Word Domain Disambiguation via Word Sense Disambiguation

Antonio Sanfilippo, Stephen Tratz, Michelle Gregory

Pacific Northwest National Laboratory

Richland, WA 99352

{Antonio.Sanfilippo, Stephen.Tratz, Michelle.Gregory}@pnl.gov

Abstract

Word subject domains have been widely used to improve the performance of word sense disambiguation algorithms. However, comparatively little effort has been devoted so far to the disambiguation of word subject domains. The few existing approaches have focused on the development of algorithms specific to word domain disambiguation. In this paper we explore an alternative approach where word domain disambiguation is achieved via word sense disambiguation. Our study shows that this approach yields very strong results, suggesting that word domain disambiguation can be addressed in terms of word sense disambiguation with no need for special purpose algorithms.

1 Introduction

Word subject domains have been ubiquitously used in dictionaries to help human readers pinpoint the specific sense of a word by specifying technical usage, e.g. see "subject field codes" in Procter (1978). In computational linguistics, word subject domains have been widely used to improve the performance of machine translation systems. For example, in a review of commonly used features in automated translation, Mowatt (1999) reports that most of the machine translation systems surveyed made use of word subject domains. Word subject domains have also been used in information systems. For example, Sanfilippo (1998) describes a summarization system where subject domains provide users with useful conceptual parameters to tailor summary requests to a user's interest.

Successful usage of word domains in applications such as machine translation and summarization is strongly dependent on the ability to assign the appropriate subject domain to a word in its context. Such an assignment requires a process of Word Domain Disambiguation (WDD) because the same word can often be assigned different subject domains out of context (e.g. the word `partner` can potentially be related to FINANCE or MARRIAGE).

Interestingly enough, word subject domains have been widely used to improve the performance of Word Sense Disambiguation (WSD) algorithms (Wilks and Stevenson 1998, Magnini et al. 2001; Gliozzo et al. 2004). However, comparatively little effort has been devoted so far to the word domain disambiguation itself. The most notable exceptions are the work of Magnini and Strapparava (2000) and Suarez & Palomar (2002). Both studies propose algorithms specific to the WDD task and have focused on the disambiguation of noun domains.

In this paper we explore an alternative approach where word domain disambiguation is achieved via word sense disambiguation. Moreover, we extend the treatment of WDD to verbs and adjectives. Initial results show that this approach yield very strong results, suggesting that WDD can be addressed in terms of word sense disambiguation with no need of special purpose algorithms.

Proceedings of the Human Language Technology Conference of the North American Chapter of the ACL, pages 141–144,
New York, June 2006.

Sense	Synset and Gloss	Domains	Semcor
#1	depository financial institution, bank, banking concern, banking company (a financial institution...)	ECONOMY	20
#2	bank (sloping land...)	GEOGRAPHY, GEOLOGY	14
#3	bank (a supply or stock held in reserve...)	ECONOMY	-
#4	bank, bank building (a building...)	ARCHITECTURE, ECONOMY	-
#5	bank (an arrangement of similar objects...)	FACTOTUM	1
#6	savings bank, coin bank, money box, bank (a container...)	ECONOMY	-
#7	bank (a long ridge or pile...)	GEOGRAPHY, GEOLOGY	2
#8	bank (the funds held by a gambling house...)	ECONOMY, PLAY	
#9	bank, cant, camber (a slope in the turn of a road...)	ARCHITECTURE	-
#10	bank (a flight maneuver...)	TRANSPORT	-

Figure 1: Senses and domains for the word *bank* in WordNet Domains, with number of occurrences in SemCor, adapted from Magnini et al. (2002).

2 WDD via WSD

Our approach relies on the use of WordNet Domains (Bagnini and Cavaglià 2000) and can be outlined in the following two steps:

1. use a WordNet-based WSD algorithm to assign a sense to each word in the input text, e.g. `doctor` → `doctor#n#1`
2. use WordNet Domains to map disambiguated words into the subject domain associated with the word, e.g. `doctor#n#1`→`doctor#n#1`#MEDICINE.

2.1 WordNet Domains

WordNet Domains is an extension of WordNet (http://wordnet.princeton.edu/) where synonym sets have been annotated with one or more subject domain labels, as shown in Figure 1. Subject domains provide an interesting and useful classification which cuts across part of speech and WordNet sub-hierarchies. For example, `doctor#n#1` and `operate#n#1` both have subject domain MEDICINE, and SPORT includes both `athlete#n#1` with top hypernym `life-form#n#1` and `sport#n#1` with top hypernym `act#n#2`.

2.2 Word Sense Disambiguation

To assign a sense to each word in the input text, we used the WSD algorithm presented in Sanfilippo et al. (2006). This WSD algorithm is based on a supervised classification approach that uses SemCor[1] as training corpus. The algorithm employs the OpenNLP MaxEnt implementation of the maximum entropy classification algorithm (Berger et al. 1996) to develop word sense recognition signatures for each lemma which predicts the most likely sense for the lemma according to the context in which the lemma occurs.

Following Dang & Palmer (2005) and Kohomban & Lee (2005), Sanfilippo et al. (2006) use contextual, syntactic and semantic information to inform our verb class disambiguation system.

- Contextual information includes the verb under analysis plus three tokens found on each side of the verb, within sentence boundaries. Tokens included word as well as punctuation.
- Syntactic information includes grammatical dependencies (e.g. subject, object) and morpho-syntactic features such as part of speech, case, number and tense.
- Semantic information includes named entity types (e.g. person, location, organization) and hypernyms.

We chose this WSD algorithm as it provides some of the best published results to date, as the comparison with top performing WSD systems in Senseval3 presented in Table 1 shows---see http://www.senseval.org and Snyder & Palmer (2004) for terms of reference on Senseval3.

[1] http://www.cs.unt.edu/~rada/downloads.html.

System	Precision	Fraction of Recall
Sanfilippo et al. 2006	61%	22%
GAMBL	59.0%	21.3%
SenseLearner	56.1%	20.2%
Baseline	52.9%	19.1%

Table 1: Results for verb sense disambiguation on Senseval3 data, adapted from Sanfilippo et al. (2006).

3 Evaluation

To evaluate our WDD approach, we used both the SemCor and Senseval3 data sets. Both corpora were stripped of their sense annotations and processed with an extension of the WSD algorithm of Sanfilippo et al. (2006) to assign a WordNet sense to each noun, verb and adjective. The extension consisted in extending the training data set so as to include a selection of WordNet examples (full sentences containing a main verb) and the Open Mind Word Expert corpus (Chklovski and Mihalcea 2002).

The original hand-coded word sense annotations of the SemCor and Senseval3 corpora and the word sense annotations assigned by the WSD algorithm used in this study were mapped into subject domain annotations using WordNet Domains, as described in the opening paragraph of section 2 above. The version of the SemCor and Senseval3 corpora where subject domain annotations were generated from hand-coded word senses served as gold standard. A baseline for both corpora was obtained by assigning to each lemma the subject domain corresponding to sense 1 of the lemma.

WDD results of a tenfold cross-validation for the SemCor data set are given in Table 2. Accuracy is high across nouns, verbs and adjectives.[2] To verify the statistical significance of these results against the baseline, we used a standard proportions comparison test (see Fleiss 1981, p. 30). According to this test, the accuracy of our system is significantly better than the baseline.

The high accuracy of our WDD algorithm is corroborated by the results for the Senseval3 data set in Table 3. Such corroboration is important as the Senseval3 corpus was not part of the data set used to train the WSD algorithm which provided the basis for subject domain assignment. The standard comparison test for the Senseval3 is not as conclusive as with SemCor. This is probably due to the comparatively smaller size of the Senseval3 corpus.

	Nouns	Verbs	Adj.s	Overall
Accuracy	0.874	0.933	0.942	0.912
Baseline	0.848	0.927	0.932	0.897
p-value	4.6e-54	1.4e-07	5.5e-08	1.4e-58

Table 2: SemCor WDD results.

	Nouns	Verbs	Adj.s	Overall
Accuracy	0.797	0.908	0.888	0.848
Baseline	0.783	0.893	0.862	0.829
p-value	0.227	0.169	0.151	0.048

Table 3: Senseval3 WDD results.

4 Comparison with Previous WDD Work

Our WDD algorithm compares favorably with the approach explored in Bagnini and Strapparava (2000), who report 0.82 p/r in the WDD tasks for a subset of nouns in SemCor.

Suarez and Palomar (2002) report WDD results of 78.7% accuracy for nouns against a baseline of 68.7% accuracy for the same data set. As in the present study, Suarez and Palomar derive the baseline by assigning to each lemma the subject domain corresponding to sense 1 of the lemma. Unfortunately, a meaningful comparison with Suarez and Palomar (2002) is not possible as they use a different data set, the DSO corpus.[3] We are currently working on repeating our study with the DSO corpus and will include the results of this evaluation in the final version of the paper to achieve commensurability with the results reported by Suarez and Palomar.

5 Conclusions and Further Work

Current approaches to WDD have assumed that special purpose algorithms are needed to model the WDD task. We have shown that very competitive and perhaps unrivaled results (pending on evaluation of our WDD algorithm with the DSO corpus) can be obtained using WSD as the basis for subject domain assignment. This improvement in WDD performance can be used to

[2] We have not worked on adverbs yet, but we expect comparable results.

[3] http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC97T12.

obtain further gains in WSD accuracy, following Wilks and Stevenson (1998), Magnini et al. (2001) and Gliozzo et al. (2004). A more accurate WSD model will in turn yield yet better WDD results, as demonstrated in this paper. Consequently, further improvements in accuracy for both WSD and WDD can be expected through a bootstrapping cycle where WDD results are fed as input to the WSD process, and the resulting improved WSD model is then used to achieve better WDD results. We intend to explore this possibility in future extensions of this work.

Acknowledgements

We would like to thank Paul Whitney for help with the evaluation of the results presented in Section 3.

References

Berger, A., S. Della Pietra and V. Della Pietra (1996) A Maximum Entropy Approach to Natural Language Processing. *Computational Linguistics,* volume 22, number 1, pages 39-71.

Chklovski, T. and R. Mihalcea (2002) Building a Sense Tagged Corpus with Open Mind Word Expert. Proceedings of the ACL 2002 Workshop on "Word Sense Disambiguation: Recent Successes and Future Directions, Philadelphia, July 2002, pp. 116-122.

Dang, H. T. and M. Palmer (2005) The Role of Semantic Roles in Disambiguating Verb Senses. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, Ann Arbor MI, June 26-28, 2005.

Fleiss, J. L. (1981) *Statistical Methods for Rates and Proportions*. 2nd edition. New York: John Wiley & Sons.

Gliozzo, A., C. Strapparava, I. Dagan (2004) Unsupervised and Supervised Exploitation of Semantic Domains in Lexical Disambiguation. *Computer Speech and Language*,18(3), Pages 275-299.

Kohomban, U. and W. Lee (2005) Learning semantic classes for word sense disambiguation. In *Proceedings of the 43rd Annual meeting of the Association for Computational Linguistics*, Ann Arbor, MI.

Magnini, B., Cavaglià, G. (2000) Integrating Subject Field Codes into WordNet. *Proceedings of LREC-2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 MAY- 2 JUNE 2000, pp. 1413-1418.

Magnini, B., Strapparava C. (2000) Experiments in Word Domain Disambiguation for Parallel Texts. *Proceedings of the ACL Workshop on Word Senses and Multilinguality*, Hong-Kong, October 7, 2000, pp. 27-33

Magnini, B., C. Strapparava, G. Pezzulo and A. Gliozzo (2001) Using Domain Information for Word Sense Disambiguation. In Proceeding of SENSEVAL-2: Second International Workshop on Evaluating Word Sense Disambiguation Systems, pp. 111-114, 5-6 July 2001, Toulouse, France.

Magnini, B., C. Strapparava, G. Pezzulo and A. Gliozzo (2002) The Role of Domain Information in Word Sense Disambiguation. *Natural Language Engineering*, 8(4):359—373.

Mowatt, D. (1999) Types of Semantic Information Necessary in a Machine Translation Lexicon. *Conférence TALN*, Cargèse, pp. 12-17.

Procter, Paul (Ed.) (1978) *Longman Dictionary o Contemporary English*. Longman Group Ltd., Essex, UK.

Sanfilippo, A. (1998) Ranking Text Units According to Textual Saliency, Connectivity and Topic Aptness. *COLING-ACL 1998*: 1157-1163.

Sanfilippo, A., S. Tratz, M. Gregory, A.Chappell, P. Whitney, C. Posse, P. Paulson, B. Baddeley, R. Hohimer, A. White. (2006) Automating Ontological Annotation with WordNet. *Proceedings of the 3rd Global WordNet Conference*, Jeju Island, South Korea, Jan 19-26 2006.

Snyder, B. and M. Palmer. 2004. The English all-words task. *SENSEVAL-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*. Barcelona, Spain.

Suárez, A., Palomar, M. (2002) Word sense vs. word domain disambiguation: a maximum entropy approach. In Sojka P., Kopecek I., Pala K., eds.: Text, Speech and Dialogue (TSD 2002). Volume 2448 of Lecture Notes in Artificial Intelligence, Springer, (2002) 131—138.

Wilks, Y. and Stevenson, M. (1998) Word sense disambiguation using optimised combinations of knowledge sources. *Proceedings of the 17th international conference on Computational Linguistics*, pp. 1398—1402.