Text categorization for improved priors of word meaning

Rob Koeling & Diana McCarthy & John Carroll
{*robk,dianam,johnca*}*@sussex.ac.uk*

Department of Informatics, University of Sussex, Brighton BN1 9QH, UK

Abstract. Distributions of the senses of words are often highly skewed. This fact is exploited by word sense disambiguation (WSD) systems which back off to the predominant (most frequent) sense of a word when contextual clues are not strong enough. 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. 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.
In this paper we address the question of whether we can automatically produce domain-specific corpora which could be used to acquire predominant senses appropriate for specific domains. We collect the corpora by automatically classifying documents from a very large corpus of newswire text. Using these corpora we estimate the predominant sense of words for each domain. We first compare with the results presented in [1]. Encouraged by the results we start exploring using text categorization for WSD by evaluating on a standard data set (documents from the SENSEVAL-2 and 3 English all-word tasks). We show that for these documents and using domain-specific predominant senses, we are able to improve on the results that we obtained with predominant senses estimated using general, non domain-specific text. We also show that the confidence of the text classifier is a good indication whether it is worthwhile using the domain-specific predominant sense or not.

1 Introduction

The fact that the distributions of word senses are often highly skewed is recognized by the word sense disambiguation (WSD) community and is often successfully exploited in WSD systems. The sense distributions can either be used as a prior in a system that collects statistical evidence from the local context of the contested word to determine the intended sense of the word, or it can be used as a back-off in those cases where the local context does not provide enough information to decide. However, manually tagging corpora with word senses is labour intensive and therefore expensive. Therefore, most researchers use the same publicly available resource, SemCor [2], to estimate word sense distributions. Despite the fact that SemCor is a fairly small corpus, it covers a reasonable range of words (and word senses) in sufficient frequencies. In WSD, the heuristic of just choosing the most frequent sense of a word is very powerful, especially for words with highly skewed sense distributions [3]. Indeed, only 5 out of the 26 systems in the recent SENSEVAL-3 English all words task [4] outperformed the heuristic of choosing the most frequent sense as derived from SemCor (which would give 61.5% precision and recall[1]). Furthermore, systems that did outperform the first sense heuristic did so only by a small margin (the top score being 65% precision and recall).

[5] 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 know *for each domain of application* the most likely sense of a word.

There are, however, several problems with obtaining hand-labelled domain-specific sense-tagged data. The first being the problem of specifying the domains. There is no such thing as a standardized definition of topical domains. The definition of a domain will be dependent on user and application. People will most likely disagree on what should be considered domains, where the borders between domains lie and finally the granularity of the domain definitions. The second problem is that even if people agreed on a domain definition, producing domain-specific sense-tagged corpora would be extremely costly, since a substantial corpus would have to be annotated by hand for every domain of interest. It would be ideal if a user could specify a topical domain, collect a substantial amount of text relevant for that domain and use that corpus for estimating domain-specific sense distributions.

In response to the second problem, we proposed a method for *automatically* inducing the predominant sense of a word from raw text [6]. The method was extensively tested on domain-neutral data and we carried out a limited test of our method on text in 2 domains to assess whether the acquired predominant sense information was broadly consistent with the domain of the text it was acquired from. In a later paper, [1], we evaluated the method on domain-specific text. In order to do this, we created a sense-annotated gold-standard for a sample of words covering 2 domains (Finance and Sport) and domain-neutral data. We showed that unsupervised estimation of the predominant sense of certain words using corpora whose topical domain has been determined by hand outperforms estimates based on domain-independent text for a sample of words and even outperforms the estimates based on counting occurrences in SemCor.

However, these results were obtained using data where the domain of the documents was determined by hand. High quality high volume domain specific corpora are not always available for a given language and a given domain. In this paper we want to address some of the questions that arose from this earlier

[1] This figure is the mean of two different estimates [4], the difference being due to multiword handling.

work. Will our method [6] be robust enough to deal with the noise that is unavoidable if you use automatically classified text? We show, [1], that the method successfully deals with a sample of words in a domain-specific setting, however, for some applications word sense disambiguation may be required for all the words in a given text. In this paper we describe the automatic construction of domain-specific text corpora using a big newswire corpus and a text classifier. We estimate the predominant senses for all polysemous nouns (as defined in WordNet) for a number of domains. We evaluate the estimated predominant senses by 1) comparing the results with the results based on hand-classified text as presented in [1] and 2) performing a WSD task on the documents used in the SENSEVAL-2 and 3 English all-words tasks. We show that our results are very comparable with [1] and, in certain cases the domain-specific predominant sense estimates outperform those based on a domain-neutral corpus. We will look at the effect the classifier has on the success and also what the influence of corpus size is.

2 Finding Predominant Senses

We use the method described in [6] 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 [6] we use $k = 50$ and obtain our thesaurus using the distributional similarity metric described by [7] 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 [8] 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 [9] since this gave reasonable results for [6] 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) [10]. The distributional thesaurus was constructed using subject, direct object adjective modifier and noun modifier relations.

3 Creating the domain corpora

3.1 The GigaWord Corpus

The GigaWord English Corpus is a comprehensive archive of newswire text data that has been acquired over several years by the Linguistic Data Consortium (LDC), at the University of Pennsylvania. The data is collected from four different sources: Agence France Press English Service, Associated Press Worldstream English Service, The New York Times Newswire Service and The Xinhua News Agency English Service. The data is roughly from the years 1994 until 2002 (not every source starts and stops in the same month). The total number of documents is 4,111,240, consisting of 1,756,504 K-words. For the experiments described in this paper, we use the first 20 months worth of data of all 4 sources. There are 4 different types of documents identified in the corpus. The vast majority of the documents are of type 'story'. We are using all the data.

3.2 The classifier

For the text classification, we adopt a previous definition of topical domains, though this could be changed in future. Since our evaluation framework and the method [6] use WN as a sense inventory, we make use of a topic domain extension for WN (WN-DOMAINS[5]). In WN-DOMAINS the Princeton English WordNet is augmented with some domain labels. Every synset in WN's sense inventory is annotated with at least one domain label, selected from a set of about 200 hierarchically organized labels. Each synsets 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' 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'. 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 are the blueprints for the domains and we used them to train a Support Vector Machine (SVM) text classifier using 'TwentyOne'[2]. 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.Selected lines of the output of the classifier are given in figure 1.

```
<CLASSIFY_SERVER>
  <N_RESULTS>48</N_RESULTS>
    <RESULT>
      <CLASS><![CDATA[medicine]]></CLASS>
      <CONF_SCORE>0.85</CONF_SCORE>
    </RESULT>
    <RESULT>
      <CLASS><![CDATA[biology]]></CLASS>
      <CONF_SCORE>0.80</CONF_SCORE>
    </RESULT>
    .....
    <RESULT>
      <CLASS><![CDATA[artisanship]]></CLASS>
      <CONF_SCORE>0.03</CONF_SCORE>
    </RESULT>
</CLASSIFY_SERVER>
```

Fig. 1. Part of the output of the 'TwentyOne' classifier.

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

3.3 The domain corpora

The 20 months worth of GigaWord corpus consists of 520501 files. Out of the 48 predefined classes, 44 are are represented in the classifier output (meaning that at least one document was classified as most likely belonging to that class). The distribution of documents is, as was to be expected, very uneven. Table 1 gives an overview of the number of documents per domain.

Number of documents in domain	Number of domains
<500	18
500 - 1000	4
1000 - 5000	6
5000 - 10000	3
>10000	13

Table 1. Distribution of documents over domains

Given the fact that we used general newswire data, it was a pleasant surprise to see so many domains well represented in the corpus. At the moment we assign a domain label to a document by simply taking the domain with the highest confidence value (the level of confidence is not considered at the moment). However, manual analysis suggests there seems to be a good case for taking the confidence level into consideration. Manual inspection of randomly selected documents suggested that documents that were assigned a confidence level under 0.74 were often assigned the wrong domain. At 0.75 the amount of noise seems to be fairly low, only to be further improved by increasing the confidence level. Evidently, the drawback of putting up a higher confidence threshold is losing data. Putting the threshold at 0.75 for the first document reduces the number of documents by some 23%. A first test using a threshold (set at 0.75) for corpus collection did not improve the results. Therefore, in the experiments in this paper we use all the data available. More experiments will be needed to explore this matter further. For the evaluation we use 6 documents from the Senseval-2 and 3 English all-words tasks (see 4). The classifier assigned the domains 'art', 'medicine' and 'psychology' to the Senseval-2 documents and 'politics' and 2 times 'psychology' to the Senseval-3 documents. The characteristics of the 4 relevant domain corpora are given in table 2.

Domain	No. of documents	No. of words
Art	11679	5729655
Medicine	14463	5644181
Psychology	44075	23748013
Politics	64106	25108055

Table 2. Size of the domain corpora

3.4 Domain rankings

The 4 resulting corpora were parsed using RASP [11] and the resulting grammatical relations were used to create a distributional similarity thesaurus, which in turn was used for computing the predominant senses (see 2). The only preprocessing we performed was stripping the XML codes from the documents. No other filtering was undertaken. This resulted in 4 sets of domain-dependent sense inventories. 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.

4 Experiments and Evaluation

To evaluate the first sense heuristic we see how the heuristic performs on a WSD task. This simply uses the skew of the data to tag every word type with one sense. In a real application, this back off heuristic should be combined with contextual WSD information. The naive WSD system evaluation approach is a very useful one. First of all, a WSD system using only the first sense heuristic where the predominant sense is estimated using hand-annotated text is a more than decent performing participant in WSD competitions. And second, [6] have shown that unsupervised estimation of predominant senses using domain-neutral text is a good approximation of the supervised alternative. We show that if you know what topic domain you are in, you can do better with domain-specific predominant senses than with domain-neutral ones and in certain cases you might even do better than when using hand-crafted domain-neutral sense distributions. In this paper, following [1], we concentrate on the evaluation of nouns, but extending our experiment from evaluating a selected set of nouns to an all-words (nouns) task. The first experiment we perform is to take the domain rankings for the Sport and Finance domain and evaluate the predominant senses on the test data used in [1]. In the second experiment we use the senseval-2 and 3 data sets as these are standard all-words datasets available for English where automatic methods can be contrasted with information from the manually produced SemCor.

4.1 Hand-labelled versus automatically classified

The first experiment is a straightforward comparison with the results reported in ([1]). The Sports and Finance corpora are collected as described in section 3. The results reported on here are based on using 20 months of GigaWord data. The resulting Sports corpus consists of 23.6M words and the Finance corpus of 48.2M words. The main aim for this experiment is to see if the good results reported in [1] can be reproduced with automatically labelled data. Table 3 presents the best results for this experiment. It shows a small (and expected) decrease in accuracy for the Finance test set and a small (surprising) increase for the Sport test set. These results are very encouraging. Despite the decrease in precision on the Finance test set, both the BNC and the SemCor results are outperformed for both test sets.

Finance	Train	WSD Precision	Sport	Train	WSD Precision
	BNC	43.3		BNC	33.2
	SemCor	35.0		SemCor	16.8
	Reuters Finance	49.9		Reuters Sport	43.7
	GigaWord Finance	44.2		GigaWord Sport	46.1

Table 3. WSD using predominant sens: hand-labelled (Reuters) versus automatically classified (GigaWord).

4.2 Senseval

The purpose of Senseval is to evaluate the strengths and weaknesses of programs that can automatically determine the sense of a word in context with respect to different words, different varieties of language, and different languages. In order to do so, a number of tasks has been set up. One of the tasks is an "all-words" task. In this task every ambiguous (according to a chosen sense inventory) word-token in a text is manually annotated with the correct sense in the context where it occurred. The predicted word senses by participants are compared to the manually annotated gold-standard. Both the Senseval-2 and 3 competitions had an English all-words task defined. Three documents were prepared for each edition. This total of 6 documents is what we use for evaluation.
We sent the 6 documents to the classifier to determine the topical domains. The results are given in table 4. The classifier's first and second (between brackets) guesses are given in this table with corresponding confidence scores in the third column. The first document has a low confidence score. The document is hard to classify, even by hand. The classifier's second and third guesses ('religion' and 'architecture') are actually equally plausible. The second document is spot on (and the classifier is confident about it). The third one would manually probably be classified as 'pedagogy', but 'psychology' is plausible. The fourth document is apparently taken from a novel. This seems to confuse the classifier, which is confident that 'psychology' is the domain. The fifth document is spot on (and again, with high confidence value). The last document is a hard to classify human interest story about the aftermath of an earthquake. The classifier's first 2 guesses are relevant, but have low confidence score.

Doc.Id.	Class	Confidence Score
Se2-d00	Art (Architecture)	0.73 (0.71)
Se2-d01	Medicine (Biology)	0.85 (0.80)
Se2-d02	Psychology (Economy)	0.79 (0.72)
Se3-d000	Psychology (Economy)	0.81 (0.72)
Se3-d001	Politics (Law)	0.82 (0.77)
Se3-d002	Psychology (Earth)	0.72 (0.70)

Table 4. Output of the classifier for the 6 Senseval documents.

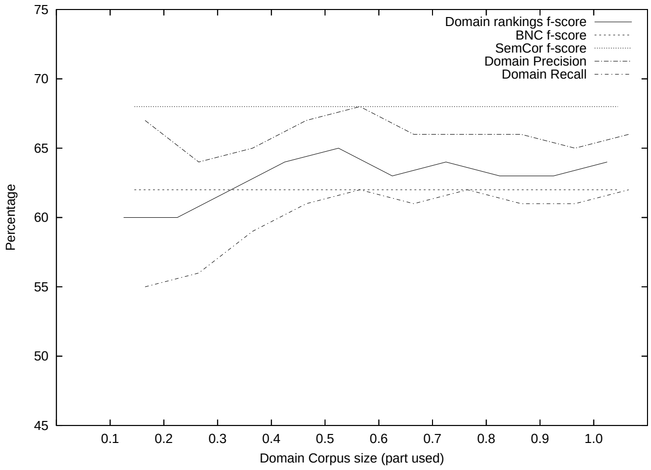

Fig. 2. Results for Senseval-2: Precision, Recall and f_1-score for varying domain corpus size (percentage of available text) versus BNC (full corpus) and SemCor

Results We produced separate results for the Senseval-2 and 3 documents because different versions of WN were used to annotate the data. The documents in Senseval-2 were annotated with WN 1.7, whilst those in Senseval-3 were annotated with WN 1.7.1. In figure 2 we show how the results develop as a function of corpus size. Different amounts of data were available for the 3 documents involved (see table 2). In this graph we want to show how the combined results (of the 3 documents) develop if you take a certain portion of the data available for each domain. We report on f_1-score[3], Precision and Recall for predominant senses estimated using increasing portions of the domain corpora and compare them with the estimated predominant senses based on the BNC (using the whole written part of the corpus; about 90M words) and the SemCor benchmark.

There are a few interesting aspects about this figure. First of all, we can see that the overall results for the domain-based are consistently better than those from the BNC. The learning curve seems to display an upward trend, although it starts to flatten out quite early on. An interesting aspect here is the fact that Precision seems to be fairly stable from the start. It is the Recall that makes the difference for the overall f-score results. A similar, though less convincing story is told in figure 3 for the Senseval-3 results. The overall results stay slightly underneath the BNC benchmark and far away from the SemCor results. Recall is going up considerably to begin with, but flattens out quite quickly and seems to remain stable from then on.

If we look at more detail at the results, we can see that the favorable Senseval-2 results are entirely due to the first and third document. The domain results for the second document starts to creep up to BNC level, but then re-

[3] $F_\beta = (1+\beta^2) * ((Prec * Recall) \ / \ (Recall + \beta^2 * Prec))$

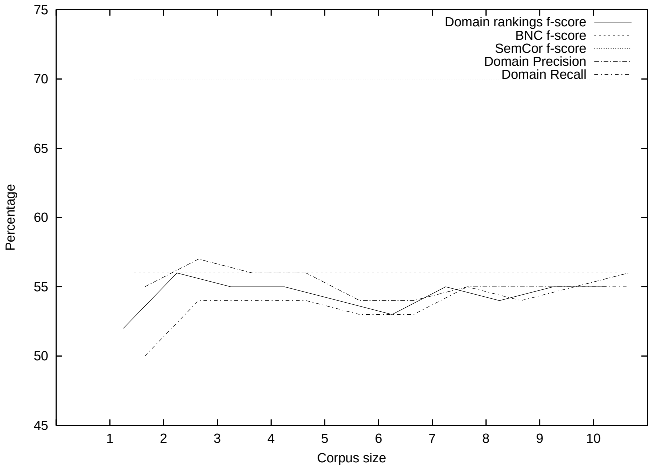

Fig. 3. Results for SENSEVAL-3: Precision, Recall and $f_1 - score$ for varying domain corpus size (percentage of available text) versus BNC (full corpus) and SemCor

mains there. A nice observation here is that fairly small corpora already produce nice results. The results start to be competitive at a corpus size of around the 2.5M words for the first and third document. Finally, not shown in this figure, is that the domain results for the second document also outperform the SemCor results (those are slightly above the BNC results). This is not the case for the other 2 documents. They stay well below the SemCor results.

Finally the results per document in the SENSEVAL-3 data is shown in figure 5. The obvious observation here is that only the second document keeps the flag flying for the domain-specific results. The domain-specific results outperforms the BNC results comfortably, albeit still below SemCor results. The 2 Psychology documents perform poorly. Although, the first one (where the classifier was fairly confident) is significantly better than the third one (where the classifier gave a very low confidence score).

4.3 Domain salient words

Words that are salient to a particular domain performed particularly well in [1]. We performed an experiment to evaluate the performance if we only consider the top 1000 salient words for each domain[4]. We trimmed the list by excluding

[4] We computed salience as a ratio of normalised document frequencies, using the formula

$$S(w,d) = \frac{N_{wd}/N_d}{N_w/N}$$

where N_{wd} is the number of documents in domain d containing the noun (lemma) w, N_d is the number of documents in domain d, N_w is the total number of documents containing the noun w and N is the total number of documents.

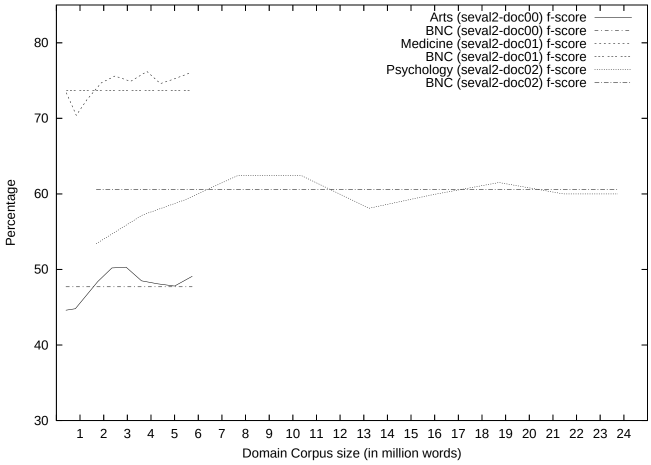

Fig. 4. Results for SENSEVAL-2: $f_1 - score$ for varying domain corpus size (in M words) versus BNC (full corpus)

any words containing capital letters and only considered words that occurred at least 10 times in the domain corpus. Just inspecting the resulting lists of salient words proved to be interesting. The lists of 'medicine' and 'politics' salient nouns indicated immediately which domain they were for. The 'art' list did that too, but also showed quite a bit of variation (e.g. music versus painting, etc). Finally the 'psychology' list was not recognizable whatsoever. It mainly consists of fairly obscure words that were not indicative of any domain in particular.
The results for this evaluation are given in table 5. The only thing we can say about the arts document is that the coverage is low. This is unsurprising, because of the fact that it is not a clear-cut arts document. It further shows that the coverage of the medicine salient word list is very high (almost half the words of the document are covered). The results for these words is very good, but equally so for BNC and SemCor. The surprising bit is that the top 1000 salient words for the psychology domain do not have *any words* in common with the psychology document. The same thing holds for the 2 psychology documents in the SENSEVAL-3 test set. The results for the politics document are outstanding: high precision, high recall, a reasonable number of words are covered and even the SemCor results are outperformed.

5 Discussion and Future research

The results show that for certain documents very good results can be obtained. The major factors that determine whether a document is a good candidate for using domain-specific sense priors seem to be:

- The classifier's confidence that the document belongs to a certain domain.

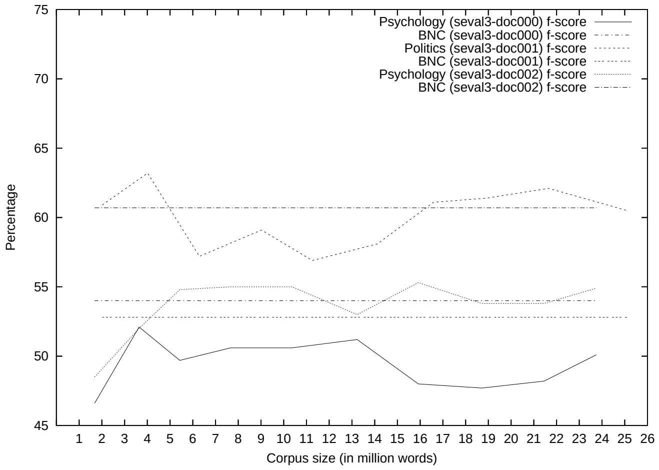

Fig. 5. Results for Senseval-3: $f_1 - score$ for varying domain corpus size (in M words) versus BNC (full corpus)

– Well defined and concise domains seem to be very helpful. Apparently, both the medical and the politics domain fit that bill. A good indication is the fact that a list of most salient words for that domain covers a reasonable size of the words in the document.

The documents classified as 'psychology' suffered from several problems. The classifier seems to be too lenient towards the psychology domain. Neither one of the 3 documents classified as 'psychology' were clear-cut examples. This might mean two things. 1) the domain is inherently too broad and we are always better of using domain-neutral sense inventories, or 2) the classifier needs a tighter definition of this domain. The latter option should be easy to explore. Only including documents with a high (where 'high' needs to be specified) confidence level can be included in the domain corpus, or we could retrain the classifier with a different, more restricted bag-of-words. A first experiment with higher confidence values for including documents in a domain corpus resulted in a significant loss of data. However, we have shown that for well-defined domains only a limited

Doc.Id.	Precision / BNC / SemCor	Recall / BNC / SemCor	No. Correct	No. Wrong	NotAttempted
Se2-d02	27.8 / 27.8 / 38.9	27.8 / 27.8 / 38.9	5	13	0
Se2-d00	87.8 / 87.8 / 88.8	87.8 / 87.8 / 86.4	194	27	0
Se2-d01	0 / 0 / 0	0 / 0 / 0	0	0	0
Se3-d000	0 / 0 / 0	0 / 0 / 0	0	0	0
Se3-d001	91.0 / 83.3 / 90.9	91.0 / 82.1 / 89.6	61	6	0
Se3-d002	0 / 0 / 0	0 / 0 / 0	0	0	0

Table 5. Evaluation results for the 1000 most salient words of each domain

amount of data is needed for good results. Certain domains (like psychology) might be more in need of tightening than others.

A testset of 6 documents is too small to draw definitive conclusions. A direct comparison with [1] taught us that we can do well with automatically created domain corpora. Even though that is a nice result, there are still many uncertainties around how and when to use the proposed technique. We don't think that it will work on every single document. One of our objectives is to find out in which conditions this technique obtains an improved prior over one obtained from, for example, a general resource (like SemCor). The results in this paper are a firm step towards a better understanding of those conditions. There is a need for more evaluation and a good possibility is to use SemCor for this task. SemCor consists of many documents from different sources. It hosts documents from many different topic domains. As soon as we have the data for all relevant domains available (parsing the documents is the bottle neck), this will be the obvious target for experiments. It will most likely give us a better understanding of the influence of domain on the results we can expect.

References

1. Koeling, R., McCarthy, D., Carroll, J.: 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., Vancouver, Canada (2005) 419–426
2. Miller, G.A., Leacock, C., Tengi, R., Bunker, R.T.: A semantic concordance. In: Proceedings of the ARPA Workshop on Human Language Technology. (1993)
3. Yarowsky, D., Florian, R.: Evaluating sense disambiguation performance across diverse parameter spaces. Natural Language Engineering **8** (2002) 293–310
4. Snyder, B., Palmer, M.: The English all-words task. In: Proceedings of SENSEVAL-3, Barcelona, Spain (2004) 41–43
5. Magnini, B., Strapparava, C., Pezzulo, G., Gliozzo, A.: The role of domain information in word sense disambiguation. Natural Language Engineering **8** (2002) 359–373
6. McCarthy, D., Koeling, R., Weeds, J., Carroll, J.: Finding predominant senses in untagged text. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, Barcelona, Spain (2004) 280–287
7. Lin, D.: Automatic retrieval and clustering of similar words. In: Proceedings of COLING-ACL 98, Montreal, Canada (1998)
8. Patwardhan, S., Pedersen, T.: The cpan wordnet::similarity package. http://search.cpan.org/s̃id/WordNet-Similarity/ (2003)
9. Jiang, J., Conrath, D.: Semantic similarity based on corpus statistics and lexical taxonomy. In: International Conference on Research in Computational Linguistics, Taiwan (1997)
10. Leech, G.: 100 million words of English: the British National Corpus. Language Research **28** (1992) 1–13
11. Briscoe, T., Carroll, J.: Robust accurate statistical annotation of general text. In: Proceedings of LREC-2002, Las Palmas de Gran Canaria (2002) 1499–1504