Knowledge-Based WSD on Specific Domains: Performing better than Generic Supervised WSD

Eneko Agirre and **Oier Lopez de Lacalle** and **Aitor Soroa**
Informatika Fakultatea, University of the Basque Country
20018, Donostia, Basque Country
{e.agirre,oier.lopezdelacalle,a.soroa}@ehu.es

Abstract

This paper explores the application of knowledge-based Word Sense Disambiguation systems to specific domains, based on our state-of-the-art graph-based WSD system that uses the information in WordNet. Evaluation was performed over a publicly available domain-specific dataset of 41 words related to Sports and Finance, comprising examples drawn from three corpora: one balanced corpus (BNC), and two domain-specific corpora (news related to Sports and Finance). The results show that in all three corpora our knowledge-based WSD algorithm improves over previous results, and also over two state-of-the-art supervised WSD systems trained on SemCor, the largest publicly available annotated corpus. We also show that using related words as context, instead of the actual occurrence contexts, yields better results on the domain datasets, but not on the general one. Interestingly, the results are higher for domain-specific corpus than for the general corpus, raising prospects for improving current WSD systems when applied to specific domains.

1 Introduction

Word Sense Disambiguation (WSD) is an enabling-technology that automatically chooses the intended sense of a word in context. It has been a central topic of research in Natural Language Processing (NLP) for years, and more recently it has been shown to be useful in several NLP tasks such as parsing, machine translation, information retrieval or question answering. WSD is considered as a key step in order to approach language understanding.

The best performing WSD systems are those based on supervised learning as attested in public evaluation exercises [Snyder and Palmer, 2004; Pradhan *et al.*, 2007], but they need large amounts of hand-tagged data. Generic WSD systems capable of disambiguating all words are typically trained on SemCor. Contrary to lexical-sample exercises (where plenty of training and testing examples for a handful of words are provided), all-words exercises (which comprise all words occurring in a running text, and where training data per word is more scarce) show that only a few systems beat the most frequent sense (MFS) baseline[1], with a very small difference (e.g. 65.2 to 62.4 F-score in Senseval-3 [Snyder and Palmer, 2004]).

The cause of only the small improvement over the MFS baseline is a mixture of the relatively small amount of training data available (*sparseness*) and the problems that arise when the supervised systems are applied to a different corpora from that used to train the system (*corpus mismatch*). The sparseness and corpus mismatch problems are specially acute when deploying a generic supervised WSD system on a specific domain [Escudero *et al.*, 2000]. Hand tagging examples for every new domain would provide the desired performance, but it is unfeasible in practice, because of the high manual cost involved.

In view of the problems of supervised systems, knowledge-based WSD is re-emerging as a powerful alternative. Knowledge-based systems exploit the information in a Lexical Knowledge Base (LKB) to perform WSD, without using any corpus evidence. In particular, graph-based methods are getting increasing attention from the WSD community [Sinha and Mihalcea, 2007; Navigli and Lapata, 2007]. These methods use well-known graph-based techniques to find and exploit the structural properties of the graph underlying a particular LKB. In [Agirre and Soroa, 2009] we proposed a graph-based algorithm using Personalized PageRank which outperformed other knowledge-based WSD systems in publicly available datasets. In this paper we explore the application of our algorithm to domain-specific corpora. For this we use the only domain-specific WSD evaluation dataset available [Koeling *et al.*, 2005].

The paper is structured as follows. Section 2 presents some related work. Section 3 briefly reviews the supervised systems used for comparison purposes. Section 4 presents the graph-based techniques, which are applied to WSD in Section 5. In Section 6 the evaluation framework and results are presented. Finally, the conclusions are drawn and further work is mentioned.

2 Related work

The dataset used for evaluating this work was first presented in [Koeling *et al.*, 2005]. Section 6 will present further details

[1] This baseline consists on tagging all occurrences in the test data with the sense that occurs most frequently in the training data.

on the dataset itself. Koeling *et al.* used the dataset in order to evaluate their predominant sense acquisition method, which consisted basically of two steps. In the first step, a corpus of untagged text from the target domain was used to construct a distributional thesaurus of related words. In the second step, each target word was disambiguated using pairwise similarity measures based on WordNet, taking as pairs the target word and each of the most related words according to the distributional thesaurus up to a certain threshold. This method aims to obtain, for each target word, the sense which is the most predominant in the target, domain-specific, corpus.

Our work here has a different focus, as our objective is mainly to compare the performance of state-of-the-art supervised and knowledge-based WSD systems on specific domains, and to study better ways to apply knowledge-based WSD methods on specific domains. Our proposal can also be seen as a continuation of [Koeling *et al.*, 2005], and we show that our WordNet-based WSD method yields better results. We also study whether the strategy to select one predominant sense for the whole corpus using the distributional thesaurus performs better than disambiguating each occurrence of the word separately.

Graph-based methods using WordNet have recently been shown to outperform other knowledge-based systems. Sinha and Mihalcea [2007] use graph-centrality measures over custom built graphs, where the senses of the words in the context are linked with edges weighted according to several similarity scores. Navigli and Lapata [2007] build a subgraph using a depth-first strategy over the whole graph of WordNet, and then apply a variety of graph-centrality measures, with degree yielding the best results. Our method (cf. Section 4) has been shown to perform better than those methods in [Agirre and Soroa, 2009]. The system and the data are publicly available in `http://ixa2.si.ehu.es/ukb/`.

Portability problems across corpora for supervised WSD systems have been well documented in a number of papers [Escudero *et al.*, 2000; Chan and Ng, 2007]. This problem of supervised systems underscores the importance of our results on domain-specific WSD. More recently [Agirre and Lopez de Lacalle, 2008; 2009] report positive results when adapting supervised WSD systems, as evaluated in the same dataset as used here. Our work is complementary to theirs, and holds promise for potential combinations.

3 Supervised WSD

As baselines, we use two state-of-the-art WSD classifiers, which use Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN), respectively. SVM and k-NN systems have been widely used in public evaluation exercises, and have attained high ranks in both lexical-sample and all-words tasks [Snyder and Palmer, 2004; Pradhan *et al.*, 2007].

k-NN is a memory based learning method, where the neighbors are the k labeled examples most similar to the test example. The similarity among instances is measured as the cosine of their featured vectors. The test instance is labeled with the sense obtaining the maximum sum of the weighted votes of the k most similar train instances. We set k to 5 based on previous work [Agirre and Lopez de Lacalle, 2008]. Regarding SVM, we used linear kernels, due to the high amount of learning features. No soft margin (C) was estimated for any baseline system and the default C was used. We used the one-versus-all strategy, as implemented in SVM-Light.

In order to train the classifiers, we relied on the usual features used in previous WSD work, including Local Collocations, Syntactic Dependencies and Bag-of-Words features. Both systems were trained on Semcor [Miller *et al.*, 1993], the most widely used sense-annotated corpora. Semcor consists of a subset of the Brown Corpus plus the novel The Red Badge of Courage. It contains a number of texts comprising about 500,000 words where all content words have been manually tagged with senses from WordNet. In the case where target word has fewer than 10 instances in SemCor we have applied the most frequent sense, as customary in all-words supervised WSD systems. For the 41 words in the evaluation dataset (cf. Section 6) 8 words had less than 10 training instances. The maximum amount of training instances was 114, with an average of 37.

4 PageRank and Personalized PageRank

In this Section we will introduce the PageRank and Personalized PageRank algorithms. The PageRank algorithm [Brin and Page, 1998] is a method for ranking the vertices on a graph according to their relative structural importance. The main idea of PageRank is that whenever a link from v_i to v_j exists on a graph, a vote from node i to node j is produced, and hence the rank of node j increases. Besides, the strength of the vote from i to j also depends on the rank of node i: the more important node i is, the more strength its votes will have. Alternatively, PageRank can also be viewed as the result of a random walk process, where the final rank of node i represents the probability of a random walk over the graph ending on node i, at a sufficiently large time.

Let G be a graph with N vertices $v_1, \ldots, v_N$ and d_i be the out degree of node i; let M be a $N \times N$ transition probability matrix, where $M_{ji} = \frac{1}{d_i}$ if a link from i to j exists, and zero otherwise. Then, the calculation of the *PageRank vector* $\mathbf{PR}$ over G is equivalent to solving Equation (1).

$$\mathbf{PR} = cM\mathbf{PR} + (1-c)\mathbf{v} \tag{1}$$

In the equation, $\mathbf{v}$ is a $N \times 1$ vector whose elements are $\frac{1}{N}$ and c is the so called *damping factor*, a scalar value between 0 and 1. The first term of the sum on the equation models the voting scheme described in the beginning of the section. The second term represents, loosely speaking, the probability of a surfer randomly jumping to any node. The damping factor c models the way in which these two terms are combined at each step.

The second term on Eq. (1) can also be seen as a smoothing factor that makes any graph fulfill the property of being aperiodic and irreducible, and thus guarantees that PageRank calculation converges to a unique stationary distribution.

In the traditional PageRank formulation the vector $\mathbf{v}$ is a stochastic normalized vector whose element values are all $\frac{1}{N}$, thus assigning equal probabilities to all nodes in the graph in case of random jumps. However, as pointed out by [Haveliwala, 2002], the vector $\mathbf{v}$ can be non-uniform and assign stronger probabilities to certain kinds of nodes, effectively biasing the resulting PageRank vector to prefer these nodes. Such a calculation is often called a *Personalized PageRank*. For example, if we concentrate all the probability mass on a unique node i, all random jumps on the walk will return to i and thus its rank will be high; moreover, the high rank of i will make all the nodes in its vicinity to also receive a high rank. Thus, the importance of node i given by the initial distribution of $\mathbf{v}$ spreads along the graph on successive iterations of the algorithm.

PageRank is actually calculated by applying an iterative algorithm which computes Eq. (1) successively until convergence below a given threshold is achieved, or, more typically, until a fixed number of iterations are executed. Following usual practice, we used a damping value of 0.85 and finish the calculations after 30 iterations. We did not optimize these parameters.

5 Application to WSD

In this section we will briefly explain how to apply PageRank and Personalized PageRank to knowledge-based WSD, as introduced in [Agirre and Soroa, 2009].

A Lexical Knowledge Base (LKB) is formed by a set of concepts and relations among them, plus a dictionary, which is a list of words (typically, word lemmas) linked to the corresponding concepts (senses) in the LKB. Such a LKB can be naturally represented as an undirected graph $G = (V, E)$ where nodes represent LKB concepts (v_i), and each relation between concepts v_i and v_j is represented by an undirected edge $e_{i,j}$. The entries in the dictionary are linked to their corresponding concepts by directed edges. In this work, we used WordNet 1.7 as the LKB, with all relations supplemented with disambiguated glosses as provided by the Extended WordNet. This setting was optimal in [Agirre and Soroa, 2009]. The WordNet version follows that of the evaluation dataset (cf. Section 6).

5.1 Static PageRank (PR), no context

If we apply traditional PageRank over the whole WordNet, we get a context-independent ranking of word senses. All concepts in WordNet get ranked according to their PageRank value. Given a target word, it suffices to check which is the relative ranking of its senses, and the WSD system would output the one ranking highest. We call this application of PageRank to WSD *Static PageRank*, as it does not change with the context, and we use it as a baseline.

As PageRank over undirected graphs is closely related to the degree, the Static PageRank returns the most predominant sense according to the number of relations the senses have. We think that this is closely related to the Most Frequent Sense attested in general corpora, as the lexicon builders would tend to assign more relations to the most predominant sense. In fact, our results (cf. Section 6.1) show that this is indeed the case, at least for the English WordNet.

5.2 Personalized PageRank (PPR), using context

Static PageRank is independent of context, but this is not what we want in a WSD system. Given an input piece of text we want to disambiguate all content words in the input according to the relationships among them. For this we can use Personalized PageRank over the whole WordNet graph.

Given an input text, e.g. a sentence, we extract the list $W_i \; i = 1 \ldots m$ of content words (i.e. nouns, verbs, adjectives and adverbs) which have an entry in the dictionary, and thus can be related to LKB concepts. Note that monosemous words will be related to just one concept, whereas polysemous words may be attached to several. As a result of the disambiguation process, every LKB concept receives a score. Then, for each target word to be disambiguated, we just choose its associated concept in G with maximum score.

In order to apply *Personalized PageRank* over the LKB graph, the context words are first inserted into the graph G as nodes, and linked with directed edges to their respective concepts. Then, the Personalized PageRank of the graph G is computed by concentrating the initial probability mass uniformly over the newly introduced word nodes. As the words are linked to the concepts by directed edges, they act as source nodes injecting mass into the concepts they are associated with, which thus become relevant nodes, and spread their mass over the LKB graph. Therefore, the resulting Personalized PageRank vector can be seen as a measure of the structural relevance of LKB concepts in the presence of the input context.

This method has one problem: if one of the target words has two senses which are related to each other by semantic relations, those senses would reinforce each other, and could thus dampen the effect of the other senses in the context. With this observation in mind we have used a variant where, for each target word W_i, the initial probability mass is concentrated in the senses of the words surrounding W_i, but not in the senses of the target word itself, avoiding to bias the initial score of concepts associated to target word W_i. In [Agirre and Soroa, 2009] we show that this variant gets the best results.

Given the fact that finding out the predominant sense seems a powerful option, we decided to try two further variants of the Personalized PageRank WSD algorithm. Instead of returning a different sense for each occurrence, we also evaluated the results of selecting the sense which is chosen most frequently by Personalized PageRank for the target word (*PPRank.maxsense* variant). Another alternative is to join all contexts of the target word into a single large context and then disambiguate the target word using this large context in a single run (*PPRank.all-in-one* variant).

5.3 Personalized PageRank (PPR), using related words

Instead of disambiguating the target word using the occurrence context, we could follow [Koeling *et al.*, 2005] and disambiguate the target word using the set of related words as collected from the target corpus (cf. Section 2). We would thus annotate all the occurrences of the target word in the test corpus with the same sense. For instance, in the Sports corpus, instead of disambiguating the word *coach* using each of its occurrences as context (e.g. "*Has never won a league title as a coach but took Parma to success in Europe ...*"), we would disambiguate *coach* using its most related words ac-

	Systems	BNC	Sports	Finances
Baselines	Random	*19.7	*19.2	*19.5
	SemCor MFS	*34.9 [33.60, 36.20]	*19.6 [18.40, 20.70]	*37.1 [35.70, 38.00]
	Static PRank	*36.6 [35.30, 38.00]	*20.1 [18.90, 21.30]	*39.6 [38.40, 41.00]
Supervised	SVM	*38.7 [37.30, 39.90]	*25.3 [24.00, 26.30]	*38.7 [37.10, 40.10]
	k-NN	42.8 [41.30, 44.10]	*30.3 [29.00, 31.20]	*43.4 [42.00, 44.80]
Context	PPRank	**43.8** [42.40, 44.90]	*35.6 [34.30, 37.00]	*46.9 [45.39, 48.10]
	PPRank.maxsense	*39.3 [38.00, 40.60]	*36.0 [34.70, 37.40]	*53.1 [51.70, 54.40]
	PPRank.all-in-one	*39.6 [38.20, 40.90]	*42.5 [41.20, 43.90]	*46.4 [44.90, 47.80]
Related words	[Koeling *et al.*, 2005]	*40.7 [39.20, 42.00]	*43.3 [42.00, 44.60]	*49.7 [48.00, 51.10]
	PPRank	*37.7 [36.30, 39.00]	**51.5** [50.00, 52.90]	**59.3** [57.80, 60.70]
	PPRank.th+ctx	*38.2 [36.70, 39.50]	49.9 [48.50, 51.60]	57.8 [56.40, 59.20]
Upperbound	Test MFS	*52.0 [50.60, 53.30]	*77.8 [76.60, 79.00]	*82.3 [81.00, 83.30]

Table 1: Recall of baselines and systems on each of the corpus (Sports, Finances and BNC), including confidence intervals. * means statistically significant compared to the best system in each column (in bold). *Context* rows correspond to Personalized PageRank using occurrence context (cf. Section 5.2). *Related words* rows correspond to systems using related words as context (cf. Section 5.3).

cording to the distributional thesaurus (e.g. *manager, captain, player, team, striker, ...*). In this work we use the distributional thesauri built by Koeling *et al.* [2005], one for each corpus of the evaluation dataset, i.e. Sports, Finances and BNC. Given a target noun w, Koeling *et al.* obtained a set of co-occurrence triples $< w, r, x >$ and associated frequencies, where r is a grammatical relation and x the co-occurring word in that relation. For every pair of nouns, they computed their distributional similarity comparing their respective triples using the measure suggested by Lin [1998]. Finally, the 50 most similar nouns are retrieved for each target noun.

6 Evaluation Framework and results

We used the evaluation dataset published in [Koeling *et al.*, 2005]. This dataset consists of examples retrieved from the Sports and Finance sections of the Reuters corpus, and also from the balanced British National Corpus (BNC). 41 words related to the Sports and Finance domains were selected, according to the following criteria: 18 words having at least one synset labeled as Sports or Finances according to WordNet Domains, 8 words which had salient frequencies in each domain (according to the normalized document frequency), and 7 words with equal salience in both domains. The selected words are quite polysemous and thus difficult to disambiguate, with an average polysemy of 6.7 senses, ranging from 2 to 13 senses.

Around 100 examples for each word and each of the three corpora where annotated by three reviewers with senses from WordNet v. 1.7.1, yielding an inter-tagger agreement of 65%. Koeling *et al.* [2005] did not clarify the method to select the "correct" sense, and we decided to choose the sense chosen by the majority of taggers. In case of ties we discarded the occurrence from the test set. This, and the fact that Koeling *et al.* discarded one of the target words, cause the small difference (0.2 at most) between the results reported in their [Koeling *et al.*, 2005] paper, and the ones we quote for them in Table 1.

6.1 Experimental results

As evaluation measure we use recall, the number of correct occurrences divided by the total number of occurrences. Recall is more informative than accuracy, as some methods failed to return results for a handful of occurrences (1% of occurrences in the worst case). Table 1 shows the results of the different WSD approaches on the different corpora (expressed in three main columns). The confidence interval is also shown, as computed using bootstrap resampling with 95% confidence. The systems in the table are divided in four groups of rows as follows.

The first rows report the baseline approaches, such as the random baseline, the most frequent sense as attested in SemCor and the results of the static PageRank. In the second group of rows, the results for supervised systems, k-NN and SVM, are shown. The next three rows report the variants for Personalized PageRank over occurrence context (cf. Section 5.2), followed by a further group with results for the approaches based on related words, including the results of Koeling *et al.* and the combination of applying our algorithm to the related words and each context (*th+ctx*). Finally, we show the most frequent sense according to the test data. We will now consider several issues in turn.

Baselines and supervised systems: The results show that SemCor MFS is very low, close to the random baseline and far from the Test MFS, especially for the domain-specific corpora but also on the general BNC corpus. Note that the most frequent sense in the test data may be considered as a kind of "upperbound", because our systems don't rely on hand-tagged data from the domain. The supervised systems scarcely improve over the Semcor MFS, which is consistent with state-of-the-art results over all-words datasets [Snyder and Palmer, 2004; Pradhan *et al.*, 2007]. They also lie well below the Test MFS, with a dramatic gap in the two domain corpora. The low results on the BNC show that the deployment of supervised systems is problematic, not only because of domain shifts, but also because of being applied to different corpora, even being both from the general domain.

	Similar			Different		
Systems	BNC	Sp.	Fin.	BNC	Sp.	Fin.
Semcor MFS	54.7	**65.5**	**79.0**	9.7	3.8	8.4
k-NN	**57.1**	64.6	69.9	24.6	18.5	25.4
Context PPR	50.0	34.9	64.2	**36.0**	35.9	35.0
Related PPR	38.1	53.1	73.7	24.8	**50.9**	**49.5**

Table 2: Results for those words with similar (and different) sense distributions. Best results in bold.

Static PageRank: Applying PageRank over the entire WordNet graph yields low results, very similar to those of SemCor MFS, and below those of all Personalized PageRank variants that we tried. In fact, Static PageRank seems to be closely related to the SemCor MFS, as we hypothesized in Section 5.1.

Personalized PageRank over context words: Surprisingly, applying our Personalized PageRank method for each occurrence yields results which are above the supervised systems in all three corpora, with larger improvements for the domain-specific ones. The results of the strategies for selecting one single sense as output (*maxsense* or *all-in-one*) are mixed, with slight improvements in Sports and Finances and degradation in the BNC.

Personalized PageRank over related words: Personalized PageRank over related words obtains the best results overall for Sports and Finances, and it is thus a preferred strategy to disambiguate domain-specific words. In the case of the balanced BNC corpus, the best results are for Personalized PageRank over the occurrence context. It seems that using related words is optimal for domain-specific WSD, but not for general purpose WSD, where a more personalized case-by-case treatment is required for each occurrence. Finally, the combination of the occurrence context and related words does not seem to be productive, as attested by its decrease in performance.

Comparison to [Koeling *et al.*, 2005]: Our Personalized PageRank algorithm over related words performs significantly better than [Koeling *et al.*, 2005] in both Sports and Finances. Our WSD method is closely related to the WordNet-based similarity method defined in [Hughes and Ramage, 2007]. In this sense, our WSD algorithm is an alternative to the one used in [Koeling *et al.*, 2005]. Instead of computing pairwise similarity and selecting the sense which yields the maximum additive similarity score with respect to each related word from the distributional thesaurus, our WSD method implicitly yields the sense with the maximum similarity score with respect to the full set of related words in one go. In fact, we initialize PPR with all the related words in the thesaurus, yielding the sense of the target word with the highest rank, i.e. the sense which is most closely related to those words. Our better results are consistent with the word similarity results reported in [Hughes and Ramage, 2007], which surpass other WordNet-based similarity methods, including those used by [Koeling *et al.*, 2005].

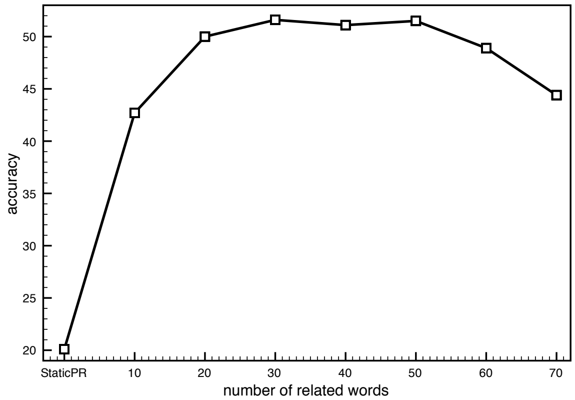

Figure 1: Learning curves for PPR (Sports) using different numbers of related words from the distributional thesaurus. Leftmost point is null context, which equals to static PR.

Results for coarse-grained senses: WordNet senses have been criticized for their fine-grainedness. Public evaluation exercises have also used coarse-grained senses as defined by the semantic files of the senses. In our case, the best results for BNC, Sports and Finances measured according to coarse senses would be of 56.9%, 61.2% and 72.0%, respectively. This is remarkable given the high polisemy of the target words. The overall results for Sports and Finances are higher than for BNC, holding promise for building effective domain-specific WSD in specific domains.

Effect of sense distributions: The performance of the supervised systems could be affected by the very different distribution of word senses in the training (Semcor) and test dataset (all three corpora), as attested by the very low performance of the Semcor MFS and the dramatic difference with respect to the Test MFS. We decided to make two groups of words for each corpus, according to the similarity of sense distributions measured as the difference between Semcor MFS and Test MFS. In the *Similar distribution* group we included those words with differences below 10 percentage points, and in the *Different distribution* group those with larger differences. Note that for each domain we acquire different set words: for Sports we had 12 *Similar* words and 29 *Different* words, for Finances we had 16 and 25, and for the BNC 19 and 22, respectively. Table 2 shows that for *Similar* sense distributions the best results are actually those of the Semcor MFS and the supervised WSD system, while the Personalized PageRank algorithms yield the best results for *Different* sense distributions.

Exploring the number of related words: Following [Koeling *et al.*, 2005] we used the 50 most similar words when doing WSD. Figure 1 shows that the number of words is an important factor, with best results for 30-50 words. These results agree with the intuition that a minimum amount of words is necessary for providing context, but introducing further terms down the list of related words involves noisier and less related words.

7 Conclusions and future work

This paper shows that, when tagging domain-specific corpora, Knowledge-Based WSD systems are a powerful alternative to generic supervised WSD systems trained on balanced corpora.

The results range from 51.5% to 59.3% (61.2% to 72% coarse-grained) for 41 domain-related and highly polysemous words. The results are higher for domain-specific corpus than for the general corpus, raising interesting prospects for improving current WSD systems when applied to specific domains.

The results also show that our knowledge-based WSD algorithm improves significantly over previous results on the same dataset. The system and the data are publicly available in `http://ixa2.si.ehu.es/ukb/`. Disambiguating related words from a distributional thesauri (instead of the actual occurrence contexts) yields better results on the domain datasets, but not on the general one.

Our analysis showed that the differences in sense distribution hurt supervised systems, and a combination of supervised and knowledge-based systems which takes this into account seems promising. In particular, our approach is complementary to supervised domain-adaptation techniques [Agirre and Lopez de Lacalle, 2008; 2009]. Given the existing difference with respect to the Test MFS, there is ample room for further improvements.

The dataset we used is a lexical-sample, and our results might depend on the actual words in the sample. For the future, we would like to confirm our findings on an all-words domain-specific corpus, such as the one planned for SemEval-2010.

Acknowledgements

Our thanks to Rob Koeling and Diana McCarthy for kindly providing the dataset, thesauri and assistance. Oier enjoys a grant from the Basque Government. This work has been partially funded by the EU Commission (FP7 project KYOTO) and Spanish Research Department (project KNOW).

References

[Agirre and Lopez de Lacalle, 2008] E. Agirre and O. Lopez de Lacalle. 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 2008.

[Agirre and Lopez de Lacalle, 2009] E. Agirre and O. Lopez de Lacalle. Supervised domain adaption for wsd. In *Proceedings of EACL-09*, Athens, Greece, 2009.

[Agirre and Soroa, 2009] E. Agirre and A. Soroa. Personalizing pagerank for word sense disambiguation. In *Proceedings of EACL-09*, Athens, Greece, 2009.

[Brin and Page, 1998] S. Brin and L. Page. The anatomy of a large-scale hypertextual web search engine. *Computer Networks and ISDN Systems*, 30(1-7), 1998.

[Chan and Ng, 2007] Y. S. Chan and H. T. Ng. 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 2007.

[Escudero *et al.*, 2000] G. Escudero, L. Márquez, and G. Rigau. An Empirical Study of the Domain Dependence of Supervised Word Sense Didanbiguation Systems. *Proceedings of the joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora, EMNLP/VLC*, 2000.

[Haveliwala, 2002] T. H. Haveliwala. Topic-sensitive pagerank. In *WWW '02: Proceedings of the 11th international conference on World Wide Web*, pages 517–526, New York, NY, USA, 2002. ACM.

[Hughes and Ramage, 2007] T. Hughes and D. Ramage. Lexical semantic relatedness with random graph walks. In *Proceedings of EMNLP-CoNLL-2007*, pages 581–589, 2007.

[Koeling *et al.*, 2005] R. Koeling, D. McCarthy, and J. Carroll. 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, 2005.

[Lin, 1998] D. Lin. Automatic retrieval and clustering of similar words. In *Proceedings of the 36th Annual Meeting of the Association of Computational Linguistics (ACL98) - joint with Computational Linguistics (Coling98)*, Montreal, Canada, August 1998.

[Miller *et al.*, 1993] G.A. Miller, C. Leacock, R. Tengi, and R.Bunker. 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, 1993.

[Navigli and Lapata, 2007] R. Navigli and M. Lapata. Graph connectivity measures for unsupervised word sense disambiguation. In *IJCAI*, 2007.

[Pradhan *et al.*, 2007] S. Pradhan, E. Loper, D. Dligach, and M.Palmer. Semeval-2007 task-17: English lexical sample srl and all words. In *Proceedings of SemEval-2007*, pages 87–92, Prague, Czech Republic, June 2007.

[Sinha and Mihalcea, 2007] R. Sinha and R. Mihalcea. Unsupervised graph-based word sense disambiguation using measures of word semantic similarity. In *Proceedings of the IEEE International Conference on Semantic Computing (ICSC 2007)*, Irvine, CA, USA, 2007.

[Snyder and Palmer, 2004] B. Snyder and M. Palmer. The English all-words task. In *Proceedings of the 3rd ACL workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL)*, Barcelona, Spain, 2004.