Knowledge Sources for Word Sense Disambiguation of Biomedical Text

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

David Martinez
Department of Computer Science
& Software Engineering
University of Melbourne
Victoria 3010
Australia
davidm@csse.unimelb.edu.au

Abstract

Like text in other domains, biomedical documents contain a range of terms with more than one possible meaning. These ambiguities form a significant obstacle to the automatic processing of biomedical texts. Previous approaches to resolving this problem have made use of a variety of knowledge sources including linguistic information (from the context in which the ambiguous term is used) and domain-specific resources (such as UMLS). In this paper we compare a range of knowledge sources which have been previously used and introduce a novel one: MeSH terms. The best performance is obtained using linguistic features in combination with MeSH terms. Results from our system outperform published results for previously reported systems on a standard test set (the NLM-WSD corpus).

1 Introduction

The number of documents discussing biomedical science is growing at an ever increasing rate, making it difficult to keep track of recent developments. Automated methods for cataloging, searching and navigating these documents would be of great benefit to researchers working in this area, as well as having potential benefits to medicine and other branches of science. Lexical ambiguity, the linguistic phenomena where a word or phrase has more than one potential meaning, makes the automatic processing of text difficult. For example, "cold" has six possible meanings in the Unified Medical Language System (UMLS) Metathesaurus (Humphreys et al., 1998) including "common cold", "cold sensation" and "Chronic Obstructive Airway Disease (COLD)". The NLM Indexing Initiative (Aronson et al., 2000) attempted to automatically index biomedical journals with concepts from the UMLS Metathesaurus and concluded that lexical ambiguity was the biggest challenge in the automation of the indexing process. Weeber et al. (2001) analysed MEDLINE abstracts and found that 11.7% of phrases were ambiguous relative to the UMLS Metathesaurus.

Word Sense Disambiguation (WSD) is the process of resolving lexical ambiguities. Previous researchers have used a variety of approaches for WSD of biomedical text. Some of them have taken techniques proven to be effective for WSD of general text and applied them to ambiguities in the biomedical domain, while others have created systems using domain-specific biomedical resources. However, there has been no direct comparison of which knowledge sources are the most useful or whether combining a variety of knowledge sources, a strategy which has been shown to be successful for WSD in the general domain (Stevenson and Wilks, 2001), improves results.

This paper compares the effectiveness of a variety of knowledge sources for WSD in the biomedical domain. These include features which have been commonly used for WSD of general text as well as information derived from domain-specific resources. One of these features is MeSH terms, which we find to be particularly effective when combined with generic features.

The next section provides an overview of various approaches to WSD in the biomedical domain. Sec-

BioNLP 2008: Current Trends in Biomedical Natural Language Processing, pages 80–87,
Columbus, Ohio, USA, June 2008.

tion 3 outlines our approach, paying particular attention to the range of knowledge sources used by our system. An evaluation of this system is presented in Section 4. Section 5 summarises this paper and provides suggestions for future work.

2 Previous Work

WSD has been actively researched since the 1950s and is regarded as an important part of the process of understanding natural language texts.

2.1 The NLM-WSD data set

Research on WSD for general text in the last decade has been driven by the SemEval evaluation frameworks[1] which provide a set of standard evaluation materials for a variety of semantic evaluation tasks. At this point there is no specific collection for the biomedical domain in SemEval, but a test collection for WSD in biomedicine was developed by Weeber et al. (2001), and has been used as a benchmark by many independent groups. The UMLS Metathesaurus was used to provide a set of possible meanings for terms in biomedical text. 50 ambiguous terms which occur frequently in MEDLINE were chosen for inclusion in the test set. 100 instances of each term were selected from citations added to the MEDLINE database in 1998 and manually disambiguated by 11 annotators. Twelve terms were flagged as "problematic" due to substantial disagreement between the annotators. There are an average of 2.64 possible meanings per ambiguous term and the most ambiguous term, "cold" has five possible meanings. In addition to the meanings defined in UMLS, annotators had the option of assigning a special tag ("none") when none of the UMLS meanings seemed appropriate.

Various researchers have chosen to evaluate their systems against subsets of this data set. Liu et al. (2004) excluded the 12 terms identified as problematic by Weeber et al. (2001) in addition to 16 for which the majority (most frequent) sense accounted for more than 90% of the instances, leaving 22 terms against which their system was evaluated. Leroy and Rindflesch (2005) used a set of 15 terms for which the majority sense accounted for less than 65% of the instances. Joshi et al. (2005) evaluated against the set union of those two sets, providing 28 ambiguous terms. McInnes et al. (2007) used the set intersection of the two sets (dubbed the "common subset") which contained 9 terms. The terms which form these various subsets are shown in Figure 1.

The 50 terms which form the NLM-WSD data set represent a range of challenges for WSD systems. 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. Despite its simplicity, the MFS heuristic is a hard baseline to beat, particularly for unsupervised systems, because it uses hand-tagged data to determine which sense is the most frequent. Analysis of the NLM-WSD data set showed that the MFS over all 50 ambiguous terms is 78%. The different subsets have lower MFS, indicating that the terms they contain are more difficult to disambiguate. The 22 terms used by (Liu et al., 2004) have a MFS of 69.9% while the set used by (Leroy and Rindflesch, 2005) has an MFS of 55.3%. The union and intersection of these sets have MFS of 66.9% and 54.9% respectively.

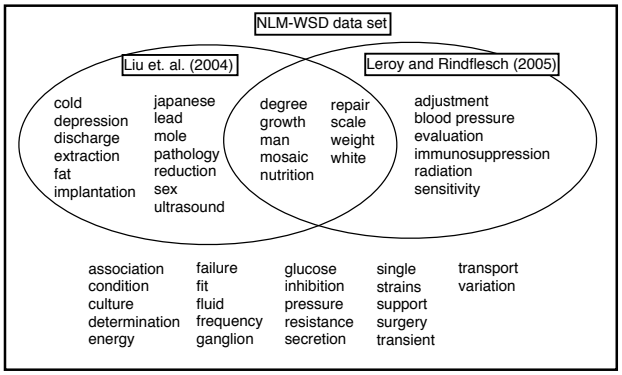

Figure 1: The NLM-WSD test set and some of its subsets. Note that the test set used by (Joshi et al., 2005) comprises the set union of the terms used by (Liu et al., 2004) and (Leroy and Rindflesch, 2005) while the "common subset" is formed from their intersection.

2.2 WSD of Biomedical Text

A standard approach to WSD is to make use of supervised machine learning systems which are trained on examples of ambiguous words in context along with the correct sense for that usage. The

[1] http://www.senseval.org

models created are then applied to new examples of that word to determine the sense being used.

Approaches which are adapted from WSD of general text include Liu et al. (2004). Their technique uses a supervised learning algorithm with a variety of features consisting of a range of collocations of the ambiguous word and all words in the abstract. They compared a variety of supervised machine learning algorithms and found that a decision list worked best. Their best system correctly disambiguated 78% the occurrences of 22 ambiguous terms in the NLM-WSD data set (see Section 2.1).

Joshi et al. (2005) also use collocations as features and experimented with five supervised learning algorithms: Support Vector Machines, Naive Bayes, decision trees, decision lists and boosting. The Support Vector Machine performed scoring 82.5% on a set of 28 words (see Section 2.1) and 84.9% on the 22 terms used by Liu et al. (2004). Performance of the Naive Bayes classifier was comparable to the Support Vector Machine, while the other algorithms were hampered by the large number of features.

Examples of approaches which have made use of knowledge sources specific to the biomedical domain include Leroy and Rindflesch (2005), who relied on information from the UMLS Metathesaurus assigned by MetaMap (Aronson, 2001). Their system used information about whether the ambiguous word is the head word of a phrase identified by MetaMap, the ambiguous word's part of speech, semantic relations between the ambiguous words and surrounding words from UMLS as well as semantic types of the ambiguous word and surrounding word. Naive Bayes was used as a learning algorithm. This approach correctly disambiguated 65.6% of word instances from a set of 15 terms (see Section 2.1). Humphrey et al. (2006) presented an unsupervised system that also used semantic types. They constructed semantic type vectors for each word from a large collection of MEDLINE abstracts. This allowed their method to perform disambiguation at a coarser level, without the need for labeled training examples. In most cases the semantic types can be mapped to the UMLS concepts but not for five of the terms in the NLM-WSD data set. Humphrey et al. (2006) reported 78.6% accuracy over the remaining 45. However, their approach could not be applied to all instances of ambiguous terms and, in particular, is unable to model the "none" tag. Their system could only assign senses to an average of 54% of the instances of each ambiguous term.

McInnes et al. (2007) made use of Concept Unique Identifiers (CUIs) from UMLS which are also assigned by MetaMap. The information contained in CUIs is more specific than in the semantic types applied by Leroy and Rindflesch (2005). For example, there are two CUIs for the term "culture" in UMLS: "C0010453: Anthropological Culture" and "C0430400: Laboratory Culture". The semantic type for the first of these is "Idea or Concept" and "Laboratory Procedure" for the second. McInnes et al. (2007) were interested in exploring whether the more specific information contained in CUIs was more effective than UMLS semantic types. Their best result was reported for a system which represented each sense by all CUIs which occurred at least twice in the abstract surrounding the ambiguous word. They used a Naive Bayes classifier as the learning algorithm. McInnes et al. (2007) reported an accuracy of 74.5% on the set of ambiguous terms tested by Leroy and Rindflesch (2005) and 80.0% on the set used by Joshi et al. (2005). They concluded that CUIs are more useful for WSD than UMLS semantic types but that they are not as robust as features which are known to work in general English, such as unigrams and bigrams.

3 Approach

Our approach is to adapt a state-of-the-art WSD system to the biomedical domain by augmenting it with additional domain-specific and domain-independent knowledge sources. Our basic system (Agirre and Martínez, 2004) participated in the Senseval-3 challenge (Mihalcea et al., 2004) with a performance close to the best system for the English and Basque lexical sample tasks. The system is based on a supervised learning approach. The features used by Agirre and Martínez (2004) are derived from text around the ambiguous word and are domain independent. We refer to these as *linguistic* features. This feature set has been adapted for the disambiguation of biomedical text by adding further linguistic features and two different types of domain-specific features: CUIs (as used by (McInnes et al., 2007)) and Medical Subject Heading (MeSH) terms.

3.1 Features

Our feature set contains a number of parameters which were set empirically (e.g. threshold for unigram frequency in the linguistic features). In addition, we use the entire abstract as the context of the ambiguous term for relevant features rather than just the sentence containing the term. Effects of varying these parameters are consistent with previous results (Liu et al., 2004; Joshi et al., 2005; McInnes et al., 2007) and are not reported in this paper.

Linguistic features: The system uses a wide range of domain-independent features which are commonly used 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[2] and (2) preceding/following lemma/word-form of the content words (adjective, adverb, noun and verb) in the same sentence with the target word. For example, consider the sentence below with the target word *adjustment*.

  > "Body surface area *adjustments* of initial heparin dosing..."

  The features would include the following: left-content-word-lemma "*area adjustment*", right-function-word-lemma "*adjustment of*", left-POS "NN NNS", right-POS "NNS IN", left-content-word-form "*area adjustments*", right-function-word-form "*adjustment of*", etc.

- Syntactic Dependencies: These features model 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. In the above example, "heparin" is noun-modifier feature of "adjustment".

[2] A maximum-entropy-based part of speech tagger was used (Ratnaparkhi, 1996) without the adaptation to the biomedical domain.

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

- Unigrams: Lemmas of unigrams which appear more frequently than a predefined threshold in the entire corpus, excluding those in a list of stopwords. We empirically set the threshold to 1. This feature was not used by Agirre and Martínez (2004), but Joshi et al. (2005) found them to be useful for this task.

Concept Unique Identifiers (CUIs): We follow the approach presented by McInnes et al. (2007) to generate features based on UMLS Concept Unique Identifiers (CUIs). The MetaMap program (Aronson, 2001) identifies all words and terms in a text which could be mapped onto a UMLS CUI. MetaMap does not disambiguate the senses of the concepts, instead it enumerates all the possible combinations of the concept names found. For example, MetaMap will segment the phrase "Body surface area adjustments of initial heparin dosing ..." into two chunks: "Body surface area adjustments" and "of initial heparin dosing". The first chunk will be mapped onto four CUIs with the concept name "Body Surface Area": "C0005902: Diagnostic Procedure" and "C1261466: Organism Attribute" and a further pair with the name "Adjustments": "C0456081: Health Care Activity" and "C0871291: Individual Adjustment". The final results from MetaMap for the first chunk will be eight combinations of those concept names, e.g. first four by second two concept names. CUIs which occur more than three times in the abstract containing the ambiguous word are included as features.

Medical Subject Headings (MeSH): The final feature is also specific to the biomedical domain. Medical Subject Headings (MeSH) (Nelson et al., 2002) is a controlled vocabulary for indexing biomedical and health-related information and documents. MeSH terms are manually assigned to abstracts by human indexers. The latest version of MeSH contains over 24,000 terms organised into an 11 level hierarchy.

The terms assigned to the abstract in which each ambiguous word occurs are used as features. For example, the abstract containing our example phrase has been assigned 16 MeSH

terms including "M01.060.116.100: Aged", "M01.060.116.100.080: Aged, 80 and over", "D27.505.954.502.119: Anticoagulants" and "G09.188.261.560.150: Blood Coagulation". To our knowledge MeSH terms have not been previously used as a feature for WSD of biomedical documents.

3.2 Learning Algorithms

We compared three machine leaning algorithms which have previously been shown to be effective for WSD tasks.

The **Vector Space Model** is a memory-based learning algorithm which was used by (Agirre and Martínez, 2004). 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.

The **Naive Bayes** classifier is based on a probabilistic model which assumes conditional independence of features given the target classification. It calculates the posterior probability that an instance belongs to a particular class given the prior probabilities of the class and the conditional probability of each feature given the target class.

Support Vector Machines have been widely used in classification tasks. SVMs map feature vectors onto a high dimensional space and construct a classifier by searching for the hyperplane that gives the greatest separation between the classes.

We used our own implementation of the Vector Space Model and Weka implementations (Witten and Frank, 2005) of the other two algorithms.

4 Results

This system was applied to the NLM-WSD data set. Experiments were carried out using each of the three types of features (linguistic, CUI and MeSH) both alone and in combination. Ten-fold cross validation was used, and the figures we report are averaged across all ten runs.

Results from this experiment are shown in Table 1 which lists the performance using combinations of learning algorithm and features. The figure shown for each configuration represents the percentage of instances of ambiguous terms which are correctly disambiguated.

These results show that each of the three types of knowledge (linguistic, CUIs and MeSH) can be used to create a classifier which achieves a reasonable level of disambiguation since performance exceeds the relevant baseline score. This suggests that each of the knowledge sources can contribute to the disambiguation of ambiguous terms in biomedical text.

The best performance is obtained using a combination of the linguistic and MeSH features, a pattern observed across all test sets and machine learning algorithms. Although the increase in performance gained from using both the linguistic and MeSH features compared to only the linguistic features is modest it is statistically significant, as is the difference between using both linguistic and MeSH features compared with using the MeSH features alone (Wilcoxon Signed Ranks Test, $p < 0.01$).

Combining MeSH terms with other features generally improves performance, suggesting that the information contained in MeSH terms is distinct from the other knowledge sources. However, the inclusion of CUIs as features does not always improve performance and, in several cases, causes it to fall. This is consistent with McInnes et al. (2007) who concluded that CUIs were a useful information source for disambiguation of biomedical text but that they were not as robust as a linguistic knowledge source (unigrams) which they had used for a previous system. The most likely reason for this is that our approach relies on automatically assigned CUIs, provided by MetaMap, while the MeSH terms are assigned manually. We do not have access to a reliable assignment of CUIs to text; if we had WSD would not be necessary. On the other hand, reliably assigned MeSH terms are readily available in Medline. The CUIs assigned by MetaMap are noisy while the MeSH terms are more reliable and prove to be a more useful knowledge source for WSD.

The Vector Space Model learning algorithm performs significantly better than both Support Vector Machine and Naive Bayes (Wilcoxon Signed Ranks Test, $p < 0.01$). This pattern is observed regardless

Data sets	Features						
	Linguistic	CUI	MeSH	CUI+ MeSH	Linguistic +MeSH	Linguistic +CUI	Linguistic+ MeSH+CUI
	Vector space model						
All words	87.2	85.8	81.9	86.9	**87.8**	87.3	87.6
Joshi subset	82.3	79.6	76.6	81.4	**83.3**	82.4	82.6
Leroy subset	77.8	74.4	70.4	75.8	**79.0**	78.0	77.8
Liu subset	84.3	81.3	78.3	83.4	**85.1**	84.3	84.5
Common subset	79.6	75.1	70.4	76.9	**80.8**	79.6	79.2
	Naive Bayes						
All words	86.2	81.2	85.7	81.1	86.4	81.4	81.5
Joshi subset	80.6	73.4	80.1	73.3	80.9	73.7	73.8
Leroy subset	76.4	66.1	74.6	65.9	76.8	66.3	66.3
Liu subset	81.9	75.4	81.7	75.3	82.2	75.5	75.6
Common subset	76.7	66.1	74.7	65.8	77.2	65.9	65.9
	Support Vector Machine						
All words	85.6	83.5	85.3	84.5	86.1	85.3	85.6
Joshi subset	79.8	76.4	79.5	78.0	80.6	79.1	79.8
Leroy subset	75.1	69.7	72.6	72.0	76.3	74.2	74.9
Liu subset	81.3	78.2	81.0	80.0	82.0	80.6	81.2
Common subset	75.7	69.8	71.6	73.0	76.8	74.7	75.2

		Previous Approaches			
	MFS baseline	Liu et. al. (2004)	Leroy and Rindflesch (2005)	Joshi et. al. (2005)	McInnes et. al. (2007)
All words	78.0	–	–	–	85.3
Joshi subset	66.9	–	–	82.5	80.0
Leroy subset	55.3	–	65.5	77.4	74.5
Liu subset	69.9	78.0	–	84.9	82.0
Common subset	54.9	–	68.8	79.8	75.7

Table 1: Results from WSD system applied to various sections of the NLM-WSD data set using a variety of features and machine learning algorithms. Results from baseline and previously published approaches are included for comparison.

of which set of features are used, and it is consistent of the results in Senseval data from (Agirre and Martínez, 2004).

4.1 Per-Word Analysis

Table 2 shows the results of our best performing system (combination of linguistic and MeSH features using the Vector Space Model learning algorithm). Comparable results for previous supervised systems are also reported where available.[3] The MFS baseline for each term is shown in the leftmost column.

The performance of Leroy and Rindflesch's system is always lower than the best result for each word. The systems reported by Joshi et al. (2005) and McInnes et al. (2007) are better than, or the same as, all other systems for 14 and 12 words respectively. The system reported here achieves results equal to or better than previously reported systems for 33 terms.

There are seven terms for which the performance of our approach is actually lower than the MFS baseline (shown in *italics*) in Table 2. (In fact, the baseline outperforms all systems for four of these terms.) The performance of our system is within 1% of the baseline for five of these terms. The remaining pair, "blood pressure" and "failure", are included in the set of problematic words identified by (Weeber et al., 2001). Examination of the possible senses show that they include pairs with similar meanings. For

[3]It is not possible to directly compare our results with Liu et al. (2004) or Humphrey et al. (2006). The first report only optimal configuration for each term (combination of feature sets and learning algorithm) while the second do not assign senses to all of the instances of each ambiguous term (see Section 2).

	MFS baseline	Leroy and Rindflesch (2005)	Joshi et. al. (2005)	McInnes et. al. (2007)	Reported system
adjustment	62	57	71	70	**74**
association	100	-	-	97	**100**
blood pressure	54	46	**53**	46	46
cold	86	-	**90**	89	88
condition	90	-	-	**89**	**89**
culture	89	-	-	94	**95**
degree	63	68	89	79	**95**
depression	85	-	86	81	**88**
determination	79	-	-	81	**87**
discharge	74	-	95	**96**	95
energy	99	-	-	**99**	98
evaluation	50	57	69	73	**81**
extraction	82	-	84	**86**	85
failure	71	-	-	**73**	67
fat	71	-	**84**	77	**84**
fit	82	-	-	87	**88**
fluid	100	-	-	99	**100**
frequency	94	-	-	**94**	**94**
ganglion	93	-	-	94	**96**
glucose	91	-	-	90	**91**
growth	63	62	**71**	69	68
immunosuppression	59	61	**80**	75	**80**
implantation	81	-	**94**	92	93
inhibition	98	-	-	**98**	**98**
japanese	73	-	**77**	76	75
lead	71	-	89	90	**94**
man	58	80	89	80	**90**
mole	83	-	**95**	87	93
mosaic	52	66	**87**	75	**87**
nutrition	45	48	52	49	**54**
pathology	85	-	**85**	84	**85**
pressure	96	-	-	93	**95**
radiation	61	72	82	81	**84**
reduction	89	-	91	**92**	89
repair	52	81	87	**93**	88
resistance	97	-	-	96	**98**
scale	65	84	81	83	**88**
secretion	99	-	-	**99**	**99**
sensitivity	49	70	88	92	**93**
sex	80	-	**88**	87	87
single	99	-	-	98	**99**
strains	92	-	-	92	**93**
support	90	-	-	**91**	89
surgery	98	-	-	94	**97**
transient	99	-	-	98	**99**
transport	93	-	-	**93**	**93**
ultrasound	84	-	**92**	85	90
variation	80	-	-	91	**95**
weight	47	68	**83**	79	81
white	49	62	**79**	74	76

Table 2: Per-word performance of best reported systems.

example, the two senses which account for 98% of the instances of "blood pressure", which refer to the blood pressure within an organism and the result obtained from measuring this quantity, are very closely related semantically.

5 Conclusion

This paper has compared a variety of knowledge sources for WSD of ambiguous biomedical terms and reported results which exceed the performance of previously published approaches. We found that accurate results can be achieved using a combination of linguistic features commonly used for WSD

of general text and manually assigned MeSH terms. While CUIs are a useful source of information for disambiguation, they do not improve the performance of other features when used in combination with them. Our approach uses manually assigned MeSH terms while the CUIs are obtained automatically using MetaMap.

The linguistic knowledge sources used in this paper comprise a wide variety of features including n-grams and syntactic dependencies. We have not explored the effectiveness of these individually and this is a topic for further work.

In addition, our approach does not make use of the fact that MeSH terms are organised into a hierarchy. It would be interesting to discover whether this information could be used to improve WSD performance. Others have developed techniques to make use of hierarchical information in WordNet for WSD (see Budanitsky and Hirst (2006)) which could be adapted to MeSH.

References

E. Agirre and D. Martínez. 2004. 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.

A. Aronson, O. Bodenreider, H. Chang, S. Humphrey, J. Mork, S. Nelson, T. Rindflesch, and W. Wilbur. 2000. The NLM Indexing Initiative. In *Proceedings of the AMIA Symposium*.

A. Aronson. 2001. Effective mapping of biomedical text to the UMLS Metathesaurus: the MetaMap program. In *Proceedings of the American Medical Informatics Association (AMIA)*, pages 17–21.

A. Budanitsky and G. Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

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.

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.

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.

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.

S. Nelson, T. Powell, and B. Humphreys. 2002. The Unified Medical Language System (UMLS) Project. In Allen Kent and Carolyn M. Hall, editors, *Encyclopedia of Library and Information Science*. Marcel Dekker, Inc.

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.

A. Ratnaparkhi. 1996. A Maximum Entropy Model for Part-of-Speech Tagging. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 133–142.

M. Stevenson and Y. Wilks. 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27(3):321–350.

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.

I. Witten and E. Frank. 2005. *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann, San Francisco.