Scaling up WSD with Automatically Generated Examples

Weiwei Cheng, Judita Preiss and **Mark Stevenson**
Department of Computer Science,
Sheffield University,
Regent Court, 211 Portobello,
Sheffield, S1 4DP
United Kingdom
{W.Cheng, J.Preiss, M.Stevenson}@dcs.shef.ac.uk

Abstract

The most accurate approaches to Word Sense Disambiguation (WSD) for biomedical documents are based on supervised learning. However, these require manually labeled training examples which are expensive to create and consequently supervised WSD systems are normally limited to disambiguating a small set of ambiguous terms. An alternative approach is to create labeled training examples automatically and use them as a substitute for manually labeled ones. This paper describes a large scale WSD system based on automatically labeled examples generated using information from the UMLS Metathesaurus. The labeled examples are generated without any use of labeled training data whatsoever and is therefore completely unsupervised (unlike some previous approaches). The system is evaluated on two widely used data sets and found to outperform a state-of-the-art unsupervised approach which also uses information from the UMLS Metathesaurus.

1 Introduction

The information contained in the biomedical literature that is available in electronic formats is useful for health professionals and researchers (Westbrook et al., 2005). The amount is so vast that it is difficult for researchers to identify information of interest without the assistance of automated tools (Krallinger and Valencia, 2005). However, processing these documents automatically is made difficult by the fact that they contain terms that are ambiguous. For example, “culture” can mean “laboratory procedure” (e.g. “In peripheral blood mononuclear cell *culture*”) or “anthropological culture” (e.g. “main accomplishments of introducing a quality management *culture*”). These lexical ambiguities are problematic for language understanding systems.

Word sense disambiguation (WSD) is the process of automatically identifying the meanings of ambiguous terms. Some WSD systems for the biomedical domain are only able to disambiguate a small number of ambiguous terms (see Section 2). However, for WSD systems to be useful in applications they should be able to disambiguate all ambiguous terms. One way to create such a WSD system is to automatically create the labeled data that is used to train supervised WSD systems. Several approaches (Liu et al., 2002; Stevenson and Guo, 2010; Jimeno-Yepes and Aronson, 2010) have used information from the UMLS Metathesaurus[1] to create labeled training data that have successfully been used to create WSD systems.

A key decision for any system that automatically generates labeled examples is the number of examples of each sense to create, known as the *bias* of the data set. It has been shown that the bias of a set of labeled examples affects the performance of the WSD system it is used to train (Mooney, 1996; Agirre and Martínez, 2004b). Some of the previous approaches to generating labeled data relied on manually annotated examples to determine the bias of the data sets and were therefore not completely unsupervised.

This paper describes the development of a large scale WSD system that is able to disambiguate all

[1] http://www.nlm.nih.gov/research/umls/

Proceedings of the 2012 Workshop on Biomedical Natural Language Processing (BioNLP 2012), pages 231–239,
Montréal, Canada, June 8, 2012. ©2012 Association for Computational Linguistics

terms that are ambiguous in the UMLS Metathesaurus. The system relies on labeled examples that are created using information from UMLS. Various bias options are explored, including ones that do not make use of information from manually labeled examples, and thus we can create a completely unsupervised system. Evaluation is carried out on two standard datasets (the NLM-WSD and MSH-WSD corpora). We find that WSD systems can be created without using any information from manually labeled examples and that their performance is better than a state-of-the-art unsupervised approach.

The remainder of this paper is organized as follows. Previous approaches to WSD in biomedical documents are described in the next Section. Section 3 presents the methods used to identify bias in the labeled examples and WSD system. Experiments in which these approaches are compared are described in Section 4 and their results in Section 5.

2 Background

Many WSD systems for the biomedical domain are based on supervised learning (McInnes et al., 2007; Xu et al., 2007; Stevenson et al., 2008; Yepes and Aronson, 2011). These systems require labeled training data, examples of an ambiguous term labeled with the correct meaning. Some sets of labeled data have been developed for the biomedical domain (Weeber et al., 2001; Savova et al., 2008; Jimeno-Yepes et al., 2011). However, these data sets only contain examples for a few hundred terms and can only be used to develop WSD systems to identify the meanings of those terms. The process of creating labeled examples is extremely time-consuming and difficult (Artstein and Poesio, 2008), making it impractical to create labeled examples of all possible ambiguous terms found in biomedical documents.

Two alternative approaches have been explored to develop systems which are able to disambiguate all ambiguous terms in biomedical documents. The first makes use of unsupervised WSD algorithms (see Section 2.1) and the second creates labeled data automatically and uses it to train a supervised WSD system (see Section 2.2).

2.1 Unsupervised WSD

Unsupervised WSD algorithms make use of information from some knowledge source, rather than relying on training data.

Humphrey et al. (2006) describe an unsupervised system which uses semantic types in UMLS to distinguish between the possible meanings of ambiguous words. However, it cannot disambiguate between senses with the same semantic type, i.e., it is not possible for the system to recognise all sense distinctions.

The *Personalised Page Rank (PPR)* system (Agirre et al., 2010; Jimeno-Yepes and Aronson, 2010) relies on a a graph-based algorithm similar to the Page Rank algorithm originally developed for use in search engines (Brin, 1998). It performs WSD by converting the UMLS Metathesaurus into a graph in which the possible meanings of ambiguous words are nodes and relations between them are edges. Disambiguation is carried out by providing the algorithm with a list of senses that appear in the text that is being disambiguated. This information is then combined with the graph and a ranked list of the possible senses for each ambiguous word generated.

Unsupervised systems have the advantage of being able to disambiguate all ambiguous terms. However, the performance of unsupervised systems that have been developed for biomedical documents is lower than that of supervised ones.

2.2 Automatic Generation of Labeled Data

Automatic generation of labeled data for WSD combines the accuracy of supervised approaches with the ability of unsupervised approaches to disambiguate all ambiguous terms. It was first suggested by Leacock et al. (1998). Their approach is based on the observation that some terms in a lexicon occur only once and, consequently, there is no doubt about their meaning. These are referred to as being *monosemous*. Examples for each possible meaning of an ambiguous term are generated by identifying the closest monosemous term (the *monosemous relative*) in the lexicon and using examples of that term. Variants of the approach have been applied to the biomedical domain using the UMLS Metathesaurus as the sense inventory.

Liu et al. (2002) were the first to apply the monosemous relatives approach to biomedical WSD and use it to disambiguate a set of 35 abbreviations. They reported high precision but low recall, indicating that labeled examples could not be created for many of the abbreviations. Jimeno-Yepes and Aronson (2010) applied a similar approach and found that it performed better than a number of alternative approaches on a standard evaluation resource (the NLM-WSD corpus) but did not perform as well as supervised WSD. Stevenson and Guo (2010) compared two techniques for automatically creating labeled data, including the monosemous relatives approach. They found that the examples which were generated were as good as manually labeled examples when used to train a supervised WSD system. However, Stevenson and Guo (2010) relied on labeled data to determine the number of examples of each sense to create, and therefore the bias of the data set. Consequently their approach is not completely unsupervised since it could not be applied to ambiguous terms that do not have labeled training data available.

3 Approach

3.1 WSD System

The WSD system is based on a supervised approach that has been adapted for the biomedical domain (Stevenson et al., 2008). The system was tested on the NLM-WSD corpus (see Section 4.1) and found to outperform alternative approaches.

The system can exploit a wide range of features, including several types of linguistic information from the context of an ambiguous term, MeSH codes and Concept Unique Identifiers (CUIs) from the UMLS Metathesaurus. However, computing these features for every example is a time consuming process and to make the system suitable for large scale WSD it was restricted to using a smaller set of features. Previous experiments (Stevenson et al., 2008) showed that this only leads to a small drop in disambiguation accuracy while significantly reducing the computational cost of generating features.

3.1.1 Features

Two types of context words are used as features: the lemmas of all content words in the same sentence as the ambiguous word and the lemmas of all content words in a ±4-word window around the ambiguous term. A list of corpus-specific stopwords was created containing terms that appear frequently in Medline abstracts but which are not useful for disambiguation (e.g. “abstract”, “conclusion”). Any lemmas found in this list were not used as features.

3.1.2 Learning algorithm

Disambiguation is carried out using the *Vector Space Model*, a memory-based learning algorithm in which each occurrence of an ambiguous word is represented as a vector created using the features extracted to represent it (Agirre and Martínez, 2004a). The Vector Space Model was found to outperform other learning algorithms when evaluated using the NLM-WSD corpus (Stevenson et al., 2008).

During the algorithm's training phase a single centroid vector, $\vec{C_{s_j}}$, is generated for each possible sense, s_j. This is shown in equation 1 where T is the set of training examples for a particular term and $sense(\vec{t})$ is the sense associated with the vector $\vec{t}$.

$$\vec{C_{s_j}} = \frac{\sum_{\vec{t_i}\ \epsilon\ T:sense(\vec{t_i})=s_j} \vec{t_i}}{|\vec{t_i}\ \epsilon\ T : sense(\vec{t_i}) = s_j|} \qquad (1)$$

Disambiguation is carried out by comparing the vector representing the ambiguous word, $\vec{a}$, against the centroid of each sense using the cosine metric, shown in equation 2, and choosing the one with the highest score.

$$score(s_j, \vec{a}) = cos(\vec{C_{s_j}}, \vec{a}) = \frac{\vec{C_{s_j}}.\vec{a}}{|\vec{C_{s_j}}||\vec{a}|} \qquad (2)$$

Note that the learning algorithm does not explicitly model the prior probability of each possible sense, unlike alternative approaches (e.g. Naive Bayes), since it was found that including this information did not improve performance.

3.2 Automatically generating training examples

The approaches used for generating training examples used here are based on the work of Stevenson and Guo (2010), who describe two approaches:

1. Monosemous relatives

2. Co-occurring concepts

Both approaches are provided with a set of ambiguous CUIs from the UMLS Metathesaurus, which represent the possible meanings of an ambiguous term, and a target number of training examples to be generated for each CUI. Each CUI is associated with at least one term and each term is labeled with a lexical unique identifier (LUI) which represents a range of lexical variants for a particular term. The UMLS Metathesaurus contains a number of data files which are exploited within these techniques, including:

`AMBIGLUI`: a list of cases where a LUI is linked to multiple CUIs.

`MRCON`: every string or concept name in the Metathesaurus appears in this file.

`MRCOC`: co-occuring concepts.

For the monosemous relatives approach, the strings of monosemous LUIs of the target CUI and its relatives are used to search Medline to retrieve training examples. The monosemous LUIs related to a CUI are defined as any LUIs associated with the CUI in the `MRCON` table and not listed in `AMBIGLUI` table.

The co-occurring concept approach works differently. Instead of using strings of monosemous LUIs of the target CUI and its relatives, the strings associated with LUIs of a number of co-occurring CUIs of the target CUI and its relatives found in MRCOC table are used. The process starts by finding the LUIs of the top n co-occurring CUIs of the target CUI. These LUIs are then used to form search queries. The query is quite restrictive in the beginning and requires all terms appear in the Medline citations files. Subsequently queries are made less restrictive by reducing the number of required terms in the query.

These techniques were used to generate labeled examples for all terms that are ambiguous in the 2010 AB version of the UMLS Metathesaurus.[2] The set of all ambiguous terms was created by analysing the `AMBIGLUI` table, to identify CUIs that are associated with multiple LUIs. The Medline Baseline Repository (MBR)[3] was also analysed and it was found that some terms were ambiguous in this resource, in the sense that more than one CUI had been assigned to an instance of a term, but could not be identified from the `AMBIGLUI` table. The final list of ambiguous CUIs was created by combining those identified from the `AMBIGLUI` table and those find in the MBR. This list contained a total of 103,929 CUIs.

Both techniques require large number of searches over the Medline database and to carry this out efficiently the MBR was indexed using the Lucene Information Retrieval system[4] and all searches executed locally.

Examples were generated using both approaches. The monosemous relatives approach generated examples for 98,462 CUIs and the co-occurring concepts for 98,540. (Examples generated using the monosemous relatives approach were preferred for the experiments reported later.) However, neither technique was able to generate examples for 5,497 CUIs, around 5% of the total. This happened when none of the terms associated with a CUI returned any documents when queried against the MBR and that CUI does not have any monosemous relatives. An example is C1281723 “Entire nucleus pulposus of intervertebral disc of third lumbar vertebra”. The lengthy terms associated with this CUI do not return any documents when used as search terms and, in addition, it is only related to one other CUI (C0223534 “Structure of nucleus pulposus of intervertebral disc of third lumbar vertebra”) which is itself only connected to C1281723. Fortunately there are relatively few CUIs for which no examples could be generated and none of them appear in the MBR, suggesting they refer to UMLS concepts that do not tend to be mentioned in documents.

3.3 Generating Bias

Three different techniques for deciding the number of training examples to be generated for each CUI (i.e. the bias) were explored.

Uniform Bias (UB) uses an equal number of training examples to generate centroid vectors for each of the possible senses of the ambiguous term.

Gold standard bias (GSB) is similar to the uniform bias but instead of being the same for all possible CUIs the number of training examples for each CUI is determined by the number of times it appears

[2] Stevenson and Guo (2010) applied them to a small set of examples from the NLM-WSD data set (see Section 4.1).

[3] `http://mbr.nlm.nih.gov`

[4] `http://lucene.apache.org/`

in a manually labeled gold standard corpus. Assume t is an ambiguous term and C_t is the set of possible meanings (CUIs). The number of training examples used to generate the centroid for that CUI, E_c, is computed according to equation 3 where G_c is the number of instances in the gold standard corpus annotated with CUI c and n is a constant which is set to 100 for these experiments.[5]

$$E_c = \frac{G_c}{\sum_{c_i \; \epsilon \; C_t} G_{c_{i,t}}}.n \tag{3}$$

The final technique, **Metamap Baseline Repository Bias (MBB)**, is based on the distribution of CUIs in the MBR. The number of training examples are generated in a similar way to the gold standard bias with MBR being used instead of a manually labeled corpus and is shown in equation 4 where M_c is the number of times the CUI c appears in the MBR.

$$E_c = \frac{M_c}{\sum_{c_i \; \epsilon \; C_t} M_{c_i}}.n \tag{4}$$

For example, consider the three possible CUIs associated with term "adjustment" in the NLM-WSD corpus: C0376209, C0456081 and C0683269[6]. The corpus contains 18 examples of C0376209, 62 examples of C0456081 and 13 of C0683269. Using equation 3, the number of training examples when GSB is applied for C0376209 is 20, 67 for C0456081 and 14 for C0683269. In the Metamap Baseline Repository files, C0376209 has a frequency count of 98046, C0456081 a count of 292809 and C0683269 a count of 83530. Therefore the number of training examples used for the three senses when applying MBB is: 21 for C0376209, 62 for C0456081 and 18 for C0683269.

4 Evaluation

4.1 Data sets

We evaluate our system on two datasets: the NLM-WSD and MSH-WSD corpora.

The NLM-WSD corpus[7] (Weeber et al., 2001) has been widely used for experiments on WSD in the biomedical domain, for example (Joshi et al., 2005; Leroy and Rindflesch, 2005; McInnes et al., 2007; Savova et al., 2008). It contains 50 ambiguous terms found in Medline with 100 examples of each. These examples were manually disambiguated by 11 annotators. The guidelines provided to the annotators allowed them to label a senses as "None" if none of the concepts in the UMLS Metathesaurus seemed appropriate. These instances could not be mapped onto UMLS Metathesaurus and were ignored for our experiments.

The larger MSH-WSD corpus (Jimeno-Yepes et al., 2011) contains 203 strings that are associated with more than one possible MeSH code in the UMLS Metathesaurus. 106 of these are ambiguous abbreviations, 88 ambiguous terms and 9 a combination of both. The corpus contains up to 100 examples for each possible sense and a total of 37,888 examples of ambiguous strings taken from Medline. Unlike the NLM-WSD corpus, all of the instances can be mapped to the UMLS Metathesaurus and none was removed from the dataset for our experiments.

The two data sets differ in the way the number of instances of each sense was determined. For the NLM-WSD corpus manual annotation is used to decide the number of instances that are annotated with each sense of an ambiguous term. However, the NLM-MSH corpus was constructed automatically and each ambiguous term has roughly the same number of examples of each possible sense.

4.2 Experiments

The WSD system described in Section 3 was tested using each of the three techniques for determining the bias, i.e. number of examples generated for each CUI. Performance is compared against various alternative approaches.

Two supervised approaches are included. The first, most frequent sense (MFS) (McCarthy et al., 2004), is widely used baseline for supervised WSD systems. It consists of assigning each ambiguous term the meaning that is more frequently observed in the training data. The second supervised approach

[5] Small values for E_c are rounded up to ensure that any rare CUIs have at least one training example.

[6] These CUIs are obtained using the mappings from NLM-WSD senses to CUIs available on the NLM website: http://wsd.nlm.nih.gov/collaboration.shtml

[7] http://wsd.nlm.nih.gov

is to train the WSD system using manually labeled examples from the NLM-WSD and MSH-WSD corpora. 10-fold cross validation is applied to evaluate this approach.

Performance of the Personalised Page Rank approach described in Section 2.1 is also provided to allow comparison with an unsupervised algorithm. Both Personalised Page Rank and the techniques we employ to generate labeled data, base disambiguation decisions on information from the UMLS Metathesaurus.

The performance of all approaches is measured in terms of the percentage of instances which are correctly disambiguated for each term with the average across all terms reported. Confidence intervals (95%) computed using bootstrap resampling (Noreen, 1989) are also shown.

5 Results

Results of the experiments are shown in Table 1 where the first three rows show performance of the approach described in Section 3 using the three methods for computing the bias (UB, MMB and GSB). MFS and Sup refer to the Most Frequent Sense supervised baseline and using manually labeled examples, respectively, and PPR to the Personalised PageRank approach.

When the performance of the approaches using automatically labeled examples (UB, MMB and GSB) is compared it is not surprising that the best results are obtained using the gold standard bias since this is obtained from manually labeled data. Results using this technique for computing bias always outperform the other two, which are completely unsupervised and do not make use of any information from manually labeled data. However, the improvement in performance varies according to the corpus, for the NLM-WSD corpus there is an improvement of over 10% in comparison to UB while the corresponding improvement for the MSH-WSD corpus is less than 0.5%.

A surprising result is that performance obtained using the uniform bias (UB) is consistently better than using the bias obtained by analysis of the MBR (MMB). It would be reasonable to expect that information about the distribution of CUIs in this corpus would be helpful for WSD but it turns out that making no assumptions whatsoever about their relative frequency, i.e., assigning a uniform baseline, produces better results.

The relative performance of the supervised (MFS, Sup and GSB) and unsupervised approaches (UB, MMB and PPR) varies according to the corpus. Unsurprisingly using manually labeled data (Sup) outperforms all other approaches on both corpora. The supervised approaches also outperform the unsupervised ones on the NLM-WSD corpus. However, for the MSH-WSD corpus all of the unsupervised approaches outperform the MFS baseline.

A key reason for the differences in these results is the different distributions of senses in the two corpora, as shown by the very different performance of the MFS approach on the two corpora. This is discussed in more detail later (Section 5.2).

Comparison of the relative performance of the unsupervised approaches (UB, MMB and PPR) shows that training a supervised system with the automatically labeled examples using a uniform bias (UB) always outperforms PPR. This demonstrates that this approach outperforms a state-of-the-art unsupervised algorithm that relies on the same information used to generate the examples (the UMLS Metathesaurus).

5.1 Performance by Ambiguity Type

The MSH-WSD corpus contains both ambiguous terms and abbreviations (see Section 4.1). Performance of the approaches on both types of ambiguity are shown in Table 2.

	MSH-WSD Ambiguity Type	
Approach	Abbreviation	Term
UB	91.40 [91.00, 91.75]	72.68 [72.06, 73.32]
MMB	84.43 [83.97, 84.89]	69.45 [68.86, 70.10]
GSB	90.82 [90.45, 91.22]	73.96 [73.40, 74.62]
MFS	52.43 [51.73, 53.05]	51.76 [51.11, 52.36]
Sup.	97.41 [97.19, 97.62]	91.54 [91.18, 91.94]
PPR	86.40 [86.00, 86.85]	68.40 [67.80, 69.14]

Table 2: WSD evaluation results for abbreviations and terms in the MSH-WSD data set.

The relative performance of the different approaches on the terms and abbreviations is similar to the entire MSH-WSD data set (see Table 1). In par-

		Corpus	
Approach	Type	NLM-WSD	MSH-WSD
UB	Unsup.	74.00 [72.80, 75.29]	83.19 [82.87, 83.54]
MMB	Unsup.	71.18 [69.94, 72.38]	78.09 [77.70, 78.46]
GSB	Sup.	84.28 [83.12, 85.36]	83.39 [83.08, 83.67]
MFS	Sup.	84.70 [83.67, 85.81]	52.01 [51.50, 52.45]
Sup	Sup.	90.69 [89.87, 91.52]	94.83 [94.63, 95.02]
PPR	Unsup.	68.10 [66.80, 69.23]	78.60 [78.23, 78.90]

Table 1: WSD evaluation results on NLM-WSD and MSH-WSD data sets.

ticular using automatically generated examples with a uniform bias (UB) outperforms using the bias derived from the Medline Baseline Repository (MBR) while using the gold standard baseline (GSB) improves results slightly for terms and actually reduces them for abbreviations.

Results for all approaches are higher when disambiguating abbreviations than terms which is consistent with previous studies that have suggested that in biomedical text abbreviations are easier to disambiguate than terms.

5.2 Analysis

An explanation of the reason for some of the results can be gained by looking at the distributions of senses in the various data sets used for the experiments. Kullback-Leibler divergence (or KL divergence) (Kullback and Leibler, 1951) is a commonly used measure for determining the difference between two probability distributions. For each term t, we define S as the set of possible senses of t, the sense probability distributions of t as D and D'. Then the KL divergence between the sense probability distributions D and D' can be calculated according to equation 5.

$$KL(D||D') = \sum_{s \, \epsilon \, S} D(s). \log \frac{D(s)}{D'(s)} \qquad (5)$$

The three techniques for determining the bias described in Section 3.3 each generate a probability distribution over senses. Table 2 shows the average KL divergence when the gold standard distribution obtained from the manually labeled data (GSB) is compared with the uniform bias (UB) and bias obtained by analysing the Medline Baseline Repository (MMB).

	Corpus	
Avg. KL Divergence	NLM-WSD	MSH-WSD
$KL(GSB\|\|MMB)$	0.5649	0.4822
$KL(GSB\|\|UB)$	0.4600	0.0406

Table 3: Average KL divergence of sense probability distributions in the NLM-WSD and MSH-WSD data sets.

The average KL divergence scores in the table are roughly similar with the exception of the much lower score obtained for the gold-standard and uniform bias for the MSH-WSD corpus (0.0406). This is due to the fact that the MSH-WSD corpus was designed to have roughly the same number of examples for each sense, making the sense distribution close to uniform (Jimeno-Yepes et al., 2011). This is evident from the MFS scores for the MSH-WSD corpus which are always close to 50%. This also provides as explanation of why performance using automatically generated examples on the MSH-WSD corpus only improves by a small amount when the gold standard bias is used (see Table 1). The gold standard bias simply does not provide much additional information to the WSD system. The situation is different in the NLM-WSD corpus, where the MFS score is much higher. In this case the additional information available in the gold standard sense distribution is useful for the WSD system and leads to a large improvement in performance.

In addition, this analysis demonstrates why performance does not improve when the bias generated from the MBR is used. The distributions which are obtained are different from the gold standard and are therefore mislead the WSD system rather than providing useful information. The difference between these distributions would be expected for the MSH-WSD corpus, since it contains roughly the same number of examples for each possible sense and does not attempt to represent the relative frequency of the different senses. However, it is surprising to observe a similar difference for the NLM-WSD corpus, which does not have this constraint. The difference suggests the information about CUIs in the MBR, which is generated automatically, has some limitations.

Table 4 shows a similar analysis for the MSH-WSD corpus when abbreviations and terms are considered separately and supports this analysis. The figures in this table show that the gold standard and uniform distributions are very similar for both abbreviations and terms, which explains the similar results for UB and GSB in Table 2. However, the gold standard distribution is different from the one obtained from the MBR. The drop in performance of MMB compared with GBS in Table 2 is a consequence of this.

	Ambiguity Type	
Avg. KL Divergence	Abbreviation	Term
$KL(GSB\|\|MMB)$	0.4554	0.4603
$KL(GSB\|\|UB)$	0.0544	0.0241

Table 4: Average KL divergence for abbreviations and terms in the MSH-WSD data set.

6 Conclusion

This paper describes the development of a large scale WSD system based on automatically labeled examples. We find that these examples can be generated for the majority of CUIs in the UMLS Metathesaurus. Evaluation on the NLM-WSD and MSH-WSD data sets demonstrates that the WSD system outperforms the PPR approach without making any use of labeled data.

Three techniques for determining the number of examples to use for training are explored. It is found that a supervised approach (which makes use of manually labeled data) provides the best results. Surprisingly it was also found that using information from the MBR did not improve performance. Analysis showed that the sense distributions extracted from the MBR were different from those observed in the evaluation data, providing an explanation for this result.

Evaluation showed that accurate information about the bias of training examples is useful for WSD systems and future work will explore other unsupervised ways of obtaining this information. Alternative techniques for generating labeled examples will also be explored. In addition, further evaluation of the WSD system will be carried out, such as applying it to an all words task and within applications.

Acknowledgements

This research has been supported by the Engineering and Physical Sciences Research Council and a Google Research Award.

References

E. Agirre and D. Martínez. 2004a. The Basque Country University system: English and Basque tasks. In Rada Mihalcea and Phil Edmonds, editors, *Proceedings of Senseval-3*, pages 44–48, Barcelona, Spain.

E. Agirre and D. Martínez. 2004b. Unsupervised WSD Based on Automatically Retrieved Examples: The Importance of Bias. In *Proceedings of EMNLP-04*, Barcelona, Spain.

E. Agirre, A. Sora, and M. Stevenson. 2010. Graph-based word sense disambiguation of biomedical documents. *Bioinformatics*, 26(22):2889–2896.

R. Artstein and M. Poesio. 2008. Inter-Coder Agreement for Computational Linguistics. *Computational Linguistics*, 34(4):555–596.

S. Brin. 1998. Extracting Patterns and relations from the Word-Wide Web. In *Proceedings of WebDB'98*.

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.

A. Jimeno-Yepes and A. Aronson. 2010. Knowledge-based biomedical word sense disambiguation: comparison of approaches. *BMC Bioinformatics*, 11(1):569.

A. Jimeno-Yepes, B. McInnes, and A. Aronson. 2011. Exploiting MeSH indexing in MEDLINE to generate a data set for word sense disambiguation. *BMC Bioinformatics*, 12(1):223.

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 IICAI-05*, pages 3449–3468, Pune, India.

M. Krallinger and A. Valencia. 2005. Text mining and information retrieval services for molecular biology. *Genome Biology*, 6(7):224.

S. Kullback and R. A. Leibler. 1951. On Information and Sufficiency. *The Annals of Mathematical Statistics*, 22(1):79–86.

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.

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.

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding Predominant Word Senses in Untagged Text. In *Proceedings of 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 AMIA Symposium*, pages 533–537, Chicago, IL.

R. Mooney. 1996. Comparative Experiments on Disambiguating Word Senses: An Illustration of the Role of Bias in Machine Learning. In *Proceedings of EMNLP-96*, pages 82–91, Philadelphia, PA.

E. W. Noreen. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

G. Savova, A. Coden, I. Sominsky, R. Johnson, P. Ogren, C. de Groen, and C. Chute. 2008. Word Sense Disambiguation across Two Domains: Biomedical Literature and Clinical Notes. *Journal of Biomedical Informatics*, 41(6):1088–1100.

M. Stevenson and Y. Guo. 2010. Disambiguation of Ambiguous Biomedical Terms using Examples Generated from the UMLS Metathesaurus. *Journal of Biomedical Informatics*, 43(5):762–773.

M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. 2008. Disambiguation of biomedical text using diverse sources of information. *BMC Bioinformatics*, 9(Suppl 11):S7.

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

J. Westbrook, E. Coiera, and A. Gosling. 2005. Do Online Information Retrieval Systems Help Experienced Clinicians Answer Clinical Questions? *Journal of the American Medical Informatics Association*, 12:315–321.

H. Xu, J. Fan, G. Hripcsak, E. Mendonça, Markatou M., and Friedman C. 2007. Gene symbol disambiguation using knowledge-based profiles. *Bioinformatics*, 23(8):1015–22.

A. Jimeno Yepes and A. Aronson. 2011. Self-training and co-training in biomedical word sense disambiguation. In *Proceedings of BioNLP 2011 Workshop*, pages 182–183, Portland, Oregon, USA, June.