Acquiring Sense Tagged Examples using Relevance Feedback

Mark Stevenson, Yikun Guo and **Robert Gaizauskas**
Department of Computer Science
University of Sheffield
Regent Court, 211 Portobello
Sheffield, S1 4DP
United Kingdom
inital.surname@dcs.shef.ac.uk

Abstract

Supervised approaches to Word Sense Disambiguation (WSD) have been shown to outperform other approaches but are hampered by reliance on labeled training examples (the *data acquisition bottleneck*). This paper presents a novel approach to the automatic acquisition of labeled examples for WSD which makes use of the Information Retrieval technique of relevance feedback. This semi-supervised method generates additional labeled examples based on existing annotated data. Our approach is applied to a set of ambiguous terms from biomedical journal articles and found to significantly improve the performance of a state-of-the-art WSD system.

1 Introduction

The resolution of lexical ambiguities has long been considered an important part of the process of understanding natural language. Supervised approaches to Word Sense Disambiguation (WSD) have been shown to perform better than unsupervised ones (Agirre and Edmonds, 2007) but require examples of ambiguous words used in context annotated with the appropriate sense (labeled examples). However these often prove difficult to obtain since manual sense annotation of text is a complex and time consuming process. In fact, Ng (1997) estimated that 16 person years of manual effort would be required to create enough labeled examples to train a wide-coverage WSD system. This limitation is commonly referred to as the *data acquisition bottleneck*. It is particularly acute in specific domains, such as biomedicine, where terms may have technical usages which only domain experts are likely to be aware of. For example, possible meanings of the term "ganglion" in UMLS (Humphreys et al., 1998) include 'neural structure' or 'benign mucinous tumour', although only the first meaning is listed in WordNet. These domain-specific semantic distinctions make manual sense annotation all the more difficult.

One approach to the data acquisition bottleneck is to generate labeled training examples automatically. Others, such as Leacock et al. (1998) and Agirre and Martínez (2004b), used information from WordNet to construct queries which were used to retrieve training examples. This paper presents a novel approach to this problem. Relevance feedback, a technique used in Information Retrieval (IR) to improve search results, is adapted to identify further examples for each sense of ambiguous terms. These examples are then used to train a semi-supervised WSD system either by combining them with existing annotated data or using them alone. The approach is applied to a set of ambiguous terms in biomedical texts, a domain for which existing resources containing labeled examples, such as the NLM-WSD data set (Weeber et al., 2001), are limited.

The next section outlines previous techniques which have been used to avoid the data acquisition bottleneck. Section 3 describes our approach based on relevance feedback. The WSD system we use is described in Section 4. Section 5 describes experiments carried out to determine the usefulness of the automatically retrieved examples. The final section summarises conclusions which can be drawn from this work and outlines future work.

2 Previous Approaches

A variety of approaches to the data acquisition bottleneck have been proposed. One is to use unsupervised algorithms, which do not require labeled training data. Examples include Lesk (1986) who disambiguated ambiguous words by examining their dictionary definitions and selecting the sense whose definition overlapped most with definitions of words in the ambiguous word's context. Leroy and Rindflesch (2005) presented an unsupervised approach to WSD in the biomedical domain using information derived from UMLS (Humphreys et al., 1998).

However, results from SemEval (Agirre et al., 2007) and its predecessors have shown that supervised approaches to WSD generally outperform unsupervised ones. It has also been shown that results obtained from supervised methods improve with access to additional labeled data for training (Ng, 1997). Consequently various techniques for automatically generating training data have been developed.

One approach makes use of the fact that different senses of ambiguous words often have different translations (e.g. Ng et al. (2003)). Parallel text is used as training data with the alternative translations serving as sense labels. However, disadvantages of this approach are that the alternative translations do not always correspond to the sense distinctions in the original language and parallel text is not always available.

Another approach, developed by Leacock et al. (1998) and extended by Agirre and Martínez (2004b), is to examine a lexical resource, WordNet in both cases, to identify unambiguous terms which are closely related to each of the senses of an ambiguous term. These "monosemous relatives" are used to as query terms for a search engine and the examples returned used as additional training data.

In the biomedical domain, Humphrey et al. (2006) use journal descriptors to train models based on the terms which are likely to co-occur with each sense. Liu et al. (2002) used information in UMLS to disambiguate automatically retrieved examples which were then used as labeled training data. The meanings of 35 ambiguous abbreviations were identified by examining the closeness of concepts in the same abstract in UMLS. Widdows et al. (2003) employ a similar approach, although their method also makes use of parallel corpora when available.

All of these approaches rely on the existence of an external resource (e.g. parallel text or a domain ontology). In this paper we present a novel approach, inspired by the relevance feedback technique used in IR, which automatically identifes additional training examples using existing labeled data.

3 Generating Examples using Relevance Feedback

The aim of relevance feedback is to generate improved search queries based on manual analysis of a set of retrieved documents which has been shown to improve search precision (Salton, 1971; Robertson and Spark Jones, 1976). Variations of relevance feedback have been developed for a range of IR models including Vector Space and probabilistic models. The formulation of relevance feedback for the Vector Space Model is most pertinent to our approach.

Given a collection of documents, C, containing a set of terms, C_{terms}, a basic premise of the Vector Space Model is that documents and queries can be represented by vectors whose dimensions represent the C_{terms}. Relevance feedback assumes that a retrieval system returns a set of documents, D, for some query, q. It is also assumed that a user has examined D and identified some of the documents as relevant to q and others as not relevant. Relevant documents are denoted by D_{+q} and the irrelevant as D_{-q}, where $D_{+q} \subseteq D$, $D_{-q} \subseteq D$ and $D_{+q} \cap D_{-q} = \emptyset$. This information is used to create a modified query, q_m, which should be more accurate than q. A standard approach to constructing q_m was described by Rocchio (1971):

$$q_m = \alpha q + \frac{\beta}{|D_{+q}|} \sum_{\forall d \in D_{+q}} d - \frac{\gamma}{|D_{-q}|} \sum_{\forall d \in D_{-q}} d \quad (1)$$

where the parameters α, β and γ are set for particular applications. Rocchio (1971) set α to 1.

Our scenario is similar to the relevance feedback problem since the sense tagged examples provide information about the documents in which a particular meaning of an ambiguous term is likely to be found. By identifying the features which distinguish the documents containing one sense from the others we can create queries which can then be used to retrieve further examples of the ambiguous words used in the same sense. However, unlike

$$score(t,s) = idf(t) \times \left(\frac{\alpha}{|D_{+s}|} \sum_{\forall d \in D_{+s}} count(t,d) - \frac{\beta}{|D_{-s}|} \sum_{\forall d \in D_{-s}} count(t,d) \right) \tag{2}$$

the relevance feedback scenario there is no original query to modify. Consequently we start with a query containing just the ambiguous term and use relevance feedback to generate queries which aim to retrieve documents where that term is being used in a particular sense.

The remainder of this section describes how this approach is applied in more detail.

3.1 Corpus Analysis

The first stage of our process is to analyse the labeled examples and identify good search terms. For each sense of an ambiguous term, s, the labeled examples are divided into two sets: those annotated with the sense in question and the remainder (annotated with another sense). In relevance feedback terminology the documents annotated with the sense in question are considered to be relevant and the remainder irrelevant. These examples are denoted by D_{+s} and D_{-s} respectively.

At its core relevance feedback, as outlined above, aims to discover how accurately each term in the collection discriminates between relevant and irrelevant documents. This approach was used to inspire a technique for identifying terms which are likely to indicate the sense in which an ambiguous word is being used. We compute a single score for each term, reflecting its indicativeness of that sense, using the formula in equation 2, where $count(t,d)$ is the number of times term t occurs in document d and $idf(t)$ is the inverse document frequency term weighting function commonly used in IR. We compute idf as follows:

$$idf(t) = log \frac{|C|}{df(t)} \tag{3}$$

where D is the set of all annotated examples (i.e. $D = D_{+s} \cap D_{-s}$) and $df(t)$ the number of documents in C which contain t.[1]

In our experiments the α and β parameters in equation 2 are set to 1. Documents are lemmatised and stopwords removed before computing relevance scores.

Table 1 shows the ten terms with the highest relevance score for two senses of the term "culture" in UMLS: 'laboratory culture' ("In peripheral blood mononuclear cell *culture* streptococcal erythrogenic toxins are able to stimulate tryptophan degradation in humans") and 'anthropological culture' ("The aim of this paper is to describe the origins, initial steps and strategy, current progress and main accomplishments of introducing a quality management *culture* within the healthcare system in Poland.").

'anthropological culture'		'laboratory culture'	
cultural	26.17	suggest	6.32
recommendation	14.82	protein	6.13
force	14.80	presence	5.86
ethnic	14.79	demonstrate	5.86
practice	14.76	analysis	5.78
man	14.76	gene	5.58
problem	13.04	compare	5.47
assessment	12.94	level	5.36
experience	11.60	response	5.35
consider	11.58	data	5.35

Table 1: Relevant terms for two senses of 'culture'

3.2 Query Generation

Unlike the traditional formulation of relevance feedback there is no initial query. To create a query designed to retrieve examples of each sense we simply combine the ambiguous term and the n terms with the highest relevance scores. We found that using the three highest ranked terms provided good results. So, for example, the queries generated for the two senses of culture shown in Table 1 would be "`culture cultural recommendation force`" and "`culture suggest protein presence`".

3.3 Example Collection

The next stage is to collect a set of examples using the generated queries. We use the Entrez retrieval system (`http://www.ncbi.nlm.nih.gov/sites/gquery`) which provides an online interface for carrying out boolean queries over the PubMed database of biomedical journal abstracts.

Agirre and Martínez (2004b) showed that it is important to preserve the bias of the original corpus when automatically retrieving examples and

[1] Our computation of $idf(t)$ is based on only information from the labeled examples, i.e. we assume $C = D_{+s} \cap D_{-s}$. Alternatively idf could be computed over a larger corpus of labeled and unlabeled examples.

consequently the number retrieved for each sense is kept in proportion to the original corpus. For example, if our existing labeled examples contain 75 usages of 'culture' in the 'laboratoy culture' sense and 25 meaning 'anthropological culture' we would ensure that 75% of the examples returned would refer to the first sense and 25% to the second.

Unsurprisingly, we found that the most useful abstracts for a particular sense are the ones which contain more of the relevant terms identified using the process in Section 3.1. However, if too many terms are included Entrez may not return any abstracts. To ensure that a sufficient number of abstracts are returned we implemented a process of query relaxation which begins by querying Entrez with the most specific query for set of terms. If that query matches enough abstracts these are retrieved and the search for labeled examples for the relevant sense considered complete. However, if that query does not match enough abstracts it is relaxed and Entrez queried again. This process is repeated until enough examples can be retrieved for a particular sense.

The process of relaxing queries is carried out as follows. Assume we have an ambiguous term, a, and a set of terms T identified using the process in Section 3.1. The first, most specific query, is formed from the conjunction of all terms in $a \cup T$, i.e. "$a\ and\ t_1\ AND\ t_2\ AND\ ...\ t_{|T|}$". This is referred to as the level $|T|$ query. If this query does not return enough abstracts the more relaxed level $|T| - 1$ query is formed. This query returns documents which include the ambiguous word and all but one of the terms in T: "$a\ AND\ ((t_1\ AND\ t_2\ AND\ ...\ AND\ t_{n-1})\ OR\ (t_1\ AND\ t_2\ AND\ ...\ t_{n-2}\ AND\ t_n)\ OR\ ...\ OR\ (t_2\ AND\ t_3\ ...\ AND\ t_n\ ))$". Similarly, level $|T| - 2$ queries return documents containing the ambiguous term and all but two of the terms in T. Level 1 queries, the most relaxed, return documents containing the ambiguous term and one of the terms in T. We do not use just the ambiguous term as the query since this does not contain any information which could discriminate between the possible meanings. Figure 1 shows the queries which are formed for the ambiguous term "culture" and the three most salient terms identified for the 'anthropological culture' sense. The "matches" column lists the number of PubMed abstracts the query matches. It can be seen that there are no matches for the level 3 query and 83 for the more relaxed level 2 query. For this sense, abstracts returned by the level 2 query would be used if 83 or fewer examples were required, otherwise abstracts returned by the level 1 query would be used.

Note that the queries submitted to Entrez are restricted so the terms only match against the title and abstract of the PubMed articles. This avoids spurious matches against other parts of the records including metadata and authors' names.

4 WSD System

The basis of our WSD system was developed by Agirre and Martínez (2004a) and participated in the Senseval-3 challenge (Mihalcea et al., 2004) with a performance which was close to the best system for the English and Basque lexical sample tasks. The system has been adapted to the biomedical domain (Stevenson et al., 2008) and has the best reported results over the NLM-WSD corpus (Weeber et al., 2001), a standard data set for evaluation of WSD algorithms in this domain.

The system uses a wide range of features which are commonly employed for WSD:

Local collocations: A total of 41 features which extensively describe the context of the ambiguous word and fall into two main types: (1) bigrams and trigrams containing the ambiguous word constructed from lemmas, word forms or PoS tags, and (2) preceding/following lemma/word-form of the content words (adjective, adverb, noun and verb) in the same sentence with the target word.

Syntactic Dependencies: This feature models longer-distance dependencies of the ambiguous words than can be represented by the local collocations. Five relations are extracted: object, subject, noun-modifier, preposition and sibling. These are identified using heuristic patterns and regular expressions applied to PoS tag sequences around the ambiguous word (Agirre and Martínez, 2004a).

Salient bigrams: Salient bigrams within the abstract with high log-likelihood scores, as described by Pedersen (2001).

Unigrams: Lemmas of all content words (nouns, verbs, adjectives, adverbs) in the target word's sentence and, as a separate feature, lemmas of all content words within a 4-word window around the target word, excluding those in a list of corpus-specific stopwords (e.g. "ABSTRACT", "CONCLUSION"). In addition, the lemmas of any

Level	Matches	Query
3	0	`culture AND (cultural AND recommendation AND force)`
2	83	`culture AND ((cultural AND recommendation) OR (cultural AND force) OR (recommendation AND force))`
1	6,358	`culture AND (cultural OR recommendation OR force)`

Figure 1: Examples of various query levels

unigrams which appear at least twice in the entire corpus which are found in the abstract are also included as features. This feature was not used by Agirre and Martínez (2004a), but Joshi et al. (2005) found them to be useful for this task.

Features are combined using the **Vector Space Model**, a memory-based learning algorithm (see Agirre and Martínez (2004a)). Each occurrence of an ambiguous word is represented as a binary vector in which each position indicates the occurrence/absence of a feature. A single centroid vector is generated for each sense during training. These centroids are compared with the vectors that represent new examples using the cosine metric to compute similarity. The sense assigned to a new example is that of the closest centroid.

5 Experiments

5.1 Setup

The NLM-WSD corpus Weeber et al. (2001) was used for evaluation. It contains 100 examples of 50 ambiguous terms which occur frequently in MEDLINE. Each example consists of the abstract from a biomedical journal article which contains an instance of the ambiguous terms which has been manually annotated with a UMLS concept.

The 50 ambiguous terms which form the NLM-WSD data set represent a range of challenges for WSD systems. Various researchers (Liu et al., 2004; Leroy and Rindflesch, 2005; Joshi et al., 2005; McInnes et al., 2007) chose to exclude some of the terms (generally those with highly skewed sense distributions or low inter-annotator agreement) and evaluated their systems against a subset of the terms. The number of terms in these subsets range between 9 and 28. The Most Frequent Sense (MFS) heuristic has become a standard baseline in WSD (McCarthy et al., 2004) and is simply the accuracy which would be obtained by assigning the most common meaning of a term to all of its instances in a corpus. The MFS for the whole NLM-WSD corpus is 78% and ranges between 69.9% and 54.9% for the various subsets. We report results across the NLM-WSD corpus and four subsets from the literature for completeness.

The approach described in Section 3 was applied to the NLM-WSD data set. 10-fold cross validation is used for all experiments. Consequently 10 instances of each ambiguous term were held back for testing during each fold and additional examples generated by examining the 90 remaining instances. Three sets of labeled examples were generated for each fold, containing 90, 180 and 270 examples for each ambiguous term. The NLM-WSD corpus represents the only reliably labeled data to which we have access and is used to evaluate all approaches (that is, systems trained on combinations of the NLM-WSD corpus and/or the automatically generated examples).

5.2 Results

Various WSD systems were created. The "basic" system was trained using only the NLM-WSD data set and was used as a benchmark. Three systems, "+90", "+180" and "+270" were trained using the combination of the NLM-WSD data set and, respectively, the 90, 180 and 270 automatically retrieved examples for each term. A further three systems, "90", "180" and "270" were trained using only the automatically retrieved examples.

The performance of our system is shown in Table 2. The part of the table labeled "Subsets properties" lists the number of terms in each subset of the NLM-WSD corpus and the relevant MFS baseline.

Adding the first 90 automatically retrieved examples ("+90" column) significantly improves performance of our system from 87.2%, over all words, to 88.5% (Wilcoxon Signed Ranks Test, $p < 0.01$). Improvements are observed over all subsets of the NLM-WSD corpus. Although the improvements may seem modest they should be understood in the context of the WSD system we are using which has exceeded previously reported performance figures and therefore represents a high baseline.

Table 2 also shows that adding more automatically retrieved examples ("+180" and "+270" columns) causes a drop in performance and re-

	Subset Properties		basic	Combined			New only		
Subset	Terms	MFS		+90	+180	+270	90	180	270
All words	50	78.0	87.2	**88.5**	87.0	86.1	85.6	84.5	82.7
Joshi et. al.	28	66.9	82.3	**83.8**	81.6	80.9	79.8	78.0	76.3
Liu et. al.	22	69.9	77.8	**79.6**	76.9	76.1	74.9	72.0	70.9
Leroy	15	55.3	84.3	**85.9**	84.4	83.6	81.2	80.0	78.0
McInnes et. al.	9	54.9	79.6	**81.8**	80.4	79.4	75.2	73.0	71.4

Table 2: Performance of system using a variety of combinations of training examples

sults using these examples are worse than using the NLM-WSD corpus alone. The query relaxation process, outlined in Section 3.3, uses less discriminating queries when more examples are required and it is likely that this is leading to noise in the training examples.

The rightmost portion of Table 2 shows performance when the system is trained using only the automatically generated examples which is consistently worse than using the NLM-WSD corpus alone. Performance also decreases as more examples are added. However, results obtained using only the automatically generated training examples are consistently better than the relevant baseline.

Table 3 shows the performance of the system trained on the NLM-WSD data set compared against training using only the 90 automatically generated examples for each ambiguous term in the NLM-WSD corpus. It can be seen that there is a wide variation between the performance of the additional examples compared with the original corpus. For 11 terms training using the additional examples alone is more effective than using the NLM-WSD corpus. However, there are several words for which the performance using the automatically acquired examples is considerably worse than using the NLM-WSD corpus.

Information about the performance of a system trained using only the 90 automatically acquired examples can be used to boost WSD performance further. In this scenario, which we refer to as *example filtering*, the system has a choice whether to make use of the additional training data or not. For each word, performance of the WSD system trained using only the 90 automatically acquired examples is compared against the one trained on the NLM-WSD data set (i.e. results shown in Table 3). If the performance is as good, or better, then the additional examples are used, otherwise only examples in the NLM-WSD corpus are used as training data. Since the annotated examples in the NLM-WSD corpus have already been examined to generate the additional examples, example filtering does not require any more labeled data.

Results obtained when example filtering is used are shown in Table 4. The columns "+90(f)", "+180(f)" and "+270(f)" show performance when the relevant set of examples is filtered. (Note that all three sets of examples are filtered against the performance of the first 90 examples, i.e. results shown in Table 3.) This table shows that example filtering improves performance when the WSD system is trained using the automatically retrieved examples. Performance using the first 90 filtered examples ("+90(f)" column) is 89%, over all words, compared with 88.5% when filtering is not used. While performance decreases as larger sets of examples are used, results using each of the three sets of filtered examples is signifcantly better than the basic system (Wilcoxon Signed Ranks Test, $p < 0.01$ for "+90(f)" and "+180(f)", $p < 0.05$ for "+270(f)").

6 Conclusion and Future Work

This paper has presented a novel approach to the data acquisition bottleneck for WSD. Our technique is inspired by the relevance feedback technique from IR. This is a semi-supervised approach which generates labeled examples using available sense annotated data and, unlike previously published approaches, does not rely on external resources such as parallel text or an ontology. Evaluation was carried out on a WSD task from the biomedical domain for which the number of labeled examples available for each ambiguous term is limited. The automatically acquired examples improve the performance of a WSD system which has already been shown to exceed previously published results.

The approach presented in this paper could be extended in several ways. Our experiments focus

	basic	+90(f)	+180(f)	+270(f)
All words	87.2	89.0	88.2	87.9
Joshi et. al.	82.3	84.6	83.5	83.3
Liu et. al.	84.3	86.6	85.7	85.5
Leroy	77.8	80.3	79.1	78.5
McInnes et. al.	79.6	82.4	81.6	80.8

Table 4: Performance using example filtering

on the biomedical domain. The relevance feedback approach could be applied to other lexical ambiguities found in biomedical texts, such as abbreviations with multiple expansions (e.g. Liu et al. (2002)), or to WSD of general text, possibly using the SemEval data for evaluation.

Future work will explore alternative methods for generating query terms including other types of relevance feedback and lexical association measures (e.g. Chi-squared and mutual information). Experiments described here rely on a boolean IR engine (Entrez). It is possible that an IR system which takes term weights into account could lead to the retrieval of more useful MEDLINE abstracts. Finally, it would be interesting to explore the relation between query relaxation and the usefulness of the retrieved abstracts.

Acknowledgments

The authors are grateful to David Martinez for the use of his WSD system for these experiments and to feedback provided by three anonymous reviewers. This work was funded by the UK Engineering and Physical Sciences Research Council, grant number EP/E004350/1.

References

E. Agirre and P. Edmonds, editors. 2007. *Word Sense Disambiguation: Algorithms and Applications*. Text, Speech and Language Technology. Springer.

E. Agirre and D. Martínez. 2004a. The Basque Country University system: English and Basque tasks. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 44–48, Barcelona, Spain, July.

E. Agirre and D. Martínez. 2004b. Unsupervised WSD based on automatically retrieved examples: The importance of bias. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*, Barcelona, Spain.

E. Agirre, L. Marquez, and R. Wicentowski, editors. 2007. *SemEval 2007: Proceedings of the 4th International Workshop on Semantic Evaluations*, Prague, Czech Republic.

S. Humphrey, W. Rogers, H. Kilicoglu, D. Demner-Fushman, and T. Rindflesch. 2006. Word Sense Disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(5):96–113.

L. Humphreys, D. Lindberg, H. Schoolman, and G. Barnett. 1998. The Unified Medical Language System: An Informatics Research Collaboration. *Journal of the American Medical Informatics Association*, 1(5):1–11.

M. Joshi, T. Pedersen, and R. Maclin. 2005. A Comparative Study of Support Vector Machines Applied to the Word Sense Disambiguation Problem for the Medical Domain. In *Proceedings of the Second Indian Conference on Artificial Intelligence (IICAI-05)*, pages 3449–3468, Pune, India.

C. Leacock, M. Chodorow, and G. Miller. 1998. Using corpus statistics and WordNet relations for sense identification. *Computational Linguistics*, 24(1):147–165.

G. Leroy and T. Rindflesch. 2005. Effects of Information and Machine Learning algorithms on Word Sense Disambiguation with small datasets. *International Journal of Medical Informatics*, 74(7-8):573–585.

M. Lesk. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC Conference*, pages 24–26, Toronto, Canada.

H. Liu, S. Johnson, and C. Friedman. 2002. Automatic Resolution of Ambiguous Terms Based on Machine Learning and Conceptual Relations in the UMLS. *Journal of the American Medical Informatics Association*, 9(6):621–636.

H. Liu, V. Teller, and C. Friedman. 2004. A Multi-aspect Comparison Study of Supervised Word Sense Disambiguation. *Journal of the American Medical Informatics Association*, 11(4):320–331.

word	basic	90	Δ
adjustment	71	70	-1
association	100	100	0
blood_pressure	48	50	2
cold	88	86	-2
condition	89	90	1
culture	96	91	-5
degree	96	86	-10
depression	88	85	-3
determination	87	82	-5
discharge	95	92	-3
energy	98	99	1
evaluation	76	75	-1
extraction	85	82	-3
failure	66	71	5
fat	85	83	-2
fit	87	85	-2
fluid	100	100	0
frequency	95	94	-1
ganglion	97	95	-2
glucose	91	92	1
growth	70	67	-3
immunosuppression	79	79	0
implantation	90	88	-2
inhibition	98	98	0
japanese	73	75	2
lead	91	90	-1
man	87	82	-5
mole	95	84	-11
mosaic	87	83	-4
nutrition	53	43	-10
pathology	85	85	0
pressure	94	96	2
radiation	84	82	-2
reduction	89	90	1
repair	87	86	-1
resistance	98	97	-1
scale	86	79	-7
secretion	99	99	0
sensitivity	93	91	-2
sex	87	84	-3
single	99	99	0
strains	92	92	0
support	86	89	3
surgery	97	98	1
transient	99	99	0
transport	93	93	0
ultrasound	87	85	-2
variation	94	89	-5
weight	77	77	0
white	73	74	1
Average	87.2	85.6	-1.58

Table 3: Comparison of performance using original training data and 90 automatically generated examples

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding Predominant Senses in Untagged Text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Lingusitics (ACL-2004)*, pages 280–287, Barcelona, Spain.

B. McInnes, T. Pedersen, and J. Carlis. 2007. Using UMLS Concept Unique Identifiers (CUIs) for Word Sense Disambiguation in the Biomedical Domain. In *Proceedings of the Annual Symposium of the American Medical Informatics Association*, pages 533–537, Chicago, IL.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3: The Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain.

H. Ng, B. Wang, and S. Chan. 2003. Exploiting Parallel Texts for Word Sense Disambiguation: an Empirical Study. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics (ACL-03)*, pages 455–462, Sapporo, Japan.

H. Ng. 1997. Getting serious about Word Sense Disambiguation. In *Proceedings of the SIGLEX Workshop "Tagging Text with Lexical Semantics: What, why and how?"*, pages 1–7, Washington, DC.

T. Pedersen. 2001. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*, pages 79–86, Pittsburgh, PA., June.

S. Robertson and K. Spark Jones. 1976. Relevance weighting of search terms. *Journal of the American Society for Information Science and Technology*, 27(3):129–146.

J. Rocchio. 1971. Relevance feedback in Information Retrieval. In G. Salton, editor, *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall, Englewood Cliffs, NJ.

G. Salton. 1971. *The SMART Retrieval System – Experiments in Automatic Document Processing*. Prentice Hall Inc., Englewood Cliffs, NJ.

M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. 2008. Knowledge Sources for Word Sense Disambiguation of Biomedical Text. In *Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing at ACL 2008*, pages 80–87.

M. Weeber, J. Mork, and A. Aronson. 2001. Developing a Test Collection for Biomedical Word Sense Disambiguation. In *Proceedings of AMAI Symposium*, pages 746–50, Washington, DC.

D. Widdows, S. Peters, S. Cedernerg, C. Chan, D. Steffen, and P. Buitelaar. 2003. Unsupervised Monolingual and Bilingual Word-sense Disambiguation of Medical Documents using UMLS. In *Workshop on "Natural Langauge Processing in Biomedicine" at ACL 2003*, pages 9–16, Sapporo, Japan.