SemEval-2010 Task 17: All-words Word Sense Disambiguation on a Specific Domain

Eneko Agirre
IXA NLP group
UBC
Donostia, Basque Country
e.agirre@ehu.es

Oier Lopez de Lacalle
IXA NLP group
UBC
Donostia, Basque Country
oier.lopezdelacalle@ehu.es

Christiane Fellbaum
Department of Computer Science
Princeton University
Princeton, USA
fellbaum@princeton.edu

Andrea Marchetti
IIT
CNR
Pisa, Italy
andrea.marchetti@iit.cnr.it

Antonio Toral
ILC
CNR
Pisa, Italy
antonio.toral@ilc.cnr.it

Piek Vossen
Faculteit der Letteren
Vrije Universiteit Amsterdam
Amsterdam, Netherlands
p.vossen@let.vu.nl

Abstract

Domain portability and adaptation of NLP components and Word Sense Disambiguation systems present new challenges. The difficulties found by supervised systems to adapt might change the way we assess the strengths and weaknesses of supervised and knowledge-based WSD systems. Unfortunately, all existing evaluation datasets for specific domains are lexical-sample corpora. With this paper we want to motivate the creation of an all-words test dataset for WSD on the environment domain in several languages, and present the overall design of this SemEval task.

1 Introduction

Word Sense Disambiguation (WSD) competitions have focused on general domain texts, as attested in the last Senseval and Semeval competitions (Kilgarriff, 2001; Mihalcea et al., 2004; Pradhan et al., 2007). Specific domains pose fresh challenges to WSD systems: the context in which the senses occur might change, distributions and predominant senses vary, some words tend to occur in fewer senses in specific domains, and new senses and terms might be involved. Both supervised and knowledge-based systems are affected by these issues: while the first suffer from different context and sense priors, the later suffer from lack of coverage of domain-related words and information.

Domain adaptation of supervised techniques is a hot issue in Natural Language Processing, including Word Sense Disambiguation. Supervised Word Sense Disambiguation systems trained on general corpora are known to perform worse when applied to specific domains (Escudero et al., 2000; Martínez and Agirre, 2000), and domain adaptation techniques have been proposed as a solution to this problem with mixed results.

Current research on applying WSD to specific domains has been evaluated on three available lexical-sample datasets (Ng and Lee, 1996; Weeber et al., 2001; Koeling et al., 2005). This kind of dataset contains hand-labeled examples for a handful of selected target words. As the systems are evaluated on a few words, the actual performance of the systems over complete texts can not be measured. Differences in behavior of WSD systems when applied to lexical-sample and all-words datasets have been observed on previous Senseval and Semeval competitions (Kilgarriff, 2001; Mihalcea et al., 2004; Pradhan et al., 2007): supervised systems attain results on the high 80's and beat the most frequent baseline by a large margin for lexical-sample datasets, but results on the all-words datasets were much more modest, on the low 70's, and a few points above the most frequent baseline.

Thus, the behaviour of WSD systems on domain-specific texts is largely unknown. While some words could be supposed to behave in similar ways, and thus be amenable to be properly treated by a generic

Proceedings of the NAACL HLT Workshop on Semantic Evaluations: Recent Achievements and Future Directions, pages 123–128,
Boulder, Colorado, June 2009.

WSD algorithm, other words have senses closely linked to the domain, and might be disambiguated using purpose-built domain adaptation strategies (cf. Section 4). While it seems that domain-specific WSD might be a tougher problem than generic WSD, it might well be that domain-related words are easier to disambiguate.

The main goal of this task is to provide a multilingual testbed to evaluate WSD systems when faced with full-texts from a specific domain, that of environment-related texts. The paper is structured as follows. The next section presents current lexical sample datasets for domain-specific WSD. Section 3 presents some possible settings for domain adaptation. Section 4 reviews the state-of-the art in domain-specific WSD. Section 5 presents the design of our task, and finally, Section 6 draws some conclusions.

2 Specific domain datasets available

We will briefly present the three existing datasets for domain-related studies in WSD, which are all lexical-sample.

The most commonly used dataset is the Defense Science Organization (DSO) corpus (Ng and Lee, 1996), which comprises sentences from two different corpora. The first is the Wall Street Journal (WSJ), which belongs to the financial domain, and the second is the Brown Corpus (BC) which is a balanced corpora of English usage. 191 polysemous words (nouns and verbs) of high frequency in WSJ and BC were selected and a total of 192,800 occurrences of these words were tagged with WordNet 1.5 senses, more than 1,000 instances per word in average. The examples from BC comprise 78,080 occurrences of word senses, and examples from WSJ consist on 114,794 occurrences. In domain adaptation experiments, the Brown Corpus examples play the role of general corpora, and the examples from the WSJ play the role of domain-specific examples.

Koeling *et al.* (2005) present a corpus were the examples are drawn from the balanced BNC corpus (Leech, 1992) and the SPORTS and FINANCES sections of the newswire Reuters corpus (Rose et al., 2002), comprising around 300 examples (roughly 100 from each of those corpora) for each of the 41 nouns. The nouns were selected because they were salient in either the SPORTS or FINANCES domains, or because they had senses linked to those domains. The occurrences were hand-tagged with the senses from WordNet version 1.7.1 (Fellbaum, 1998). In domain adaptation experiments the BNC examples play the role of general corpora, and the FINANCES and SPORTS examples the role of two specific domain corpora.

Finally, a dataset for biomedicine was developed by Weeber *et al.* (2001), and has been used as a benchmark by many independent groups. The UMLS Metathesaurus was used to provide a set of possible meanings for terms in biomedical text. 50 ambiguous terms which occur frequently in MEDLINE were chosen for inclusion in the test set. 100 instances of each term were selected from citations added to the MEDLINE database in 1998 and manually disambiguated by 11 annotators. Twelve terms were flagged as "problematic" due to substantial disagreement between the annotators. In addition to the meanings defined in UMLS, annotators had the option of assigning a special tag ("none") when none of the UMLS meanings seemed appropriate.

Although these three corpora are useful for WSD research, it is difficult to infer which would be the performance of a WSD system on full texts. The corpus of Koeling et al., for instance, only includes words which where salient for the target domains, but the behavior of WSD systems on other words cannot be explored. We would also like to note that while the biomedicine corpus tackles scholarly text of a very specific domain, the WSJ part of the DSO includes texts from a financially oriented newspaper, but also includes news of general interest which have no strict relation to the finance domain.

3 Possible settings for domain adaptation

When performing supervised WSD on specific domains the first setting is to train on a general domain data set and to test on the specific domain (**source setting**). If performance would be optimal, this would be the ideal solution, as it would show that a generic WSD system is robust enough to tackle texts from new domains, and domain adaptation would not be necessary.

The second setting (**target setting**) would be to train the WSD systems only using examples from

the target domain. If this would be the optimal setting, it would show that there is no cost-effective method for domain adaptation. WSD systems would need fresh examples every time they were deployed in new domains, and examples from general domains could be discarded.

In the third setting, the WSD system is trained with examples coming from both the general domain and the specific domain. Good results in this setting would show that **supervised domain adaptation** is working, and that generic WSD systems can be supplemented with hand-tagged examples from the target domain.

There is an additional setting, where a generic WSD system is supplemented with untagged examples from the domain. Good results in this setting would show that **semi-supervised domain adaptation** works, and that generic WSD systems can be supplemented with untagged examples from the target domain in order to improve their results.

Most of current all-words generic supervised WSD systems take SemCor (Miller et al., 1993) as their source corpus, i.e. they are trained on SemCor examples and then applied to new examples. SemCor is the largest publicly available annotated corpus. It's mainly a subset of the Brown Corpus, plus the novel The Red Badge of Courage. The Brown corpus is balanced, yet not from the general domain, as it comprises 500 documents drawn from different domains, each approximately 2000 words long. Although the Brown corpus is balanced, SemCor is not, as the documents were not chosen at random.

4 State-of-the-art in WSD for specific domains

Initial work on domain adaptation for WSD systems showed that WSD systems were not able to obtain better results on the source or adaptation settings compared to the target settings (Escudero et al., 2000), showing that a generic WSD system (i.e. based on hand-annotated examples from a generic corpus) would not be useful when moved to new domains.

Escudero et al. (2000) tested the supervised adaptation scenario on the DSO corpus, which had examples from the Brown Corpus and Wall Street Journal corpus. They found that the source corpus did not help when tagging the target corpus, showing that tagged corpora from each domain would suffice, and concluding that hand tagging a large general corpus would not guarantee robust broad-coverage WSD. Agirre and Martínez (2000) used the same DSO corpus and showed that training on the subset of the source corpus that is topically related to the target corpus does allow for domain adaptation, obtaining better results than training on the target data alone.

In (Agirre and Lopez de Lacalle, 2008), the authors also show that state-of-the-art WSD systems are not able to adapt to the domains in the context of the Koeling *et al.* (2005) dataset. While WSD systems trained on the target domain obtained 85.1 and 87.0 of precision on the sports and finances domains, respectively, the same systems trained on the BNC corpus (considered as a general domain corpus) obtained 53.9 and 62.9 of precision on sports and finances, respectively. Training on both source and target was inferior that using the target examples alone.

Supervised adaptation

Supervised adaptation for other NLP tasks has been widely reported. For instance, (Daumé III, 2007) shows that a simple feature augmentation method for SVM is able to effectively use both labeled target and source data to provide the best domain-adaptation results in a number of NLP tasks. His method improves or equals over previously explored more sophisticated methods (Daumé III and Marcu, 2006; Chelba and Acero, 2004). In contrast, (Agirre and Lopez de Lacalle, 2009) reimplemented this method and showed that the improvement on WSD in the (Koeling et al., 2005) data was marginal.

Better results have been obtained using purpose-built adaptation methods. Chan and Ng (2007) performed supervised domain adaptation on a manually selected subset of 21 nouns from the DSO corpus. They used active learning, count-merging, and predominant sense estimation in order to save target annotation effort. They showed that adding just 30% of the target data to the source examples the same precision as the full combination of target and source data could be achieved. They also showed that using the source corpus significantly improved results when only 10%-30% of the target corpus was used for training. In followup work (Zhong et

Projections for 2100 suggest that temperature in Europe will have risen by between 2 to 6.3 C above 1990 levels. The sea level is projected to rise, and a greater frequency and intensity of extreme weather events are expected. Even if emissions of greenhouse gases stop today, these changes would continue for many decades and in the case of sea level for centuries. This is due to the historical build up of the gases in the atmosphere and time lags in the response of climatic and oceanic systems to changes in the atmospheric concentration of the gases.

Figure 1: Sample text from the environment domain.

al., 2008), the feature augmentation approach was combined with active learning and tested on the OntoNotes corpus, on a large domain-adaptation experiment. They significantly reduced the effort of hand-tagging, but only obtained positive domain-adaptation results for smaller fractions of the target corpus.

In (Agirre and Lopez de Lacalle, 2009) the authors report successful adaptation on the (Koeling et al., 2005) dataset on supervised setting. Their method is based on the use of unlabeled data, reducing the feature space with SVD, and combination of features using an ensemble of kernel methods. They report 22% error reduction when using both source and target data compared to a classifier trained on target the target data alone, even when the full dataset is used.

Semi-supervised adaptation

There are less works on semi-supervised domain adaptation in NLP tasks, and fewer in WSD task. Blitzer et al. (2006) used Structural Correspondence Learning and unlabeled data to adapt a Part-of-Speech tagger. They carefully select so-called pivot features to learn linear predictors, perform SVD on the weights learned by the predictor, and thus learn correspondences among features in both source and target domains. Agirre and Lopez de Lacalle (2008) show that methods based on SVD with unlabeled data and combination of distinct feature spaces produce positive semi-supervised domain adaptation results for WSD.

Unsupervised adaptation

In this context, we take unsupervised to mean Knowledge-Based methods which do not require hand-tagged corpora. The predominant sense acquisition method was succesfully applied to specific domains in (Koeling et al., 2005). The methos has two steps: In the first, a corpus of untagged text from the target domain is used to construct a thesaurus of similar words. In the second, each target word is disambiguated using pairwise WordNet-based similarity measures, taking as pairs the target word and each of the most related words according to the thesaurus up to a certain threshold. This method aims to obtain, for each target word, the sense which is the most predominant for the target corpus. When a general corpus is used, the most predominant sense in general is obtained, and when a domain-specific corpus is used, the most predominant sense for that corpus is obtained (Koeling et al., 2005). The main motivation of the authors is that the most frequent sense is a very powerful baseline, but it is one which requires hand-tagging text, while their method yields similar information automatically. The results show that they are able to obtain good results. In related work, (Agirre et al., 2009) report improved results using the same strategy but applying a graph-based WSD method, and highlight the domain-adaptation potential of unsupervised knowledge-based WSD systems compared to supervised WSD.

5 Design of the WSD-domain task

This task was designed in the context of Kyoto (Piek Vossen and VanGent, 2008)[1], an Asian-European project that develops a community platform for modeling knowledge and finding facts across languages and cultures. The platform operates as a Wiki system with an ontological support that social communities can use to agree on the meaning of terms in specific domains of their interest. Kyoto will focus on the environmental domain because it poses interesting challenges for information sharing, but the techniques and platforms will be independent of the application domain. Kyoto

[1] http://www.kyoto-project.eu/

will make use of semantic technologies based on ontologies and WSD in order to extract and represent relevant information for the domain, and is thus interested on measuring the performance of WSD techniques on this domain.

The WSD-domain task will comprise comparable all-words test corpora on the environment domain. Texts from the European Center for Nature Conservation[2] and Worldwide Wildlife Forum[3] will be used in order to build domain specific test corpora. We will select documents that are written for a general but interested public and that involve specific terms from the domain. The document content will be comparable across languages. Figure 1 shows an example in English related to global warming.

The data will be available in a number of languages: English, Dutch, Italian and Chinese. The sense inventories will be based on wordnets of the respective languages, which will be updated to include new vocabulary and senses. The test data will comprise three documents of around 2000 words each for each language. The annotation procedure will involve double-blind annotation plus adjudication, and inter-tagger agreement data will be provided. The formats and scoring software will follow those of Senseval-3[4] and SemEval-2007[5] English all-words tasks.

There will not be training data available, but participants are free to use existing hand-tagged corpora and lexical resources (e.g. SemCor and previous Senseval and SemEval data). We plan to make available a corpus of documents from the same domain as the selected documents, as well as wordnets updated to include the terms and senses in the selected documents.

6 Conclusions

Domain portability and adaptation of NLP components and Word Sense Disambiguation systems present new challenges. The difficulties found by supervised systems to adapt might change the way we assess the strengths and weaknesses of supervised and knowledge-based WSD systems. Unfortunately, all existing evaluation datasets for specific domains are lexical-sample corpora. With this paper we have motivated the creation of an all-words test dataset for WSD on the environment domain in several languages, and presented the overall design of this SemEval task.

Further details can be obtained from the Semeval-2010[6] website, our task website[7], and in our distribution list[8]

7 Acknowledgments

The organization of the task is partially funded by the European Commission (KYOTO FP7 ICT-2007-211423) and the Spanish Research Department (KNOW TIN2006-15049-C03-01).

References

Eneko Agirre and Oier Lopez de Lacalle. 2008. On robustness and domain adaptation using SVD for word sense disambiguation. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 17–24, Manchester, UK, August. Coling 2008 Organizing Committee.

Eneko Agirre and Oier Lopez de Lacalle. 2009. Supervised domain adaptation for wsd. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (EACL-09)*.

E. Agirre, O. Lopez de Lacalle, and A. Soroa. 2009. Knowledge-based WSD and specific domains: Performing over supervised WSD. In *Proceedings of IJCAI*, Pasadena, USA.

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, pages 120–128, Sydney, Australia, July. Association for Computational Linguistics.

Yee Seng Chan and Hwee Tou Ng. 2007. Domain adaptation with active learning for word sense disambiguation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 49–56, Prague, Czech Republic, June. Association for Computational Linguistics.

Ciprian Chelba and Alex Acero. 2004. Adaptation of maximum entropy classifier: Little data can help a lot. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Barcelona, Spain.

[2] http://www.ecnc.org

[3] http://www.wwf.org

[4] http://www.senseval.org/senseval3

[5] http://nlp.cs.swarthmore.edu/semeval/

[6] http://semeval2.fbk.eu/

[7] http://xmlgroup.iit.cnr.it/SemEval2010/

[8] http://groups.google.com/groups/wsd-domain

Hal Daumé III and Daniel Marcu. 2006. Domain adaptation for statistical classifiers. *Journal of Artificial Intelligence Research*, 26:101–126.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 256–263, Prague, Czech Republic, June. Association for Computational Linguistics.

Gerard Escudero, Lluiz Márquez, and German Rigau. 2000. An Empirical Study of the Domain Dependence of Supervised Word Sense Disambiguation Systems. *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP/VLC*.

C. Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

A. Kilgarriff. 2001. English Lexical Sample Task Description. In *Proceedings of the Second International Workshop on evaluating Word Sense Disambiguation Systems*, Toulouse, France.

R. Koeling, D. McCarthy, and J. 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. HLT/EMNLP*, pages 419–426, Ann Arbor, Michigan.

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

David Martínez and Eneko Agirre. 2000. One Sense per Collocation and Genre/Topic Variations. *Conference on Empirical Method in Natural Language*.

R. Mihalcea, T. Chklovski, and Adam Killgariff. 2004. The Senseval-3 English lexical sample task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain.

G.A. Miller, C. Leacock, R. Tengi, and R.Bunker. 1993. A Semantic Concordance. In *Proceedings of the ARPA Human Language Technology Workshop. Distributed as* Human Language Technology *by San Mateo, CA: Morgan Kaufmann Publishers.*, pages 303–308, Princeton, NJ.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting of the Association for Computationla Linguistics (ACL)*, pages 40–47.

Nicoletta Calzolari Christiane Fellbaum Shu-kai Hsieh Chu-Ren Huang Hitoshi Isahara Kyoko Kanzaki Andrea Marchetti Monica Monachini Federico Neri Remo Raffaelli German Rigau Maurizio Tescon Piek Vossen, Eneko Agirre and Joop VanGent. 2008. Kyoto: a system for mining, structuring and distributing knowledge across languages and cultures. In European Language Resources Association (ELRA), editor, *Proceedings of the Sixth International Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, may.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. Semeval-2007 task-17: English lexical sample, srl and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic.

Tony G. Rose, Mark Stevenson, and Miles Whitehead. 2002. The Reuters Corpus Volumen 1: from Yesterday's News to Tomorrow's Language Resources. In *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002)*, pages 827–832, Las Palmas, Canary Islands.

Marc Weeber, James G. Mork, and Alan R. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *Proceedings of the AMAI Symposium*, pages 746–750, Washington, DC.

Zhi Zhong, Hwee Tou Ng, and Yee Seng Chan. 2008. Word sense disambiguation using OntoNotes: An empirical study. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 1002–1010, Honolulu, Hawaii, October. Association for Computational Linguistics.