors of the subsequent POS tagging and noun phrase chunking.

## 5. Conclusion

In this paper we have compared the performance of five ATR algorithms capable of recognising both single- and multi-word terms on a corpus of animal domain and a corpus of biological science. These include ‘termhood’ only approaches such as TFIDF and weirdness, and hybrid approaches that combine ‘termhood and unit-hood’ such as C-value, Glossex and Termex; among which Termex demonstrated best performance on the Wikipedia corpus and C-value outperforms other algorithms on the Genia corpus. In general we showed that hybrid methods work better than ‘termhood’ only methods.

Our experiments showed that single-word terms can be equally important and occupy a fairly large proportion in certain domains. As a result, algorithms that ignore single-word terms may cause problems to tasks built on top of ATR.

More generally effective ATR systems need to take into account both the unstructured text and the structured aspects such as layout, italics and other textual features, where accessible. The Termex algorithm claims to do so but does not provide details in the relevant publications, so we were not able to replicate this. Many of the items identified as terms fall into the category of items that information extraction has traditionally extracted from text, and thus our future work will be to integrate IE techniques into the ATR process including machine learning and Active Learning with user feedback.

## 6. Acknowledgement

This research was supported by ‘Abraxas’ ontology learning from text project, which was funded by EPSRC.

## 7. References

Ahmad, K.; Gillam, L.; and Tostevin, L. 1999. University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder). In *The Eighth Text REtrieval Conference (TREC-8).*

Ananiadou, S. 1994. A methodology for automatic term recognition. In *COLING 15th International Conference on Computational Linguistics,* 1034–1038.

Bhagdev, R.; Butters, J.; Chakravarthy, A.; Chapman, S.; Dadzie, A.-S.; Greenwood, M. A.; Iria, J.; and Ciravegna, F. 2007. Doris: Managing document-based

6(3):145–180.

Kim, J.-D.; Ohta, T.; Tateisi, Y.; and Tsujii, J. 2003. Genia corpus–semantically annotated corpus for bio-textmining. *Bioinformatics* 19 Suppl 1 :i180–i182.

Klein, D.; Toutanova, K.; Ilhan, H. T.; Kamvar, S. D.; and Manning, C. D. 2002. Combining heterogeneous classifiers for word-sense disambiguation. In *Word Sense Disambiguation Workshop: Recent Successes and Future Directions (held in conjunction with ACL conference).*

Kozakov, L.; Park, Y.; Fin, T.-H.; Drissi, Y.; Doganata, Y. N.; and Cofino, T. 2004. Glossary extraction and

knowledge in large organisations via semantic web technologies. In *Proceedings of the 6th International Semantic Web Conference and the 2nd Asian Semantic Web Conference (Se-mantic Web Challenge Track).*

Bourigault, D. 1992. Surface grammatical analysis for the extraction of terminological noun phrases. In *14th International Conference on Computational Linguistics - COLING 92,* 977–98 1.

Brewster, C.; Iria, J.; Zhang, Z.; Ciravegna, F.; Guthrie, L.; and Wilks, Y. 2007. Dynamic iterative ontology learning. In *Recent Advances in Natural Language Processing (RANLP 07).*

Caraballo, S. A., and Charniak, E. 1999. Determining the specificity of nouns from text. In *Proceedings of the 1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora.*

Cohen, J. D. 1995. Highlights: Language- and domain-independent automatic indexing terms for abstracting. *Journal of the American Society for Information Science* 46(3): 162–174.

Daille, B. 1996. Study and implementation of combined techniques for automatic extraction of terminology. In Klavans, J., and Resnik, P., eds., *The Balancing Act: Combining Symbolic and Statistical Approaches to Language.* Cambridge, Massachusetts: The MIT Press. 49–66.

Deane, P. 2005. A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the Conference 43rd Annual Meeting of the Association for Computational Linguistics.* University of Michigan, USA: The Association for Computer Linguistics.

Evans, D. A., and Lefferts, R. G. 1995. Clarit-trec experiments. *Information Processing and Management* 3 1(3):385–395.

Fahmi, I.; Bouma, G.; and van der Plas, L. 2007. Improving statistical method using known terms for automatic term extraction. In *Computational Linguistics in the Netherlands - CLIN 17.*

Frantzi, K. T., and Ananiadou, S. 1999. The c/nc value domain independent method for multi-word term extraction. *Journal of Natural Language Processing*

utilization in the information search and delivery system for ibm technical support. *IBM Systems Journal* 43(3):546–563.

Krauthammer, M., and Nenadic, G. 2004. Term identification in the biomedical literature. *J Biomed Inform* 37(6):512–526.

Medelyan, O., and Witten, I. H. 2006. Thesaurus based automatic keyphrase indexing. In Marchionini, G.; Nelson, M. L.; and Marshall, C. C., eds., *Porceedings of the ACM/IEEE Joint Conference on Digital Libraries, JCDL 2006,*, 296–297. Chapel Hill, NC, USA: ACM.

Nakagawa, H., and Mori, T. 1998. Nested collocation and compound noun for term extraction. In *Proceedings of the First Workshop on Comutational Terminology(COMPUTERM'98),* 64–70.

Park, Y.; Byrd, R. J.; and Boguraev, B. 2002. Automatic glossary extraction: Beyond terminology identification. In *19th International Conference on Computational Linguistics - COLING 02.* Taipei, Taiwan: Howard International House and Academia Sinica.

Park, Y.; Byrd, R. J.; and Boguraev, B. K. 2003. Towards ontologies on demand. In Ashish, N., and Goble, C., eds., *Proceedings of the Workshop on Semantic Web Technologies for Searching and Retrieving Scientific Data. Colocated with the Second International Semantic Web Conference (ISWC-03).*

Schone, P., and Jurafsky, D. 2001. Is knowledge-free induction of multiword unit dictionary headwords a solved problem? In *Proceedings of Empirical Methods in Natural Language Processing.*

Sclano, F., and Velardi, P. 2007. Termextractor: a web application to learn the shared terminology of emergent web communities. In *Proceedings of the 3rd International Conference on Interoperability for Enterprise Software andApplications (I-ESA 2007).*

Sinha, R., and Mihalcea, R. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC '07: Proceedings of the Inter-national Conference on Semantic Computing (ICSC 2007),* 363–369. Washington, DC, USA: IEEE Computer Society.

Su´arez, A., and Palomar, M. 2002. A maximum entropy-based word sense disambiguation system. In *Proceedings of the 19th international conference on Computational linguistics (COLING 02),* 1–7. Taipei, Taiwan: Association for Computational Linguistics.

Wermter, J., and Hahn, U. 2005. Finding new terminology in very large corpora. In Clark, P., and Schreiber, G., eds., *Proceedings of the 3rd International Conference on Knowledge Capture (K-CAP 2005),* 137–144. Banff, Alberta, Canada: ACM.