CFILT: Resource Conscious Approaches for All-Words Domain Specific WSD

Anup Kulkarni **Mitesh M. Khapra** **Saurabh Sohoney** **Pushpak Bhattacharyya**
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay,
Powai, Mumbai 400076,
India
{anup,miteshk,saurabhsohoney,pb}@cse.iitb.ac.in

Abstract

We describe two approaches for *All-words Word Sense Disambiguation on a Specific Domain*. The first approach is a knowledge based approach which extracts domain-specific largest connected components from the Wordnet graph by exploiting the semantic relations between all candidate synsets appearing in a domain-specific untagged corpus. Given a test word, disambiguation is performed by considering only those candidate synsets that belong to the *top-k* largest connected components.

The second approach is a weakly supervised approach which relies on the "*One Sense Per Domain*" heuristic and uses a few hand labeled examples for the most frequently appearing words in the target domain. Once the most frequent words have been disambiguated they can provide strong clues for disambiguating other words in the sentence using an iterative disambiguation algorithm. Our weakly supervised system gave the **best performance** across all systems that participated in the task even when it used as few as 100 hand labeled examples from the target domain.

1 Introduction

Domain specific WSD exhibits high level of accuracy even for the all-words scenario (Khapra et al., 2010) - provided training and testing are on the same domain. However, the effort of creating the training corpus - annotated sense marked corpora - for every domain of interest has always been a matter of concern. Therefore, attempts have been made to develop unsupervised (McCarthy et al., 2007; Koeling et al., 2005) and knowledge based techniques (Agirre et al., 2009) for WSD which do not need sense marked corpora. However, such approaches have not proved effective, since they typically do not perform better than the Wordnet first sense baseline accuracy in the all-words scenario.

Motivated by the desire to develop *annotation-lean* all-words domain specific techniques for WSD we propose two resource conscious approaches. The first approach is a knowledge based approach which focuses on retaining only domain specific synsets in the Wordnet using a two step pruning process. In the first step, the Wordnet graph is restricted to only those synsets which contain words appearing in an untagged domain-specific corpus. In the second step, the graph is pruned further by retaining only the largest connected components of the pruned graph. Each target word in a given sentence is then disambiguated using an iterative disambiguation process by considering only those candidate synsets which appear in the *top-k* largest connected components. Our knowledge based approach performed better than current state of the art knowledge based approach (Agirre et al., 2009). Also, the precision was better than the Wordnet first sense baseline even though the F-score was slightly lower than the baseline.

The second approach is a weakly supervised approach which uses a few hand labeled examples for the most frequent words in the target domain in addition to the publicly available mixed-domain SemCor (Miller et al., 1993) corpus. The underlying assumption is that words exhibit *"One Sense Per Domain"* phenomenon and hence even as few as 5 training examples per word would be sufficient to identify the predominant sense of the most frequent words in the target domain. Further, once the most frequent words have been disambiguated using the predominant sense, they can provide strong clues for disambiguating other words in the

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 421–426,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

sentence. Our weakly supervised system gave the **best performance** across all systems that participated in the task even when it used **as few as 100 hand labeled examples from the target domain**.

The remainder of this paper is organized as follows. In section 2 we describe related work on domain-specific WSD. In section 3 we discuss an Iterative Word Sense Disambiguation algorithm which lies at the heart of both our approaches. In section 4 we describe our knowledge based approach. In section 5 we describe our weakly supervised approach. In section 6 we present results and discussions followed by conclusion in section 7.

2 Related Work

There are two important lines of work for domain specific WSD. The first focuses on target word specific WSD where the results are reported on a handful of target words (41-191 words) on three lexical sample datasets, *viz.*, DSO corpus (Ng and Lee, 1996), MEDLINE corpus (Weeber et al., 2001) and the corpus of Koeling et al. (2005). The second focuses on all-words domain specific WSD where the results are reported on large annotated corpora from two domains, *viz.*, TOURISM and HEALTH (Khapra et al., 2010).

In the target word setting, it has been shown that unsupervised methods (McCarthy et al., 2007) and knowledge based methods (Agirre et al., 2009) can do better than wordnet first sense baseline and in some cases can also outperform supervised approaches. However, since these systems have been tested only for certain target words, the question of their utility in all words WSD it still open .

In the all words setting, Khapra et al. (2010) have shown significant improvements over the wordnet first sense baseline using a fully supervised approach. However, the need for sense annotated corpus in the domain of interest is a matter of concern and provides motivation for adapting their approach to annotation scarce scenarios. Here, we take inspiration from the target-word specific results reported by Chan and Ng (2007) where by using just 30% of the target data they obtained the same performance as that obtained by using the entire target data.

We take the fully supervised approach of (Khapra et al., 2010) and convert it to a weakly supervised approach by using only a handful of hand labeled examples for the most frequent words appearing in the target domain. For the remaining words we use the sense distributions learnt from SemCor (Miller et al., 1993) which is a publicly available mixed domain corpus. Our approach is thus based on the *"annotate-little from the target domain"* paradigm and does better than all the systems that participated in the shared task.

Even our knowledge based approach does better than current state of the art knowledge based approaches (Agirre et al., 2009). Here, we use an untagged corpus to prune the Wordnet graph thereby reducing the number of candidate synsets for each target word. To the best of our knowledge such an approach has not been tried earlier.

3 Iterative Word Sense Disambiguation

The Iterative Word Sense Disambiguation (IWSD) algorithm proposed by Khapra et al. (2010) lies at the heart of both our approaches. They use a scoring function which combines corpus based parameters (such as, sense distributions and corpus co-occurrence) and Wordnet based parameters (such as, semantic similarity, conceptual distance, *etc.*) for ranking the candidates synsets of a word. The algorithm is iterative in nature and involves the following steps:

- Tag all monosemous words in the sentence.
- Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
- At each stage rank the candidate senses of a word using the scoring function of Equation (1).

$$S^* = \arg\max_i (\theta_i V_i + \sum_{j \in J} W_{ij} * V_i * V_j) \qquad (1)$$

where,

$$\begin{aligned}
i &\in \textit{Candidate Synsets} \\
J &= \textit{Set of disambiguated words} \\
\theta_i &= BelongingnessToDominantConcept(S_i) \\
V_i &= P(S_i|word) \\
W_{ij} &= CorpusCooccurrence(S_i, S_j) \\
&\quad * 1/WNConceptualDistance(S_i, S_j) \\
&\quad * 1/WNSemanticGraphDistance(S_i, S_j)
\end{aligned}$$

The scoring function as given above cleanly separates the self-merit of a synset ($P(S_i|word)$)

as learnt from a tagged corpus and its interaction-merit in the form of corpus co-occurrence, conceptual distance, and wordnet-based semantic distance with the senses of other words in the sentence. The scoring function can thus be easily adapted depending upon the amount of information available. For example, in the weakly supervised setting, $P(S_i|word)$ will be available for some words for which either manually hand labeled training data from environment domain is used or which appear in the SemCor corpus. For such words, all the parameters in Equation (1) will be used for scoring the candidate synsets and for remaining words only the interaction parameters will be used. Similarly, in the knowledge based setting, $P(S_i|word)$ will never be available and hence only the wordnet based interaction parameters (*i.e.*, $WNConceptualDistance(S_i, S_j)$ *and* $WNSemanticGraphDistance(S_i, S_j)$) will be used for scoring the pruned list of candidate synsets. Please refer to (Khapra et al., 2010) for the details of how each parameter is calculated.

4 Knowledge-Based WSD using Graph Pruning

Wordnet can be viewed as a graph where synsets act as nodes and the semantic relations between them act as edges. It should be easy to see that given a domain-specific corpus, synsets from some portions of this graph would be more likely to occur than synsets from other portions. For example, given a corpus from the HEALTH domain one might expect synsets belonging to the sub-trees of *"doctor"*, *"medicine"*, *"disease"* to appear more frequently than the synsets belonging to the sub-tree of *"politics"*. Such dominance exhibited by different components can be harnessed for domain-specific WSD and is the motivation for our work.

The crux of the approach is to identify such domain specific components using a two step pruning process as described below:

Step 1: First, we use an untagged corpus from the environment domain to identify the unique words appearing in the domain. Note that, by unique words we mean all content words which appear at least once in the environment corpus (these words may or may not appear in a general mixed domain corpus). This untagged corpus containing 15 documents (22K words) was downloaded from the websites of WWF[1] and ECNC[2] and contained articles on *Climate Change, Deforestation, Species Extinction, Marine Life and Ecology*. Once the unique words appearing in this environment-specific corpus are identified, we restrict the Wordnet graph to only those synsets which contain one or more of these unique words as members. This step thus eliminates all spurious synsets which are not related to the environment domain.

Step 2: In the second step, we perform a *Breadth-First-Search* on the pruned graph to identify the connected components of the graph. While traversing the graph we consider only those edges which correspond to the *hypernymy-hyponymy* relation and ignore all other semantic relations as we observed that such relations add noise to the components. The top-5 largest components thus identified were considered to be environment-specific components. A subset of synsets appearing in one such sample component is listed in Table 1.
Each target word in a given sentence is then disambiguated using the IWSD algorithm described in section 3. However, now the $\arg\max$ of Equation (1) is computed only over those candidate synsets which belong to the top-5 largest components and all other candidate synsets are ignored. The suggested pruning technique is indeed very harsh and as a result there are many words for which none of their candidate synsets belong to these top-5 largest components. These are typically domain-invariant words for which pruning does not make sense as the synsets of such generic words do not belong to domain-specific components of the Wordnet graph. In such cases, we consider all the candidate synsets of these words while computing the $\arg\max$ of Equation (1).

5 Weakly Supervised WSD

Words are known to exhibit *"One Sense Per Domain"*. For example, in the HEALTH domain the word *cancer* will invariably occur in the *disease* sense and almost never in the sense of *a zodiac sign*. This is especially true for the most frequently appearing nouns in the domain as these are typically domain specific nouns. For example, nouns such as *farmer, species, population, conservation, nature, etc.* appear very frequently in the environment domain and exhibit a clear predominant

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

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

{ **safety**} - NOUN - the state of being certain that adverse effects will not be caused by some agent under defined conditions; ”insure the safety of the children”; ”the reciprocal of safety is risk”

{**preservation, saving**} - NOUN - the activity of protecting something from loss or danger

{**environment**} - NOUN - the totality of surrounding conditions; ”he longed for the comfortable environment of his living room”

{**animation, life, living, aliveness**} - NOUN - the condition of living or the state of being alive; ”while there’s life there’s hope”; ”life depends on many chemical and physical processes”

{**renovation, restoration, refurbishment**} - NOUN - the state of being restored to its former good condition; ”the inn was a renovation of a Colonial house”

{**ecology**} - NOUN - the environment as it relates to living organisms; ”it changed the ecology of the island”

{**development**} - NOUN - a state in which things are improving; the result of developing (as in the early part of a game of chess); ”after he saw the latest development he changed his mind and became a supporter”; ”in chess your should take care of your development before moving your queen”

{**survival, endurance**} - NOUN - a state of surviving; remaining alive

... ...
... ...

Table 1: Environment specific component identified after pruning

sense in the domain. As a result as few as 5 hand labeled examples per noun are sufficient for finding the predominant sense of these nouns. Further, once these most frequently occurring nouns have been disambiguated they can help in disambiguating other words in the sentence by contributing to the interaction-merit of Equation (1) (note that in Equation (1), $J =$ *Set of disambiguated words*).

Based on the above intuition, we slightly modified the IWSD algorithm and converted it to a weakly supervised algorithm. The original algorithm as described in section 3 uses monosemous words as seed input (refer to the first step of the algorithm). Instead, we use the most frequently appearing nouns as the seed input. These nouns are disambiguated using their pre-dominant sense as calculated from the hand labeled examples. Our weakly supervised IWSD algorithm can thus be summarized as follows

- If a word w in a test sentence belongs to the list of most frequently appearing domain-specific nouns then disambiguate it first using its self-merit (*i.e.,* $P(S_i|word)$) as learnt from the hand labeled examples.
- Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
- While disambiguating the remaining words rank the candidate senses of a word using the self-merit learnt from SemCor and the interaction-merit based on previously disambiguated words.

The most frequent words and the corresponding examples to be hand labeled are extracted from the same 15 documents (22K words) as described in section 4.

6 Results

We report the performance of our systems in the SEMEVAL task on *All-words Word Sense Disambiguation on a Specific Domain* (Agirre et al., 2010). The task involved sense tagging 1398 nouns and verbs from 3 documents extracted from the environment domain. We submitted one run for the knowledge based system and 2 runs for the weakly supervised system. For the weakly supervised system, in one run we used 5 training examples each for the 80 most frequently appearing nouns in the domain and in the second run we used 5 training examples each for the 200 most frequently appearing nouns. Both our submissions in the weakly supervised setting performed better than all other systems that participated in the shared task. Post-submission we even experimented with using 5 training examples each for **as few as 20 most frequent nouns** and even in this case we found that our weakly supervised system **performed better than all other systems** that participated in the shared task.

The precision of our knowledge based system was slightly better than the most frequent sense (MFS) baseline reported by the task organizers but the recall was slightly lower than the baseline. Also, our approach does better than the current state of the art knowledge based approach (Personalized Page Rank approach of Agirre et al. (2009)).

All results are summarized in Table 2. The following guide specifies the systems reported:

- ***WS-k:*** Weakly supervised approach using 5 training examples for the *k* most frequently appearing nouns in the environment domain.
- ***KB:*** Knowledge based approach using graph based pruning.
- ***PPR:*** Personalized PageRank approach of Agirre et al. (2009).
- ***MFS:*** Most Frequent Sense baseline provided by the task organizers.
- ***Random:*** Random baseline provided by the task organizers.

System	Precision	Recall	Rank in shared task
WS-200	0.570	0.555	1
WS-80	0.554	0.540	2
WS-20	0.548	0.535	3 (Post submission)
KB	0.512	0.495	7
PPR	0.373	0.368	24 (Post submission)
MFS	0.505	0.505	6
Random	0.23	0.23	30

Table 2: The performance of our systems in the shared task

In Table 3 we provide the results of WS-200 for each POS category. As expected, the results for nouns are much better than those for verbs mainly because nouns are more likely to stick to the "One sense per domain" property than verbs.

Category	Precision	Recall
Verbs	45.37	42.89
Nouns	59.64	59.01

Table 3: The performance of WS-200 on each POS category

7 Conclusion

We presented two resource conscious approaches for *All-words Word Sense Disambiguation on a Specific Domain*. The first approach is a knowledge based approach which retains only domain specific synsets from the Wordnet by using a two step pruning process. This approach does better than the current state of the art knowledge based approaches although its performance is slightly lower than the Most Frequent Sense baseline. The second approach which is a weakly supervised approach based on the *"annotate-little from the target domain"* paradigm performed better than all systems that participated in the task even when it used as few as 100 hand labeled examples from the target domain. This approach establishes the veracity of the *"One sense per domain"* phenomenon by showing that even as few as five examples per word are sufficient for predicting the predominant sense of a word.

Acknowledgments

We would like to thank Siva Reddy and Abhilash Inumella (from IIIT Hyderabad, India) for providing us the results of Personalized PageRank (PPR) for comparison.

References

Eneko Agirre, Oier Lopez De Lacalle, and Aitor Soroa. 2009. Knowledge-based wsd on specific domains: Performing better than generic supervised wsd.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Shu kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain.

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

Mitesh Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2010. Domain-specific word sense disambiguation combining corpus based and wordnet based parameters. In *5th International Conference on Global Wordnet (GWC2010)*.

Rob Koeling, Diana McCarthy, and John Carroll. 2005. Domain-specific sense distributions and predominant sense acquisition. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, pages 419–426, Morristown, NJ, USA. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2007. Unsupervised acquisition of predominant word senses. *Comput. Linguist.*, 33(4):553–590.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *HLT '93: Proceedings of the workshop on Human Language Technology*, pages 303–308, Morristown, NJ, USA. Association for Computational Linguistics.

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

Marc Weeber, James G. Mork, and Alan R. Aronson. 2001. Developing a test collection for biomedical word sense disambiguation. In *In Proceedings of the American Medical Informatics Association Annual Symposium (AMIA 2001)*, pages 746–750.