Sussx: WSD using Automatically Acquired Predominant Senses

Rob Koeling and Diana McCarthy
Department of Informatics
University of Sussex
Brighton BN1 9QJ, UK
robk,dianam@sussex.ac.uk

1 Introduction

We introduced a method for discovering the predominant sense of words automatically using raw (unlabelled) text in (McCarthy et al., 2004) and participated with this system in SENSEVAL3. Since then, we worked on further developing ideas to improve upon the base method. In the current paper we target two areas where we believe there is potential for improvement. In the first one we address the fine-grained structure of WordNet's (WN) sense inventory (i.e. the topic of the task in this particular track). The second issue we address here, deals with topic domain specilisation of the base method.

Error analysis tought us that the method is sensitive to the fine-grained nature of WN. When two distinct senses in the WN sense inventory are closely related, the method often has difficulties discriminating between the two senses. If, for example, sense 1 and sense 7 for a word are closely related, choosing sense 7 in stead of sense 1 has serious consequences if you are using a first-sense heuristic (considering the highly skewed distribution of word senses). We expect that applying our method on a coarser grained sense inventory might help us resolve some of the more unfortunate errors.

(Magnini et al., 2002) have shown that information about the domain of a document is very useful for WSD. This is because many concepts are specific to particular domains, and for many words their most likely meaning in context is strongly correlated to the domain of the document they appear in. Thus, since word sense distributions are skewed and depend on the domain at hand we would like to explore if we can estimate the most likely sense of a word *for each domain of application* and exploit this in a WSD system.

2 Predominant Sense Acquisition

We use the method described in (McCarthy et al., 2004) for finding predominant senses from raw text. The method uses a thesaurus obtained from the text by parsing, extracting grammatical relations and then listing each word (w) with its top k nearest neighbours, where k is a constant. Like (McCarthy et al., 2004) we use $k = 50$ and obtain our thesaurus using the distributional similarity metric described by (Lin, 1998) and we use WordNet (WN) as our sense inventory. The senses of a word w are each assigned a ranking score which sums over the distributional similarity scores of the neighbours and weights each neighbour's score by a WN Similarity score (Patwardhan and Pedersen, 2003) between the sense of w and the sense of the neighbour that maximises the WN Similarity score. This weight is normalised by the sum of such WN similarity scores between all senses of w and the senses of the neighbour that maximises this score. We use the WN Similarity **jcn** score (Jiang and Conrath, 1997) since this gave reasonable results for (McCarthy et al., 2004) and it is efficient at run time given precompilation of frequency information. The **jcn** measure needs word frequency information, which we obtained from the British National Corpus (BNC) (Leech, 1992). The distributional thesaurus was constructed using subject, direct object adjective modifier and noun modifier relations.

Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), pages 314–317,
Prague, June 2007.

3 Coarse Sense Inventory Adaptation

We contrasted ranking of the original WordNet senses with ranking produced using the coarse grained mapping between WordNet senses and the clusters provided for this task. In the first, which we refer to as fine-grained training (SUSSX-FR), we use the original method as described in section 2 using WordNet 2.1 as our sense inventory. For the second method which we refer to as coarse-grained training (SUSSX-CR), we use the clusters of the target word as our senses. The distributional similarity of each neighbour is apportioned to these clusters using the maximum WordNet similarity between any of the WordNet senses in the cluster and any of the senses of the neighbour. This WordNet similarity is normalised as in the original method, but for the denominator we use the sum of the WordNet similarity scores between this neighbour and all clusters of the target word.

4 Domain Adaptation

The topic domain of a document has a strong influence on the sense distribution of words. Unfortunately, it is not feasible to produce large manually sense-annotated corpora for every domain of interest. Since the method described in section 2 works with raw text, we can specialize our sense rankings for a particular topic domain, simply by feeding a domain specific corpus to the algorithm. Previous experiments have shown that unsupervised estimation of the predominant sense of certain words using corpora whose domain has been determined by hand outperforms estimates based on domain-independent text for a subset of words and even outperforms the estimates based on counting occurrences in an annotated corpus (Koeling et al., 2005). A later experiment (using SENSEVAL2 and 3 data) showed that using domain specific predominant senses can slightly improve the results for some domains (Koeling et al., 2007). However, a firm idea of when domain specilisation should be considered could not (yet) be given.

4.1 Creating the Domain Corpora

In order to estimate topic domain specific sense rankings, we need to specify what we consider 'domains' and we need to collect corpora of texts for these domains. We decided to use text classification for determining the topic domain and adopted the domain hierarchy as defined for the topic domain extension for WN (Subject Field Codes or WordNet Domains (WN-DOMAINS) (Magnini et al., 2002)).

Domains In WN-DOMAINS the Princeton English WordNet is augmented with domain labels. Every synset in WN's sense inventory is annotated with at least one domain label, selected from a set of about 200 labels hierarchically organized (based on the Dewey Decimal Classification (Diekema,)). Each synset of Wordnet 1.6 was labeled with one or more labels. The label 'factotum' was assigned if any other was inadequate. The first level consists of 5 main categories (e.g. 'doctrines' and 'social_science') and 'factotum'. 'doctrines', for example, has subcategories such as 'art', 'religion' and 'psychology'. Some subcategories are divided in sub-subcategories, e.g. 'dance', 'music' or 'theatre' are subcategories of 'art'.

Classifier We extracted bags of domain-specific words from WordNet for all the defined domains by collecting all the word senses (synsets) and corresponding glosses associated with a certain domain label. These bags of words define the domains and we used them to train a Support Vector Machine (SVM) text classifier using 'TwentyOne'[1].

The classifier distinguishes between 48 classes (first and second level of the WN-DOMAINS hierarchy). When a document is evaluated by the classifier, it returns a list of all the classes (domains) it recognizes and an associated *confidence score* reflecting the certainty that the document belongs to that particular domain.

Corpora We used the Gigaword English Corpus as our data source. This corpus is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium, at the University of Pennsylvania. For the experiments described in this paper, we use the first 20 months worth of data of all four sources (Agence France Press English Service, Associated Press Worldstream English Service, The New York Times Newswire Service and The Xinhua News Agency English Service). There are 4 different types

[1] TwentyOne Classifier is an Irion Technologies product: www.irion.ml/products/english/products_classify.html

Doc.Id.	Class	Conf. Score
d001	Medicine (Economy)	0.75 (0.75)
d002	Economy (Politics)	0.76 (0.74)
d003	Transport (Biology)	0.75 (0.68)
d004	Comp-Sci (Architecture)	0.81 (0.68)
d005	Psychology (Art)	0.78 (0.74)

Table 1: Output of the classifier for the 5 documents. The classifiers second choice is given between brackets.

of documents identified in the corpus. The vast majority of the documents are of type 'story'. We are using all the data.

The five documents were fed to the classifier. The results are given in table 1. Unfortunately, only one document (d004) was considered to be a clear-cut example of a particular domain by the classifier (i.e. a high score is given to the first class and a much lower score to the following classes).

4.2 Domain rankings

We created domain corpora by feeding the GigaWord documents to the classifier and adding each document to the domain corpus corresponding to the classifier's first choice. The five corpora we needed for these documents were parsed using RASP (Briscoe and Carroll, 2002) and the resulting grammatical relations were used to create a distributional similarity thesaurus, which in turn was used for computing the predominant senses (see section 2). The only pre-processing we performed was stripping the XML codes from the documents. No other filtering was undertaken. This resulted in five sets of sense inventories with domain-dependent sense rankings. Each of them has a slightly different set of words. The words they have in common do have the same senses, but not necessarily the same estimated most frequently used sense.

5 Results from Semeval

Coarse Disambiguation of coarse-grained senses is obviously an easier task than fine grained training. We had hoped that the coarse-grained training might show superior performance by removing the noise created by related but less frequent senses. Since the mapping between fine-grained senses and clusters is used anyway in the scorer the noise from related senses does not seem to be an issue. Related senses are scored correctly. Indeed the performance of the fine-grained training is superior to that of the coarse-grained training. We believe this is because predominant meanings have more related senses. There are therefore more chances that the distributional similarity of the neighbours will get apportioned to one of the related senses when there are more related senses. The coarse grained ranking would have an advantage on occasions when in the fine-grained ranking the credit between related senses is split and an unrelated sense ends up with a higher ranking score. Since the coarse-grained ranker lumps the credit for related sense together it would be at an advantage. Clearly this doesn't happen enough in the data to outweigh the beneficial effect of the number of related senses compensating for other noise in the data.

Doc.Id.	Class	SUSSX-FR	SUSSX-C-WD
d001	Medicine	0.556	0.560
d002	Economy	0.508	0.515
d003	Transport	0.487	0.454
d004	Comp-Sci	0.407	0.424
d005	Psychology	0.356	0.372

Table 2: Impact of domain specialisation for each of the five documents (F_1 scores).

Domain Unfortuately, the system specialised for domain (SUSSX-C-WD) did not improve the results over the 5 documents significantly. However, if we look at the contributions made by each document, we might learn something about the relation beteen the output of the classifier and the impact on the WSDresults. Table 2 shows the per-document results for the systems SUSSX-FR and SUSSX-C-WD. The first two documents show very little difference with the domain independent results. The documents 'd004' and 'd005' show a small but clear improved performance for the domain results. Unfortunately, document 'd003' displays a very disappointing drop of more than 3% in performance, and cancels out all the gains made by the last two documents.

The output of the classifier seems to be indicative of the results for all documents except 'd003'. The classifier doesn't seem to find enough evidence for a marked preference for a particular domain

for documents 'd001' and 'd002'. This could be an indication that there is no strong domain effect to be expected. The strong preference for the 'computer_science' domain for 'd004' is reflected in good performance of SUSSX-C-WD and even though the confidence scores for the first 2 alternatives of 'd005' are fairly close, there is a clear drop in confidence score for the third alternative, which might indicate that the topic of this document is related to both first choices of the classifier. It will be interesting to evaluate the results for 'd005' using the 'Art' sense rankings. One would expect those results to be similar to the results found here. Finally, the results for 'd003' are hard to explain. We will need to do an extensive error analysis as soon as the gold-standard is available.

6 Conclusions

In this paper we investigated two directions where we expect potential for improving the performance of our method for acquiring predominant senses. In order to fully appreciate what the effects of the coarse grained sense inventory are (i.e. whether some of the more unfortunate errors are resolved), we will have to do an extensive error analysis as soon as the gold standard becomes available. Considering the fairly low number of attempted tokens (only 72.8% of the tokens are attempted), we are at a disadvantage compared to systems that back-off to (for example) the first sense in WN. However, we are well pleased with the high precision (71.7%) of the method SUSSX-FR, considering this is a completely unsupervised method. There seems to be potential gains for domain adaptation, but applying it to each document does not seem to be advisable. More research needs to be done to identify in which cases a performance boost can be expected. Five documents is not enough to fully investigate the matter. At the moment we are performing a larger scale experiment with the documents in SemCor. These documents seem to cover a fairly wide range of domains (according to our text classifier) and many domains are represented by several documents.

Acknowledgements

This work was funded by UK EPSRC project EP/C537262 "Ranking Word Senses for Disambiguation: Models and Applications", and by a UK Royal Society Dorothy Hodgkin Fellowship to the second author. We would also like to thank Piek Vossen for giving us access to the Irion Technologies text categoriser.

References

Ted Briscoe and John Carroll. 2002. Robust accurate statistical annotation of general text. In *Proceedings of LREC-2002*, pages 1499–1504, Las Palmas de Gran Canaria.

Anne Diekema. http://www.oclc.org/dewey/.

Jay Jiang and David Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *International Conference on Research in Computational Linguistics*, Taiwan.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *Proceedings of the Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing.*, pages 419–426, Vancouver, Canada.

Rob Koeling, Diana McCarthy, and John Carroll. 2007. Text categorization for improved priors of word meaning. In *Proceedings of the Eighth International Conference on Intelligent Text Processing and Computational Linguistics (Cicling 2007)*, pages 241–252, Mexico City, Mexico.

Geoffrey Leech. 1992. 100 million words of English: the British National Corpus. *Language Research*, 28(1):1–13.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of COLING-ACL 98*, Montreal, Canada.

Bernardo Magnini, Carlo Strapparava, Giovanni Pezzulo, and Alfio Gliozzo. 2002. The role of domain information in word sense disambiguation. *Natural Language Engineering*, 8(4):359–373.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant senses in untagged text. In *Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287, Barcelona, Spain.

Siddharth Patwardhan and Ted Pedersen. 2003. The cpan wordnet::similarity package. http://search.cpan.org/šid/WordNet-Similarity/.