On Robustness and Domain Adaptation using SVD for Word Sense Disambiguation

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

Abstract

In this paper we explore robustness and domain adaptation issues for Word Sense Disambiguation (WSD) using Singular Value Decomposition (SVD) and unlabeled data. We focus on the semi-supervised domain adaptation scenario, where we train on the source corpus and test on the target corpus, and try to improve results using unlabeled data. Our method yields up to 16.3% error reduction compared to state-of-the-art systems, being the first to report successful semi-supervised domain adaptation. Surprisingly the improvement comes from the use of unlabeled data from the source corpus, and not from the target corpora, meaning that we get robustness rather than domain adaptation. In addition, we study the behavior of our system on the target domain.

1 Introduction

In many Natural Language Processing (NLP) tasks we find that a large collection of manually-annotated text is used to train and test supervised machine learning models. While these models have been shown to perform very well when tested on the text collection related to the training data (what we call the **source** domain), the performance drops considerably when testing on text from other domains (called **target** domains).

In order to build models that perform well in new (target) domains we usually find two settings (Daumé III, 2007): In the **semi-supervised** setting the goal is to improve the system trained on the source domain using unlabeled data from the target domain, and the baseline is that of the system trained on the source domain. In the **supervised setting**, training data from both source and target domains are used, and the baseline is provided by the system trained on the target domain. The semi-supervised setting is the most attractive, as it would save developers the need to hand-annotate target corpora every time a new domain is to be processed.

The main goal of this paper is to use unlabeled data in order to get better domain-adaptation results for Word Sense Disambiguation (WSD) in the semi-supervised setting. Singular Value Decomposition (SVD) has been shown to find correlations between terms which are helpful to overcome the scarcity of training data in WSD (Gliozzo et al., 2005). This paper explores how this ability of SVD can be applied to the domain-adaptation of WSD systems, and we show that SVD and unlabeled data improve the results of two state-of-the-art WSD systems (k-NN and SVM). For the sake of this paper we call this set of experiments the **domain adaptation scenario**.

In addition, we also perform some related experiments on just the target domain. We use unlabeled data in order to improve the results of a system trained and tested in the target domain. These results are complementary to the domain adaptation experiments, and also provide an upperbound for semi-supervised domain adaptation. We call these experiments the **target domain scenario**. Note that both scenarios are semi-supervised, in that our focus is on the use of unlabeled data in addition to the available labeled data.

The experiments were performed on a publicly available corpus which was designed to study the effect of domain in WSD (Koeling et al., 2005). It comprises 41 nouns closely related to the SPORTS and FINANCES domains with 300 examples for each. The 300 examples were drawn from the British National Corpus (Leech, 1992) (BNC), the SPORTS section of the Reuters corpus (Leech,

Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008), pages 17–24
Manchester, August 2008

1992), and the FINANCES section of Reuters in equal number.

The paper is structured as follows. Section 2 reviews prior work in the area. Section 3 presents the datasets used, and Section 4 the learning methods, including the application of SVD. The experimental results are presented in Section 5, for the semi-supervised domain adaptation scenario, and Section 6, for the target scenario. Section 7 presents the discussion and Section 8 the conclusions and future work.

2 Prior Work

Domain adaptation is a subject attracting more and more attention. In the semi-supervised setting, Blitzer et al. (2006) use 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. Our technique also uses SVD, but we directly apply it to all features, and thus avoid the need to define pivot features. In preliminary work we unsuccessfully tried to carry along the idea of pivot features to WSD. Zelikovitz and Hirsh (2001) use unlabeled data (so-called background knowledge) with Latent Semantic Indexing (also based on SVD) on a Text Classification task with positive results. They use related unlabeled text and include it in the term-by-document matrix to expand it and capture better the interesting properties of the data. Their approach is similar to our SMA method in Section 4.2).

In the supervised setting, a recent paper by Daumé III (2007) shows that, using a very simple feature augmentation method coupled with Support Vector Machines, he is able to effectively use both labeled target and source data to provide the best 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).

Regarding WSD, some initial works made basic analysis of the particular issues. Escudero et al. (2000) tested the supervised adaptation setting 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) also used the DSO corpus in the supervised setting to show that training on a subset of the source corpora that is topically related to the target corpus does allow for some domain adaptation. Their work used the fact that the genre tags of Brown allowed to detect which parts of the corpus were related to the target corpus.

More recently, Koeling et al. (2005) presented an unsupervised system to learn the predominant senses of particular domains. Their system was based on the use of a similarity thesaurus induced from the domain corpus and WordNet. They used the same dataset as in this paper for evaluation. 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 allowed to significantly improve results when only 10%-30% of the target corpus was used for training. No data was given about the use of both tagged corpora.

Though not addressing domain adaptation, other works on WSD also used SVD and are closely related to the present paper. Gliozzo et al. (2005) used SVD to reduce the space of the term-to-document matrix, and then computed the similarity between train and test instances using a mapping to the reduced space (similar to our SMA method in Section 4.2). They combined other knowledge sources into a complex kernel using SVM. They report improved performance on a number of languages in the Senseval-3 lexical sample dataset. Our present paper differs from theirs in that we propose an additional method to use SVD (the OMT method, Section 4.2), and that we evaluate the contribution of unlabeled data and SVD in isolation, leaving combination for future work.

Ando (2006) used Alternative Structured Optimization, which is closely related to Structural Learning (cited above). He first trained one linear predictor for each target word, and then performed SVD on 7 carefully selected submatrices

of the feature-to-predictor matrix of weights. The system attained small but consistent improvements (no significance data was given) on the Senseval-3 lexical sample datasets using SVD and unlabeled data.

We have previously shown (Agirre et al., 2005; Agirre and Lopez de Lacalle, 2007) that performing SVD on the feature-to-documents matrix is a simple technique that allows to improve performance with and without unlabeled data. The use of several k-NN classifiers trained on a number of reduced and original spaces was shown to rank first in the Senseval-3 dataset and second in the SemEval 2007 competition. The present work extends our own in that we present a comprehensive study on a domain adaptation dataset, producing additional insight on our method and the relation between SVD, features and unlabeled data.

3 Data sets

The dataset we use was designed for domain-related WSD experiments by Koeling et al. (2005), and is publicly available. The examples come from the BNC (Leech, 1992) and the SPORTS and FINANCES sections of the 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 (WN) version 1.7.1 (Fellbaum, 1998).

Compared to the DSO corpus used in prior work (cf. Section 2) this corpus has been explicitly created for domain adaptation studies. DSO contains texts coming from the Brown corpus and the Wall Street Journal, but the texts are not classified according to specific domains (e.g. Sports, Finances), which make DSO less suitable to study domain adaptation.

In addition to the labeled data, we also use unlabeled data coming from the three sources used in the labeled corpus: the 'written' part of the BNC (89.7M words), the FINANCES part of Reuters (117,734 documents, 32.5M words), and the SPORTS part (35,317 documents, 9.1M words).

4 Learning features and methods

In this section, we review the learning features, the two methods to apply SVD, and the two learning algorithms used in the experiments.

4.1 Learning features

We relied on the usual features used in previous WSD work, grouped in three main sets. **Local collocations** comprise the bigrams and trigrams formed around the target word (using either lemmas, word-forms, and PoS tags[1]), those formed with the previous/posterior lemma/word-form in the sentence, and the content words in a $\pm$4-word window around the target. **Syntactic dependencies**[2] use the object, subject, noun-modifier, preposition, and sibling lemmas, when available. Finally, **Bag-of-words features** are the lemmas of the content words in the whole context, plus the salient bigrams in the context (Pedersen, 2001).

4.2 Features from the reduced space

Apart from the original space of features, we have the so called **SVD features**, obtained from the projection of the feature vectors into the reduced space (Deerwester et al., 1990). Basically, we set a term-by-document or feature-by-example matrix M from the corpus (see section below for more details). SVD decomposes it into three matrices, $M = U\Sigma V^T$. If the desired number of dimensions in the reduced space is p, we select p rows from Σ and V, yielding Σ_p and V_p respectively. We can map any feature vector $\vec{t}$ (which represents either a train or test example) into the p-dimensional space as follows: $\vec{t_p} = \vec{t}^T V_p \Sigma_p^{-1}$. Those mapped vectors have p dimensions, and each of the dimensions is what we call a SVD feature. We can now use the mapped vectors ($\vec{t_p}$) to train and test any learning method, as usual. We have explored two different variants in order to build the reduced matrix and obtain the SVD features, as follows.

Single Matrix for All target words (SVD-SMA). The method comprises the following steps: (i) extract bag-of-word features (terms in this case) from unlabeled corpora, (ii) build the term-by-document matrix, (iii) decompose it with SVD, and (iv) project the labeled data (train/test). This technique is very similar to previous work on SVD (Gliozzo et al., 2005; Zelikovitz and Hirsh, 2001). The dimensionality reduction is performed once, over the whole unlabeled corpus, and it is then applied to the labeled data of each word. The reduced

[1] The PoS tagging was performed with the fnTBL toolkit (Ngai and Florian, 2001)

[2] This software was kindly provided by David Yarowsky's group, from Johns Hopkins University.

space is constructed only with terms, which correspond to bag-of-words features, and thus discards the rest of the features. Given that the WSD literature has shown that all features, including local and syntactic features, are necessary for optimal performance (Pradhan et al., 2007), we propose the following alternative to construct the matrix.

One Matrix per Target word (SVD-OMT). For each word: (i) construct a corpus with its occurrences in the labeled and, if desired, unlabeled corpora, (ii) extract all features, (iii) build the feature-by-example matrix, (iv) decompose it with SVD, and (v) project all the labeled training and test data for the word. Note that this variant performs one SVD process for each target word separately, hence its name. We proposed this technique in (Agirre et al., 2005).

An important parameter when doing SVD is the number of dimensions in the reduced space (p). We tried two different values for p (25 and 200) in the BNC domain, and the results were consistent in that 25 performed better for SVD-OMT and 200 better for SVD-SMA. Those values were chosen for testing in the SPORTS and FINANCES domains, i.e. 25 for SVD-OMT and 200 for SVD-SMA.

4.3 Building Matrices

The methods in the previous section can be applied to the following matrices M:

- TRAIN: The matrix comprises features from labeled train examples alone. This matrix can only be used to obtain OMT features.
- TRAIN $\cup$ BNC: In addition to TRAIN, we matrix also includes unlabeled examples from the source corpus (BNC). Both OMT and SMA features can be obtained.
- TRAIN $\cup$ {SPORTS,FINANCES}: Like the previous, but using unlabeled examples from one of the target corpora (FINANCES or SPORTS) instead. Both OMT and SMA feature can be obtained.

Based on previous work (Agirre et al., 2005), we used 50% of the respective unlabeled corpora for OMT features, and the whole corpora for SMA.

4.4 Learning methods

We used two well known classifiers, Support Vector Machines (SVM) and k-Nearest Neighbors (k-NN). Regarding SVM we used linear kernels implemented in SVM-Light (Joachims, 1999). We estimated the soft margin (C) for each feature space and each word using a greedy process in a preliminary experiment on the source training data using cross-validation. The same C value was used in the rest of the settings.

k-NN is a memory based learning method, where the neighbors are the k most similar labeled examples to the test example. The similarity among instances is measured by the cosine of their vectors. The test instance is labeled with the sense obtaining the maximum the sum of the weighted vote of the k most similar contexts. We set k to 5 based on previous results (Agirre and Lopez de Lacalle, 2007).

5 Domain adaptation scenario

In this scenario we try to adapt a general purpose supervised WSD system trained on the source corpus (BNC) to a target corpus (either SPORTS or FINANCES) using unlabeled corpora only.

5.1 Experimental results

Table 1 shows the precision results for this scenario. Note that all methods have full coverage, i.e. they return a sense for all test examples, and therefore precision suffices to compare among systems. We have computed significance ranges for all results in this paper using bootstrap resampling (Noreen, 1989). F_1 scores outside of these intervals are assumed to be significantly different from the related F_1 score ($p < 0.05$).

The table has two main parts, each regarding to one of the target domains, SPORTS and FINANCES. The use of two target domains allows to test whether the methods behave similarly in both domains. The columns denote the classifier and SVD method used: the MFS column corresponds to the most frequent sense, k-NN-ORIG (SVM-ORIG) corresponds to performing k-NN (SVM) on the original feature space, k-NN-OMT (SVM-OMT) corresponds to k-NN (SVM) on the reduced dimensions of the OMT strategy, and k-NN-SMA (SVM-SMA) corresponds to k-NN (SVM) on the reduced dimensions of the SMA strategy (cf. Section 4.2). The rows correspond to the matrix used for SVD (cf. Section 4.3). Note that some of the cells have no result, because that combination is not applicable, e.g. using the TRAIN $\cup$ BNC in the original space.

In the first row (TRAIN) of Table 1 we can see that in both domains SVM on the original space outperforms k-NN with statistical signifi-

BNC → SPORTS							
matrix configuration	MFS	k-NN-ORIG	k-NN-OMT	k-NN-SMA	SVM-ORIG	SVM-OMT	SVM-SMA
TRAIN	39.0±1.3	51.7±1.3	53.0±1.6	-	53.9±1.3	47.4±1.5	-
TRAIN ∪ SPORTS	-	-	47.8±1.5	49.7±1.5	-	51.8±1.5	53.8±1.5
TRAIN ∪ BNC	-	-	**61.4**±1.4	57.1±1.5	-	57.1±1.6	57.2±1.5
BNC → FINANCES							
matrix configuration	MFS	k-NN-ORIG	k-NN-OMT	k-NN-SMA	SVM-ORIG	SVM-OMT	SVM-SMA
TRAIN	51.2±1.6	60.4±1.6	62.5±1.4	-	62.9±1.6	59.4±1.5	-
TRAIN ∪ FINANCES	-	-	57.4±1.9	60.6±1.5	-	60.4±1.4	62.7±1.4
TRAIN ∪ BNC	-	-	65.9±1.5	**68.3**±1.4	-	67.0±1.3	66.8±1.5

Table 1: Precision for the domain adaptation scenario: training on labeled source corpus, plus unlabeled corpora.

cance. Those are the baseline systems. On the same row, working on the reduced space of the TRAIN matrix with OMT allows to improve the results of k-NN, but not for SVM.

Contrary to our expectations, adding target unlabeled corpora (TRAIN ∪ SPORTS and TRAIN ∪ FINANCES rows respectively) does not improve the results over the baseline. But using the source unlabeled data (TRAIN ∪ BNC), we find that for both domains and in all four columns the results are significantly better than for the best baseline in both SPORTS and FINANCES corpora.

The best results on the TRAIN ∪ BNC row depend on the domain corpus. While k-NN-OMT obtains the best results for SPORTS, in FINANCES k-NN-SMA is best. k-NN, in principle a weaker method that SVM, is able to attain the same or superior performance than SVM on the reduced spaces.

Table 3 summarizes the main results, and also shows the error reduction figures, which range between 6.9% and 16.3%. As the most important conclusion, we want to stress that, in this scenario, we are able to build a very robust system just adding unlabeled source material, and that we fail to adapt to the domain using the target corpus. These results are relevant to improve a generic WSD system to be more robust when ported to new domains.

5.2 Controlling size

In the original experiments reported in the previous sections, the size of the unlabeled corpora was not balanced. Due to the importance of the amount of unlabeled data, we performed two control experiments for the OMT and SMA matrices on the domain adaptation scenario, focusing on the k-NN method. Regarding OMT, we used the minimum number of instances per word between BNC and each of the target domains. The system obtained 60.0 of precision using unlabeled data from BNC and 49.5 for SPORTS data (compared to 61.4 and 47.8 in table 1, respectively). We did the same in the FINANCES domain, and we obtained 65.6 of precision for BNC and 54.4 for FINANCES (compared to 65.7 and 57.4 in table 1, respectively). Although the contribution of BNC unlabeled data is slightly lower in this experiment, due to the smaller amount of data, it still outperforms the target unlabeled data by a large margin.

In the case of the SMA matrix, we used 25% of the BNC, which is comparable to the SPORTS and FINANCES sizes. The results, 56.9 of precision in SPORTS domain and 68.1 in FINANCES (compared to 57.1 and 68.3 in table 1, respectively), confirm that the size is not an important factor for SMA either.

6 Target scenario

In this second scenario we focus on the target domain. We train and test on the target domain, and use unlabeled data in order to improve the result. The goal of these experiments is to check the behavior of our method when applied to the target domain, in order to better understand the results on the domain adaptation scenario. They also provide an upperbound for semi-supervised domain adaptation.

6.1 Experimental results

The results are presented in table 2. All experiments in this section have been performed using 3-fold cross-validation. Again, we have full coverage in all cases, and the significance ranges correspond to the 95% confidence level. The table has two main parts, each regarding to one of the target domains, SPORTS and FINANCES. As in Table 1, the columns specify the classifier and SVD method used, and the rows correspond to the matrices used

SPORTS → SPORTS (*xval*)							
matrix configuration	MFS	k-NN-ORIG	k-NN-OMT	k-NN-SMA	SVM-ORIG	SVM-OMT	SVM-SMA
TRAIN	77.8±1.2	84.5±1.0	85.0±1.1	-	85.1±1.0	81.0±1.5	-
TRAIN ∪ SPORTS	-	-	**86.1**±0.9	82.7±1.1	-	85.1±1.1	80.3±1.5
TRAIN ∪ BNC	-	-	84.4±1.0	80.4±1.5	-	84.3±0.9	79.8±1.2
FINANCES → FINANCES (*xval*)							
matrix configuration	MFS	k-NN-ORIG	k-NN-OMT	k-NN-SMA	SVM-ORIG	SVM-OMT	SVM-SMA
TRAIN	82.3±1.3	87.1±1.0	87.4±1.0	-	87.0±1.0	85.5±1.1	-
TRAIN ∪ SPORTS	-	-	**87.8**±0.8	84.3±1.4	-	86.4±0.9	82.9±1.1
TRAIN ∪ BNC	-	-	87.4±1.2	83.5±1.2	-	85.7±0.9	84.3±1.1

Table 2: Precision for the target scenario: training on labeled target corpora, plus unlabeled corpora.

to obtain the features.

Table 2 shows that k-NN-OMT using the target corpus (SPORTS and FINANCES, respectively) slightly improves over the k-NN-ORIG and SVM-ORIG classifiers, with significant difference in the SPORTS domain. Contrary to the results on the previous section, the source unlabeled corpus degrades performance, but the target corpus does allow for small improvements. Note that, in this scenario, both SVM and k-NN perform similarly in the original space, but only k-NN is able to profit from the reduced space. Table 3 summarizes the best result, alongside the error reduction.

The results of these experiments allow to contrast both scenarios, and to get deeper insight about the relation between the labeled and unlabeled data when performing SVD, as we will examine in the next section.

7 Discussion

The main contribution of this paper is to show that we obtain robustness when faced with domain shifts using a semi-supervised strategy. We show that we can obtain it using a large, general, unlabeled corpus. Note that our semi-supervised method to attain robustness for domain shifts is very cost-effective, as it does not require costly hand-tagged material nor even large numbers of unlabeled data from each target domain. These results are more valuable given the lack of substantial positive results on the literature on semi-supervised or supervised domain adaptation for WSD (Escudero et al., 2000; Martínez and Agirre, 2000; Chan and Ng, 2007).

Compared to other settings, our semi-supervised results improve over the completely unsupervised system in (Koeling et al., 2005), which had 43.7% and 49.9% precision for the SPORTS and FINANCES domains respectively, but lag well behind the target domain scenario, showing that there is still room for improvement in the semi-supervised setting.

While these results are based on a lexical sample, and thus not directly generalizable to an all-words corpus, we think that they reflect the main trends for nouns, as the 41 nouns where selected among those exhibiting domain dependence (Koeling et al., 2005). We can assume, though it would be needed to be explored empirically, that other nouns exhibiting domain independence would degrade less when moving to other domains, and thus corroborate the robustness effect we have discovered.

The fact that we attain robustness rather than domain adaptation proper deserves some analysis. In the domain adaptation scenario only source unlabeled data helped, but the results on the target scenario show that it is the target unlabeled data which is helping, and not the source one. Given that SVD basically finds correlations among features, it seems that constructing the term-by-document (or feature-by-example) matrix with the training data and the unlabeled corpus related to the training data is the key factor in play here.

The reasons for this can be traced back as follows. Our source corpus is the BNC, which is a balanced corpus containing a variety of genres and domains. The 100 examples for each word that have been hand-tagged were gathered at random, and thus cover several domains. For instance, the OMT strategy for building the matrix extracts hundreds of other examples from the BNC, and when SVD collapses the features into a reduced space, it effectively captures the most important correlations in the feature-by-example matrix. When faced with examples from a new domain, the reduced matrix is able to map some of the features found in the test example to those in the train example. Such overlap is more difficult if only 100 examples from the source domain are available.

SPORTS	FINANCES	sign.	E.R (%)	method
53.9±1.3	62.9±1.6	-	-	labeled source (SVM-ORIG: baseline)
57.1±1.5	**68.3±1.4**	++	6.9/14.5	**labeled source + SVD on unlabeled source (k-NN-SMA)**
61.4±1.4	65.9±1.5	++	16.3/8.1	**labeled source + SVD on unlabeled source (k-NN-OMT)**
85.1±1.0	87.0±1.0	-	-	labeled target (SVM-ORIG: baseline)
86.1±0.9	**87.8±0.8**	+	6.7/6.1	**labeled target + SVD on unlabeled target (k-NN-OMT)**

Table 3: Summary with the most important results for the two scenarios (best results for each in bold). The significance column shows significance over baselines: ++ (significant in both target domains), + (significant in a single domain). The E.R column shows the error reduction in percentages over the baseline methods.

The unlabeled data and SVD process allow to capture correlations among the features occurring in the test data and those in the training data.

On the other hand, we are discarding all original features, as we focus on the features from the reduced space alone. The newly found correlations come at the price of possibly ignoring effective original features, causing information loss. Only when the correlations found in the reduced space outweigh this information loss do we get better performance on the reduced space than in the original space. The experiment in Section 6 is important in that it shows that the improvement is much smaller and only significant in the target domain scenario, which is in accordance with the hypothesis above. This information loss is a motivation for the combination of the features from the reduced space with the original features, which will be the focus of our future work.

Regarding the learning method and the two strategies to apply SVD, the results show that k-NN profits from the reduced spaces more than SVM, even if its baseline performance is lower than SVM. Regarding the matrix building system, in the domain adaptation scenario, k-NN-OMT obtains the best results (with statistical significance) in the SPORTS corpus, and k-NN-SMA yields the best results (with statistical significance) in the FINANCES domain. Averaging over both domains, k-NN-OMT is best. The target scenario results confirm this trend, as k-NN-OMT is superior to k-NN-SMA in both domains. These results are in accordance with our previous experience on WSD (Agirre et al., 2005), where our OMT method got better results than SMA and those of (Gliozzo et al., 2005) (who also use a method similar to SMA) on the Senseval-3 lexical sample. While OMT reduces the feature-by-example matrix of each target word, SMA reduces a single term-by-document matrix. SMA is able to find important correlations among similar terms in the corpus, but it misses the rich feature set used by WSD systems, as it focuses on bag-of-words alone. OMT on the other hand is able to find correlations between all features which are relevant to the target word only.

8 Conclusions and Future Work

In this paper we explore robustness and domain adaptation issues for Word Sense Disambiguation using SVD and unlabeled data. We focus on the semi-supervised scenario, where we train on the source corpus (BNC), test on two target corpora (SPORTS and FINANCES sections of Reuters), and improve the results using unlabeled data.

Our method yields up to 16.3% error reduction compared to SVM and k-NN on the labeled data alone, showing the first positive results on domain adaptation for WSD. In fact, we show that our results are due to the use of a large, general, unlabeled corpus, and rather than domain-adaptation proper we show robustness in face of a domain shift. This kind of robustness is even more cost-effective than semi-supervised domain adaptation, as it does not require large unlabeled corpora and repeating the computations for each new target domain.

This paper shows that the OMT technique to apply SVD that we proposed in (Agirre et al., 2005) compares favorably to SMA, which has been previously used in (Gliozzo et al., 2005), and that k-NN excels SVM on the features from the reduced space. We also show that the unlabeled data needs to be related to the training data, and that the benefits of our method are larger when faced with a domain shift (compared to test data coming from the same domain as the training data).

In the future, we plan to combine the features from the reduced space with the rest of features, either using a combination of k-NN classifiers (Agirre et al., 2005; Agirre and Lopez de Lacalle, 2007) or a complex kernel (Gliozzo et al., 2005).

A natural extension of our work would be to apply our techniques to the supervised domain adaptation scenario.

Acknowledgments

We wish to thank Diana McCarthy and Rob Koeling for kindly providing us the Reuters tagged corpora, David Martínez for helping us with the learning features, and Walter Daelemans for his advice on domain adaptation. Oier Lopez de Lacalle has a PhD grant from the Basque Government. This work is partially funded by the Education Ministry (KNOW TIN2006-15049, OpenMT TIN2006-15307-C03-02) and the Basque Country University (IT-397-07).

References

Agirre, E. and O. Lopez de Lacalle. 2007. UBC-ALM: Combining k-NN with SVD for WSD. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*. Association for Computational Linguistics.

Agirre, E., O. Lopez de Lacalle, and D. Martínez. 2005. Exploring feature spaces with svd and unlabeled data for Word Sense Disambiguation. In *Proceedings of the Conference on Recent Advances on Natural Language Processing (RANLP'05)*.

Ando, R. Kubota. 2006. Applying alternating structure optimization to word sense disambiguation. In *Proceedings of the 10th Conference on Computational Natural Language Learning (CoNLL)*.

Blitzer, J., R. McDonald, and F. Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*.

Chan, Yee Seng 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*.

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

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

Daumé III, H. 2007. Frustratingly easy domain adaptation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*.

Deerwester, S., S. Dumais, G. Furnas, T. Landauer, and R. Harshman. 1990. Indexing by Latent Semantic Analysis. *Journal of the American Society for Information Science*.

Escudero, G., L. Márquez, and G. Rigau. 2000. 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*.

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

Gliozzo, A. M., C. Giuliano, and C. Strapparava. 2005. Domain Kernels for Word Sense Disambiguation. *43nd Annual Meeting of the Association for Computational Linguistics. (ACL-05)*.

Joachims, T. 1999. Making Large–Scale SVM Learning Practical. *Advances in Kernel Methods — Support Vector Learning*, Cambridge, MA. MIT Press.

Koeling, R., 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*.

Leech, G. 1992. 100 million words of English: the British National Corpus. *Language Research*.

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

Ngai, G. and R. Florian. 2001. Transformation-Based Learning in the Fast Lane. *Proceedings of the Second Conference of the North American Chapter of the Association for Computational Linguistics*.

Noreen, E. W. 1989. *Computer-Intensive Methods for Testing Hypotheses*. John Wiley & Sons.

Pedersen, T. 2001. A Decision Tree of Bigrams is an Accurate Predictor of Word Sense. In *Proceedings of the Second Meeting of the North American Chapter of the Association for Computational Linguistics (NAACL-01)*.

Pradhan, S., E. Loper, D. Dligach, and M. 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)*.

Rose, T. G., M. Stevenson, and M. 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)*.

Zelikovitz, S. and H. Hirsh. 2001. Using LSI for text classification in the presence of background text. In *Proceedings of CIKM-01, 10th ACM International Conference on Information and Knowledge Management*. US.