IIITH: Domain Specific Word Sense Disambiguation

Siva Reddy
IIIT Hyderabad
India
gvsreddy@students.iiit.ac.in

Abhilash Inumella
IIIT Hyderabad
India
abhilashi@students.iiit.ac.in

Diana McCarthy
Lexical Computing Ltd.
United Kingdom
diana@dianamccarthy.co.uk

Mark Stevenson
University of Sheffield
United Kingdom
m.stevenson@dcs.shef.ac.uk

Abstract

We describe two systems that participated in SemEval-2010 task 17 (All-words Word Sense Disambiguation on a Specific Domain) and were ranked in the third and fourth positions in the formal evaluation. Domain adaptation techniques using the background documents released in the task were used to assign ranking scores to the words and their senses. The test data was disambiguated using the Personalized PageRank algorithm which was applied to a graph constructed from the whole of WordNet in which nodes are initialized with ranking scores of words and their senses. In the competition, our systems achieved comparable accuracy of 53.4 and 52.2, which outperforms the most frequent sense baseline (50.5).

1 Introduction

The senses in WordNet are ordered according to their frequency in a manually tagged corpus, SemCor (Miller et al., 1993). Senses that do not occur in SemCor are ordered arbitrarily after those senses of the word that have occurred. It is known from the results of SENSEVAL2 (Cotton et al., 2001) and SENSEVAL3 (Mihalcea and Edmonds, 2004) that first sense heuristic outperforms many WSD systems (see McCarthy et al. (2007)). The first sense baseline's strong performance is due to the skewed frequency distribution of word senses. WordNet sense distributions based on SemCor are clearly useful, however in a given domain these distributions may not hold true. For example, the first sense for "bank" in WordNet refers to "sloping land beside a body of river" and the second to "financial institution", but in the domain of "finance" the "financial institution" sense would be expected to be more likely than the "sloping land beside a body of river" sense. Unfortunately, it is not feasible to produce large manually sense-annotated corpora for every domain of interest. McCarthy et al. (2004) propose a method to predict sense distributions from raw corpora and use this as a first sense heuristic for tagging text with the predominant sense. Rather than assigning predominant sense in every case, our approach aims to use these sense distributions collected from domain specific corpora as a knowledge source and combine this with information from the context.

Our approach focuses on the strong influence of domain for WSD (Buitelaar et al., 2006) and the benefits of focusing on words salient to the domain (Koeling et al., 2005). Words are assigned a ranking score based on its keyness (salience) in the given domain. We use these word scores as another knowledge source.

Graph based methods have been shown to produce state-of-the-art performance for unsupervised word sense disambiguation (Agirre and Soroa, 2009; Sinha and Mihalcea, 2007). These approaches use well-known graph-based techniques to find and exploit the structural properties of the graph underlying a particular lexical knowledge base (LKB), such as WordNet. These graph-based algorithms are appealing because they take into account information drawn from the entire graph as well as from the given context, making them superior to other approaches that rely only on local information individually derived for each word.

Our approach uses the Personalized PageRank algorithm (Agirre and Soroa, 2009) over a graph

Proceedings of the 5th International Workshop on Semantic Evaluation, ACL 2010, pages 387–391,
Uppsala, Sweden, 15-16 July 2010. ©2010 Association for Computational Linguistics

representing WordNet to disambiguate ambiguous words by taking their context into consideration. We also combine domain-specific information from the knowledge sources, like sense distribution scores and keyword ranking scores, into the graph thus personalizing the graph for the given domain.

In section 2, we describe domain sense ranking. Domain keyword ranking is described in Section 3. Graph construction and personalized page rank are described in Section 4. Evaluation results over the SemEval data are provided in Section 5.

2 Domain Sense Ranking

McCarthy et al. (2004) propose a method for finding predominant senses from raw text. The method uses a thesaurus acquired from automatically parsed text based on the method described by Lin (1998). This provides the top k nearest neighbours for each target word w, along with the distributional similarity score between the target word and each neighbour. The senses of a word w are each assigned a score by summing over the distributional similarity scores of its neighbours. These are weighted by a semantic similarity score (using WordNet Similarity score (Pedersen et al., 2004) between the sense of w and the sense of the neighbour that maximizes the semantic similarity score.

More formally, let $N_w = \{n_1, n_2, \ldots n_k\}$ be the ordered set of the top k scoring neighbours of w from the thesaurus with associated distributional similarity scores $\{dss(w, n_1), dss(w, n_2), \ldots dss(w, n_k)\}$. Let $senses(w)$ be the set of senses of w. For each sense of w ($ws_i \in senses(w)$) a ranking score is obtained by summing over the $dss(w, n_j)$ of each neighbour ($n_j \in N_w$) multiplied by a weight. This weight is the WordNet similarity score ($wnss$) between the target sense (ws_i) and the sense of n_j ($ns_x \in senses(n_j)$) that maximizes this score, divided by the sum of all such WordNet similarity scores for $senses(w)$ and n_j. Each sense $ws_i \in senses(w)$ is given a sense ranking score $srs(ws_i)$ using

$$srs(ws_i) =$$

$$\sum_{n_j \epsilon N_w} dss(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum\limits_{ws_i \epsilon senses(w)} wnss(ws_i, n_j)}$$

where $wnss(ws_i, n_j) =$

$$max_{ns_x \in senses(n_j)}(wnss(ws_i, ns_x))$$

Since this approach requires only raw text, sense rankings for a particular domain can be generated by simply training the algorithm using a corpus representing that domain. We used the background documents provided to the participants in this task as a domain specific corpus. In general, a domain specific corpus can be obtained using domain-specific keywords (Kilgarriff et al., 2010). A thesaurus is acquired from automatically parsed background documents using the Stanford Parser (Klein and Manning, 2003). We used $k = 5$ to built the thesaurus. As we increased k we found the number of non-domain specific words occurring in the thesaurus increased and negatively affected the sense distributions. To counter this, one of our systems IIITH2 used a slightly modified ranking score by multiplying the effect of each neighbour with its domain keyword ranking score. The modified sense ranking $msrs(ws_j)$ score of sense ws_i is

$$msrs(ws_i) =$$

$$\sum_{n_j \epsilon N_w} dss(w, n_j) \times \frac{wnss(ws_i, n_j)}{\sum\limits_{ws_i \epsilon senses(w)} wnss(ws_i, n_j)} \times krs(n_j)$$

where $krs(n_j)$ is the keyword ranking score of the neighbour n_j in the domain specific corpus. In the next section we describe the way in which we compute $krs(n_j)$.

WordNet::Similarity::lesk (Pedersen et al., 2004) was used to compute word similarity wnss. IIITH1 and IIITH2 systems differ in the way senses are ranked. IIITH1 uses $srs(ws_j)$ whereas IIITH2 system uses $msrs(ws_j)$ for computing sense ranking scores in the given domain.

3 Domain Keyword Ranking

We extracted keywords in the domain by comparing the frequency lists of domain corpora (background documents) and a very large general corpus, ukWaC (Ferraresi et al., 2008), using the method described by Rayson and Garside (2000). For each word in the frequency list of the domain corpora, words(domain), we calculated the log-likelihood (LL) statistic as described in Rayson and Garside (2000). We then normalized LL to compute keyword ranking score $krs(w)$ of word w $words(domain)$ using

$$krs(w) = \frac{LL(w)}{\sum\limits_{w_i \in words(domain)} LL(w_i)}$$

The above score represents the keyness of the word in the given domain. Top ten keywords (in descending order of krs) in the corpora provided for this task are *species, biodiversity, life, habitat, natura*[1], *EU, forest, conservation, years, amp*[2].

4 Personalized PageRank

Our approach uses the Personalized PageRank algorithm (Agirre and Soroa, 2009) with WordNet as the lexical knowledge base (LKB) to perform WSD. WordNet is converted to a graph by representing each synset as a node (synset node) and the relationships in WordNet (hypernymy, hyponymy etc.) as edges between synset nodes. The graph is initialized by adding a node (word node) for each context word of the target word (including itself) thus creating a context dependent graph (personalized graph). The popular PageRank (Page et al., 1999) algorithm is employed to analyze this personalized graph (thus the algorithm is referred as personalized PageRank algorithm) and the sense for each disambiguous word is chosen by choosing the synset node which gets the highest weight after a certain number of iterations of PageRank algorithm.

We capture domain information in the personalized graph by using sense ranking scores and keyword ranking scores of the domain to assign initial weights to the word nodes and their edges (word-synset edge). This way we personalize the graph for the given domain.

4.1 Graph Initialization Methods

We experimented with different ways of initializing the graph, described below, which are designed to capture domain specific information.

Personalized Page rank (PPR): In this method, the graph is initialized by allocating equal probability mass to all the word nodes in the context including the target word itself, thus making the graph context sensitive. This does not include domain specific information.

Keyword Ranking scores with PPR (KRS + PPR): This is same as PPR except that context words are initialized with krs.

Sense Ranking scores with PPR (SRS + PPR): Edges connecting words and their synsets are assigned weights equal to srs. The initialization of word nodes is same as in PPR.

KRS + SRS + PPR: Word nodes are initialized with krs and edges are assigned weights equal to srs.

In addition to the above methods of unsupervised graph initialization, we also initialized the graph in a *semi-supervised* manner. WordNet (version 1.7 and above) have a field *tag_cnt* for each synset (in the file *index.sense*) which represents the number of times the synset is tagged in various semantic concordance texts. We used this information, *concordance score* (cs) of each synset, with the above methods of graph initialization as described below.

Concordance scores with PPR (CS + PPR): The graph initialization is similar to PPR initialization additionally with concordance score of synsets on the edges joining words and their synsets.

CS + KRS + PPR: The initialization graph of KRS + PPR is further initialized by assigning concordance scores to the edges connecting words and their synsets.

CS + SRS + PPR: Edges connecting words and their synsets are assigned weights equal to sum of the concordance scores and sense ranking scores i.e. $cs + srs$. The initialization of word nodes is same as in PPR.

CS + KRS + SRS + PPR: Word nodes are initialized with krs and edges are assigned weights equal to $cs + srs$.

PageRank was applied to all the above graphs to disambiguate a target word.

4.2 Experimental details of PageRank

Tool: We used UKB tool[3] (Agirre and Soroa, 2009) which provides an implementation of personalized PageRank. We modified it to incorporate our methods of graph initialization. The LKB used in our experiments is WordNet3.0 + Gloss which is provided in the tool. More details of the tools used can be found in the Appendix.

Normalizations: Sense ranking scores (srs) and keyword ranking scores (krs) have diverse ranges. We found srs generally in the range between 0 to

[1] In background documents this word occurs in reports describing Natura 2000 networking programme.

[2] This new word *"amp"* is created by our programs while extracting body text from background documents. The HTML code *"&amp;"* which represents the symbol*"&"* is converted into this word.

[3] `http://ixa2.si.ehu.es/ukb/`

	Precision	Recall
Unsupervised Graph Initialization		
PPR	37.3	36.8
KRS + PPR	38.1	37.6
SRS + PPR	48.4	47.8
KRS + SRS + PPR	48.0	47.4
Semi-supervised Graph Initialization		
CS + PPR	50.2	49.6
CS + KRS + PPR	50.1	49.5
* CS + SRS + PPR	53.4	52.8
CS + KRS + SRS + PPR	**53.6**	**52.9**
Others		
1^{st}sense	50.5	50.5
PSH	49.8	43.2

Table 1: Evaluation results on English test data of SemEval-2010 Task-17. * represents the system which we submitted to SemEval and is ranked 3rd in public evaluation.

1 and krs in the range 0 to 0.02. Since these scores are used to assign initial weights in the graph, these ranges are scaled to fall in a common range of [0, 100]. Using any other scaling method should not effect the performance much since PageRank (and UKB tool) has its own internal mechanisms to normalize the weights.

5 Evaluation Results

Test data released for this task is disambiguated using IIITH1 and IIITH2 systems. As described in Section 2, IIITH1 and IIITH2 systems differ in the way the sense ranking scores are computed. Here we project only the results of IIITH1 since IIITH1 performed slightly better than IIITH2 in all the above settings. Results of 1^{st}*sense* system provided by the organizers which assigns first sense computed from the annotations in hand-labeled corpora is also presented. Additionally, we also present the results of Predominant Sense Heuristic (PSH) which assigns every word w with the sense ws_j $(ws_j \in senses(w))$ which has the highest value of $srs(ws_j)$ computed in Section 2 similar to (McCarthy et al., 2004).

Table 1 presents the evaluation results. We used TreeTagger [4] to Part of Speech tag the test data. POS information was used to discard irrelevant senses. Due to POS tagging errors, our precision values were not equal to recall values. In the competition, we submitted IIITH1 and IIITH2 systems with CS + SRS + PPR graph initialization. IIITH1 and IIIH2 gave performances of 53.4 % and 52.2 % precision respectively. In our later experiments, we found CS + KRS + SRS + PPR has given the best performance of 53.6 % precision.

From the results, it can be seen when srs information is incorporated in the graph, precision improved by 11.1% compared to PPR in unsupervised graph initialization and by 3.19% compared to CS + PPR in semi-supervised graph initialization. Also little improvements are seen when krs information is added. This shows that domain specific information like sense ranking scores and keyword ranking scores play a major role in domain specific WSD.

The difference between the results in unsupervised and semi-supervised graph initializations may be attributed to the additional information the semi-supervised graph is having i.e. the sense distribution knowledge of non-domain specific words (common words).

6 Conclusion

This paper proposes a method for domain specific WSD. Our method is based on a graph-based algorithm (Personalized Page Rank) which is modified to include information representing the domain (sense ranking and key word ranking scores). Experiments show that exploiting this domain specific information within the graph based methods produces better results than when this information is used individually.

[4] http://www.ims.uni-stuttgart.de/projekte/corplex/TreeTagger/

Acknowledgements

The authors are grateful to Ted Pedersen for his helpful advice on the WordNet Similarity Package. We also thank Rajeev Sangal for supporting the authors Siva Reddy and Abhilash Inumella.

References

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 33–41, Morristown, NJ, USA. Association for Computational Linguistics.

Paul Buitelaar, Bernardo Magnini, Carlo Strapparava, and Piek Vossen. 2006. Domain-specific wsd. In *Word Sense Disambiguation. Algorithms and Applications, Editors: Eneko Agirre and Philip Edmonds*. Springer.

Scott Cotton, Phil Edmonds, Adam Kilgarriff, and Martha Palmer. 2001. Senseval-2. `http://www.sle.sharp.co.uk/senseval2`.

A. Ferraresi, E. Zanchetta, M. Baroni, and S. Bernardini. 2008. Introducing and evaluating ukwac, a very large web-derived corpus of english. In *Proceed-ings of the WAC4 Workshop at LREC 2008*, Marrakesh, Morocco.

Adam Kilgarriff, Siva Reddy, Jan Pomikálek, and Avinesh PVS. 2010. A corpus factory for many languages. In *LREC 2010*, Malta.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *ACL '03: Proceedings of the 41st Annual Meeting on Association for Computational Linguistics*, pages 423–430, Morristown, NJ, USA. Association for Computational Linguistics.

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.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics*, pages 768–774, Morristown, NJ, USA. Association for Computational Linguistics.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 279, Morristown, NJ, USA. Association for Computational Linguistics.

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

Rada Mihalcea and Phil Edmonds, editors. 2004. *Proceedings Senseval-3 3rd International Workshop on Evaluating Word Sense Disambiguation Systems.* ACL, Barcelona, Spain.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the ARPA Workshop on Human Language Technology*, pages 303–308. Morgan Kaufman.

Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. The pagerank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, November.

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. Wordnet::similarity: measuring the relatedness of concepts. In *HLT-NAACL '04: Demonstration Papers at HLT-NAACL 2004 on XX*, pages 38–41, Morristown, NJ, USA. Association for Computational Linguistics.

Paul Rayson and Roger Garside. 2000. Comparing corpora using frequency profiling. In *Proceedings of the workshop on Comparing corpora*, pages 1–6, Morristown, NJ, USA. Association for Computational Linguistics.

Ravi Sinha and Rada Mihalcea. 2007. Unsupervised graph-basedword sense disambiguation using measures of word semantic similarity. In *ICSC '07: Proceedings of the International Conference on Semantic Computing*, pages 363–369, Washington, DC, USA. IEEE Computer Society.

Appendix

Domain Specific Thesaurus, Sense Ranking Scores and Keyword Ranking Scores are accessible at

`http://web.iiit.ac.in/~gvsreddy/SemEval2010/`

Tools Used:

- UKB is used with options *–ppr –dict_weight*. Dictionary files which UKB uses are automatically generated using sense ranking scores srs.
- Background document words are canonicalized using KSTEM, a morphological analyzer
- The Stanford Parser is used to parse background documents to build thesaurus
- Test data is part of speech tagged using TreeTagger.