The Effect of Ambiguity on the Automated Acquisition of WSD Examples

Mark Stevenson and Yikun Guo
Department of Computer Science,
University of Sheffield,
Regent Court, 211 Portobello,
Sheffield, S1 4DP
United Kingdom
m.stevenson@dcs.shef.ac.uk and g.yikun@dcs.shef.ac.uk

Abstract

Several methods for automatically generating labeled examples that can be used as training data for WSD systems have been proposed, including a semi-supervised approach based on relevance feedback (Stevenson et al., 2008a). This approach was shown to generate examples that improved the performance of a WSD system for a set of ambiguous terms from the biomedical domain. However, we find that this approach does not perform as well on other data sets. The levels of ambiguity in these data sets are analysed and we suggest this is the reason for this negative result.

1 Introduction

Several studies, for example (Mihalcea et al., 2004; Pradhan et al., 2007), have shown that supervised approaches to Word Sense Disambiguation (WSD) outperform unsupervised ones. But these rely on labeled training data which is difficult to create and not always available (e.g. (Weeber et al., 2001)). Various techniques for creating labeled training data automatically have been suggested in the literature. Stevenson et al. (2008a) describe a semi-supervised approach that used relevance feedback (Rocchio, 1971) to analyse existing labeled examples and use the information produced to generate further ones. The approach was tested on the biomedical domain and the additional examples found to improve performance of a WSD system. However, biomedical documents represent a restricted domain. In this paper the same approach is tested against two data sets that are not limited to a single domain.

2 Application to a Range of Data Sets

In this paper the relevance feedback approach described by Stevenson et al. (2008a) is evaluated using three data sets: the **NLM-WSD** corpus (Weeber et al., 2001) which Stevenson et al. (2008a) used for their experiments, the **Senseval-3** lexical sample task (Mihalcea et al., 2004) and the coarse-grained version of the **SemEval** English lexical sample task (Pradhan et al., 2007).

2.1 Generating Examples

To generate examples for a particular sense of an ambiguous term all of the examples where the term is used in that sense are considered to be "relevant documents" while the examples in which any other sense of the term is used are considered to be "irrelevant documents". Relevance feedback (Rocchio, 1971) is used to generate a set of query terms designed to identify relevant documents, and therefore instances of the sense. The top five query terms are used to retrieve documents and these are used as labeled examples of the sense. Further details of this process are described by Stevenson et al. (2008a).

This process requires a collection of documents that can be queried to generate the additional examples. For the NLM-WSD data set we used PubMed, a database of biomedical journal abstracts queried using the Entrez retrieval system (http://www.ncbi.nlm.nih.gov/sites/gquery). The British National Corpus (BNC) was used for Senseval-3 and SemEval.[1] Lucene (http://lucene.apache.org) was used to index the BNC and retrieve examples.

[1] We also experimented with the English WaCky corpus (Baroni et al., 2009) which contains nearly 2 billion words automatically retrieved from the web. However, results were not as good as when the BNC was used.

Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 353–356,
Los Angeles, California, June 2010.

2.2 WSD System

We use a WSD system that has been shown to perform well when evaluated against ambiguities found in both general text and the biomedical domain (Stevenson et al., 2008b). Medical Subject Headings (MeSH), a controlled vocabulary used for document indexing, are obtained from PubMed and used as additional features for the NLM-WSD data set since they have been shown to improve performance. The features are combined using the Vector Space Model, a simple memory-based learning algorithm.

2.3 Experiment

Experiments were carried out comparing performance when the WSD system was trained using either the examples in the original data set (**original**), the examples generated from these using the relevance feedback approach (**additional**) or a combination of these (**combined**). The Senseval-3 and SemEval corpora are split into training and test portions so the training portion is used as the original data set and the WSD system evaluated against the held-back data. As there is no such recognised standard split for the NLM-WSD corpus, 10-fold cross-validation was used. For each fold the training portion is used as the original data set and automatically generated examples created by examining just that part of the data. Evaluation is carried out against the fold's test data and the average result across the 10 folds reported.

Table 1 shows the results of this experiment.[2] Examples generated using the relevance feedback approach only improve results for one data set, the NLM-WSD corpus. In this case there is a significant improvement (Mann-Whitney, $p < 0.01$) when the original and automatically generated examples are combined. There is no such improvement for the other two data sets: WSD results using the additional data are noticeably worse than when the original data is used alone and, although performance improves when these examples are combined with the original data, results are still lower than using the original data. When examples are combined there is a drop in performance of 1.2% and 2.9% for SemEval and Senseval-3 respectively.

Corpus	Original	Additional	Combined
NLM-WSD	87.9	87.6	89.2
SemEval	83.7	74.6	82.5
Senseval-3	68.8	56.3	65.9

Table 1: Results of relevance feedback approach applied to three data sets

These results indicate that the relevance feedback approach described by Stevenson et al. (2008a) is not able to generate useful examples for the Senseval-3 and SemEval data sets, although it can for the NLM-WSD data set. We hypothesise that these corpora contain different levels of ambiguity which effect suitability of the approach.

3 Analysis of Ambiguities

The three data sets are compared using measures designed to determine the level of ambiguity they contain. Section 3.1 reports results using various widely used measures based on the distribution of senses. Section 3.2 introduces a measure based on the semantic similarity between the possible senses of ambiguous terms.

3.1 Sense Distributions

Three measures for characterising the difficulty of WSD data sets based on their sense distribution were used. The first is the widely applied most frequent sense (MFS) baseline (McCarthy et al., 2004), i.e. the proportion of examples for an ambiguous term that are labeled with the commonest sense. The second is number of senses per ambiguous term. The final measure, the entropy of the sense distribution, has been shown to be a good indication of disambiguation difficulty (Kilgarriff and Rosenzweig, 2000). For two of these measures (number of senses and entropy) a higher figure indicates greater ambiguity while for the MFS measure a lower figure indicates a more difficult data set.

Table 2 shows the results of computing these measures averaged across all terms in the corpus. For two measures (number of senses and entropy) the NLM-WSD corpus is least ambiguous, Senseval-3 the most ambiguous with SemEval between them. The MFS scores are very similar for two data sets (NLM-WSD and SemEval), both of which are much higher than for Senseval-3.

[2] Results reported here for the NLM-WSD corpus are slightly different from those reported by (Stevenson et al., 2008a). We used an additional feature (MeSH headings), which improved the baseline performance, and more query terms which improved the quality of the additional examples for all three data sets.

These measures suggest that the NLM-WSD corpus is less ambiguous than the other two and also that the Senseval-3 corpus is the most ambiguous of the three.

Corpus	MFS	Senses	Entropy
NLM-WSD	78.0	2.63	0.73
SemEval	78.4	3.60	0.91
Senseval-3	53.8	6.43	1.75

Table 2: Properties of Data Sets using sense distribution measures

3.2 Semantic Similarity

We also developed a measure that takes into account the similarity in meaning between the possible senses for an ambiguous term. This measure is similar to the one used by Passoneau et al. (2009) to analyse levels of inter-annotator agreement in word sense annotation. Our measure is shown in equation 1 where $Senses$ is the set of possible senses for an ambiguous term, $|Senses| = n$ and $\binom{Senses}{2}$ is the set of all subsets of $Senses$ containing two of its members (i.e the set of unordered pairs). The similarity between a pair of senses, $sim(x, y)$, can be computed using any lexical similarity measure, see Pedersen et al. (2004). Essentially this measure computes the mean of the similarities between each pair of senses for the term.

$$sim\_measure = \frac{\sum_{\{x,y\}\epsilon\binom{Senses}{2}} sim(x, y)}{\binom{n}{2}} \quad (1)$$

One problem with comparing the data sets used here is that they use a range of sense inventories. Although lexical similarity measures have been applied to WordNet (Pedersen et al., 2004) and UMLS (Pedersen et al., 2007), it is not clear that the scores they produce can be meaningfully compared. To avoid this problem we mapped the sense inventories onto a single resource: WordNet version 3.0.

The mapping was most straightforward for Senseval-3 which uses WordNet 1.7.1 and could be automatically mapped onto WordNet 3.0 senses using publicly available mappings (Daudé et al., 2000). The SemEval data contains a mapping from the OntoNotes senses to groups of WordNet 2.1 senses. The first sense from this group was mapped to WordNet 3.0 using the same mappings.

Mapping the NLM-WSD corpus was more problematic and had to be carried out manually by comparing sense definitions in UMLS and WordNet 3.0. We had expected this process to be difficult but found clear mappings for the majority of senses. There were even found cases in which the sense definitions were identical in both resources. (The most likely reason for this is that some of the resources that are included in the UMLS were used to compile WordNet.) Another, more serious, problem is related to the annotation scheme used in the NLM-WSD corpus. If none of the possible senses in UMLS were judged to be appropriate the annotators could label the sense as "None". We did not map these senses since it would require examining each instance to determine the most appropriate sense or senses in WordNet and we expected this to be error prone. In addition, there is no guarantee that all of the instances of a particular term labeled with "None" refer to the same meaning. All of the "None" senses were removed from the NLM-WSD data set and any terms where there were more than ten instances marked as "None" were also rejected from the similarity analysis. This allowed us to compute the similarity score for just 20 examples (40% of the total) although we felt that this was a large enough sample to provide insight into the data set.

The `WordNet::Similarity` package (Pedersen et al., 2004) was used to compute similarity scores. Results are reported for three of the measures in this package. (Other measures produced similar results.) The simple **path** measure computes the similarity between a pair of nodes in WordNet as the reciprocal of the number of edges in the shortest path between them, the **LCh** measure (Leacock et al., 1998) also uses information about the length of the shortest path between a pair of nodes and combines this with information about the maximum depth in WordNet and the **JCn** measure (Jaing and Conrath, 1997) makes use of information theory to assign probabilities to each of the nodes in the WordNet hierarchy and computes similarity based on these scores.

Table 3 shows the values of equation 1 for the three similarity measures with scores averaged across terms. These results indicate that for all measures the Senseval-3 data set contains the most ambiguity and NLM-WSD the least. This analysis is consistent with the one carried out using measures based on sense distributions (Section 3.1)

Corpus	Measure		
	Path	JCn	LCh
NLM-WSD	0.074	0.032	1.027
SemEval	0.136	0.061	1.292
Senseval-3	0.159	0.063	1.500

Table 3: Semantic similarity for each data set using a variety of measures

and suggest that the senses in the NLM-WSD data set are more clearly distinguished than the other two.

4 Conclusion

This paper has explored a semi-supervised approach to the generation of labeled training data for WSD that is based on relevance feedback (Stevenson et al., 2008a). It was tested on three data sets but was only found to generate examples that were accurate enough to improve WSD performance for one of these. The data set in which a performance improvement was observed represented a limited domain (biomedicine) while the other two were not restricted in this way. Measures designed to quantify the level of ambiguity were applied to these data sets including ones based on the distribution of senses and another designed to quantify similarities between senses. These measures provided evidence that the corpus for which the relevance feedback approach was successful contained less ambiguity than the other two and this suggests that the relevance feedback approach is most appropriate when the level of ambiguity is low.

The experiments described in this paper highlight the importance of the level of ambiguity on the relevance feedback approach's ability to generate useful labeled examples. Since it is semi-supervised the ambiguity level can be checked using the measures used in this paper (Section 3) and the performance of any automatically generated examples can be compared with the manually labeled ones (see Section 2.3) before deciding whether or not they should be applied.

References

M. Baroni, S. Bernardini, A. Ferraresi, and E. Zanchetta. 2009. The wacky wide web: a collection of very large linguistically processed web-crawled corpora. *Language Resources and Evaluation*, 43(3):209–226.

J. Daudé, L. Padró, and G. Rigau. 2000. Mapping wordnets using structural information. In *Proceedings of ACL '00*, pages 504–511, Hong Kong.

J. Jaing and D. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proceedings of International Conference on Research in Computational Linguistics*, Taiwan.

A. Kilgarriff and J. Rosenzweig. 2000. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2):15–48.

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

D. McCarthy, R. Koeling, J. Weeds, and J. Carroll. 2004. Finding predominant word senses in untagged text. In *Proceedings of ACL'04*, pages 279–286, Barcelona, Spain.

R. Mihalcea, T. Chklovski, and A. Kilgarriff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of Senseval-3*, pages 25–28, Barcelona, Spain.

R. Passoneau, A. Salleb-Aouissi, and N. Ide. 2009. Making sense of word sense variation. In *Proceedings of SEW-2009*, pages 2–9, Boulder, Colorado.

T. Pedersen, S. Patwardhan, and Michelizzi. 2004. Wordnet::similarity - measuring the relatedness of concepts. In *Proceedings of AAAI-04*, pages 1024–1025, San Jose, CA.

T. Pedersen, S. Pakhomov, S. Patwardhan, and C. Chute. 2007. Measures of semantic similarity and relateness in the biomedical domain. *Journal of Biomedical Informatics*, 40(3):288–299.

S. Pradhan, E. Loper, D. Dligach, and M. Palmer. 2007. SemEval-2007 Task-17: English Lexical Sample, SRL and All Words. In *Proceedings of SemEval-2007*, pages 87–92, Prague, Czech Republic.

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.

M. Stevenson, Y. Guo, and R. Gaizauskas. 2008a. Acquiring Sense Tagged Examples using Relevance Feedback. In *Proceedings of the Coling 2008*, pages 809–816, Manchester, UK, August.

M. Stevenson, Y. Guo, R. Gaizauskas, and D. Martinez. 2008b. 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.