All Words Domain Adapted WSD: Finding a Middle Ground between Supervision and Unsupervision

Mitesh M. Khapra Anup Kulkarni Saurabh Sohoney Pushpak Bhattacharyya
Indian Institute of Technology Bombay,
Mumbai - 400076, India.
{miteshk,anup,saurabhsohoney,pb}@cse.iitb.ac.in

Abstract

In spite of decades of research on word sense disambiguation (WSD), all-words general purpose WSD has remained a distant goal. Many supervised WSD systems have been built, but the effort of creating the training corpus - *annotated sense marked corpora* - has always been a matter of concern. Therefore, attempts have been made to develop unsupervised and knowledge based techniques for WSD which do not need sense marked corpora. However such approaches have not proved effective, since they typically do not better Wordnet first sense baseline accuracy. Our research reported here proposes to stick to the supervised approach, but with far less demand on annotation. We show that if we have ANY sense marked corpora, be it from mixed domain or a specific domain, a small amount of annotation in ANY other domain can deliver the goods almost as if exhaustive sense marking were available in that domain. We have tested our approach across Tourism and Health domain corpora, using also the well known mixed domain SemCor corpus. Accuracy figures close to self domain training lend credence to the viability of our approach. Our contribution thus lies in finding a convenient middle ground between pure supervised and pure unsupervised WSD. Finally, our approach is not restricted to any specific set of target words, a departure from a commonly observed practice in domain specific WSD.

1 Introduction

Amongst annotation tasks, sense marking surely takes the cake, demanding as it does high level of language competence, topic comprehension and domain sensitivity. This makes supervised approaches to WSD a difficult proposition (Agirre et al., 2009b; Agirre et al., 2009a; McCarthy et al., 2007). Unsupervised and knowledge based approaches have been tried with the hope of creating WSD systems with no need for sense marked corpora (Koeling et al., 2005; McCarthy et al., 2007; Agirre et al., 2009b). However, the accuracy figures of such systems are low.

Our work here is motivated by the desire to develop *annotation-lean all-words* domain adapted techniques for supervised WSD. It is a common observation that 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. Also domain adaptation - in which training happens in one domain and testing in another - often is able to attain good levels of performance, albeit on a specific set of target words (Chan and Ng, 2007; Agirre and de Lacalle, 2009). To the best of our knowledge there does not exist a system that solves the combined problem of *all words domain adapted WSD*. We thus propose the following:

a. For any target domain, create a small amount of sense annotated corpus.

b. Mix it with an existing sense annotated corpus – from a mixed domain or specific domain – to train the WSD engine.

This procedure tested on four adaptation scenarios, *viz.*, (i) SemCor (Miller et al., 1993) to Tourism, (ii) SemCor to Health, (iii) Tourism to Health and (iv) Health to Tourism has consistently yielded good performance (to be explained in sections 6 and 7).

The remainder of this paper is organized as follows. In section 2 we discuss previous work in the area of domain adaptation for WSD. In section 3

we discuss three state of art supervised, unsupervised and knowledge based algorithms for WSD. Section 4 discusses the injection strategy for domain adaptation. In section 5 we describe the dataset used for our experiments. We then present the results in section 6 followed by discussions in section 7. Section 8 examines whether there is any need for intelligent choice of injections. Section 9 concludes the paper highlighting possible future directions.

2 Related Work

Domain specific WSD for selected target words has been attempted by Ng and Lee (1996), Agirre and de Lacalle (2009), Chan and Ng (2007), Koeling et al. (2005) and Agirre et al. (2009b). They report results on three publicly available lexical sample datasets, *viz.*, DSO corpus (Ng and Lee, 1996), MEDLINE corpus (Weeber et al., 2001) and the corpus made available by Koeling et al. (2005). Each of these datasets contains a handful of target words (41-191 words) which are sense marked in the corpus.

Our main inspiration comes from the target-word specific results reported by Chan and Ng (2007) and Agirre and de Lacalle (2009). The former showed that adding just 30% of the target data to the source data achieved the same performance as that obtained by taking the entire source and target data. Agirre and de Lacalle (2009) reported a 22% error reduction when source and target data were combined for training a classifier, as compared to the case when only the target data was used for training the classifier. However, both these works focused on *target word specific* WSD and do not address all-words domain specific WSD.

In the unsupervised setting, McCarthy et al. (2007) showed that their predominant sense acquisition method gives good results on the corpus of Koeling et al. (2005). In particular, they showed that the performance of their method is comparable to the most frequent sense obtained from a tagged corpus, thereby making a strong case for unsupervised methods for domain-specific WSD. More recently, Agirre et al. (2009b) showed that knowledge based approaches which rely only on the semantic relations captured by the Wordnet graph outperform supervised approaches when applied to specific domains. The good results obtained by McCarthy et al. (2007) and Agirre et al. (2009b) for unsupervised and knowledge based approaches respectively have cast a doubt on the viability of supervised approaches which rely on sense tagged corpora. However, these conclusions were drawn only from the performance on certain target words, leaving open the question of their utility in all words WSD.

We believe our work contributes to the WSD research in the following way: (i) it shows that there is promise in supervised approach to all-word WSD, through the instrument of domain adaptation; (ii) it places in perspective some very recently reported unsupervised and knowledge based techniques of WSD; (ii) it answers some questions arising out of the debate between supervision and unsupervision in WSD; and finally (iv) it explores a convenient middle ground between unsupervised and supervised WSD – the territory of "annotate-little and inject" paradigm.

3 WSD algorithms employed by us

In this section we describe the knowledge based, unsupervised and supervised approaches used for our experiments.

3.1 Knowledge Based Approach

Agirre et al. (2009b) showed that a graph based algorithm which uses only the relations between concepts in a Lexical Knowledge Base (LKB) can outperform supervised approaches when tested on specific domains (for a set of chosen target words). We employ their method which involves the following steps:

1. Represent Wordnet as a graph where the concepts (*i.e.*, synsets) act as nodes and the relations between concepts define edges in the graph.
2. Apply a context-dependent *Personalized PageRank* algorithm on this graph by introducing the context words as nodes into the graph and linking them with their respective synsets.
3. These nodes corresponding to the context words then inject probability mass into the synsets they are linked to, thereby influencing the final relevance of all nodes in the graph.

We used the publicly available implementation of this algorithm[1] for our experiments.

[1] http://ixa2.si.ehu.es/ukb/

3.2 Unsupervised Approach

McCarthy et al. (2007) used an untagged corpus to construct a thesaurus of related words. They then found the predominant sense (i.e., the most frequent sense) of each target word using pair-wise Wordnet based similarity measures by pairing the target word with its top-k neighbors in the thesaurus. Each target word is then disambiguated by assigning it its predominant sense – the motivation being that the predominant sense is a powerful, hard-to-beat baseline. We implemented their method using the following steps:

1. Obtain a domain-specific untagged corpus (we crawled a corpus of approximately 9M words from the web).
2. Extract grammatical relations from this text using a dependency parser[2] (Klein and Manning, 2003).
3. Use the grammatical relations thus extracted to construct features for identifying the k nearest neighbors for each word using the distributional similarity score described in (Lin, 1998).
4. Rank the senses of each target word in the test set using a weighted sum of the distributional similarity scores of the neighbors. The weights in the sum are based on Wordnet Similarity scores (Patwardhan and Pedersen, 2003).
5. Each target word in the test set is then disambiguated by simply assigning it its predominant sense obtained using the above method.

3.3 Supervised approach

Khapra et al. (2010) proposed a supervised algorithm for domain-specific WSD and showed that it beats the most frequent corpus sense and performs on par with other state of the art algorithms like PageRank. We implemented their iterative algorithm which involves the following steps:

1. Tag all monosemous words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in increasing order of their degree of polysemy.
3. At each stage rank the candidate senses of a word using the scoring function of Equation (1) which combines corpus based parameters (such as, sense distributions and corpus co-occurrence) and Wordnet based parameters (such as, semantic similarity, conceptual distance, *etc.*)

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

where,

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

4. Select the candidate synset with maximizes the above score as the winner sense.

4 Injections for Supervised Adaptation

This section describes the main interest of our work i.e. *adaptation using injections*. For supervised adaptation, we use the supervised algorithm described above (Khapra et al., 2010) in the following 3 settings as proposed by Agirre et al. (2009a):

a. **Source setting:** We train the algorithm on a mixed-domain corpus (SemCor) or a domain-specific corpus (say, Tourism) and test it on a different domain (say, Health). A good performance in this setting would indicate robustness to domain-shifts.

b. **Target setting:** We train and test the algorithm using data from the same domain. This gives the skyline performance, i.e., the best performance that can be achieved if sense marked data from the target domain were available.

c. **Adaptation setting:** This setting is the main focus of interest in the paper. We augment the training data which could be from one domain or mixed domain with a small amount of data from the target domain. This combined data is then used for training. The aim here is to reach as close to the skyline performance using as little data as possible. For injecting data from the target domain we randomly select some sense marked words from the target domain and add

[2]We used the Stanford parser - http://nlp.stanford.edu/software/lex-parser.shtml

Category	Polysemous words		Monosemous words	
	Tourism	Health	Tourism	Health
Noun	53133	15437	23665	6979
Verb	15528	7348	1027	356
Adjective	19732	5877	10569	2378
Adverb	6091	1977	4323	1694
All	94484	30639	39611	11407

Table 1: Polysemous and Monosemous words per category in each domain

Category	Avg. no. of instances per polysemous word		
	Health	Tourism	SemCor
Noun	7.06	12.56	10.98
Verb	7.47	9.76	11.95
Adjective	5.74	12.07	8.67
Adverb	9.11	19.78	25.44
All	6.94	12.17	11.25

Table 2: Average number of instances per polysemous word per category in the 3 domains

Category	Avg. degree of Wordnet polysemy for polysemous words		
	Health	Tourism	SemCor
Noun	5.24	4.95	5.60
Verb	10.60	10.10	9.89
Adjective	5.52	5.08	5.40
Adverb	3.64	4.16	3.90
All	6.49	5.77	6.43

Table 3: Average degree of Wordnet polysemy of polysemous words per category in the 3 domains

Category	Avg. degree of Corpus polysemy for polysemous words		
	Health	Tourism	SemCor
Noun	1.92	2.60	3.41
Verb	3.41	4.55	4.73
Adjective	2.04	2.57	2.65
Adverb	2.16	2.82	3.09
All	2.31	2.93	3.56

Table 4: Average degree of Corpus polysemy of polysemous words per category in the 3 domains

them to the training data. An obvious question which arises at this point is "Why were the words selected at random?" or "Can selection of words using some active learning strategy yield better results than a random selection?" We discuss this question in detail in Section 7 and show that a random set of injections performs no worse than a craftily selected set of injections.

5 DataSet Preparation

Due to the lack of any publicly available all-words domain specific sense marked corpora we set upon the task of collecting data from two domains, *viz.*, *Tourism and Health*. The data for Tourism domain was downloaded from Indian Tourism websites whereas the data for Health domain was obtained from two doctors. This data was manually sense annotated by two lexicographers adept in English. Princeton Wordnet 2.1[3] (Fellbaum, 1998) was used as the sense inventory. A total of 1,34,095 words from the Tourism domain and 42,046 words from the Health domain were manually sense marked. Some files were sense marked by both the lexicographers and the Inter Tagger Agreement (ITA) calculated from these files was 83% which is comparable to the 78% ITA reported on the SemCor corpus considering the domain-specific nature of the corpus.

We now present different statistics about the corpora. Table 1 summarizes the number of polysemous and monosemous words in each category. Note that we do not use the monosemous words while calculating precision and recall of our algorithms.

Table 2 shows the average number of instances per polysemous word in the 3 corpora. We note that the number of instances per word in the Tourism domain is comparable to that in the SemCor corpus whereas the number of instances per word in the Health corpus is smaller due to the overall smaller size of the Health corpus.

Tables 3 and 4 summarize the average degree of Wordnet polysemy and corpus polysemy of the polysemous words in the corpus. Wordnet polysemy is the number of senses of a word as listed in the Wordnet, whereas corpus polysemy is the number of senses of a word actually appearing in the corpus. As expected, the average degree of corpus polysemy (Table 4) is much less than the average degree of Wordnet polysemy (Table 3). Further, the average degree of corpus polysemy (Table 4) in the two domains is less than that in the mixed-domain SemCor corpus, which is expected due to the domain specific nature of the corpora.

Finally, Table 5 summarizes the number of unique polysemous words per category in each domain.

Category	No. of unique polysemous words		
	Health	Tourism	SemCor
Noun	2188	4229	5871
Verb	984	1591	2565
Adjective	1024	1635	2640
Adverb	217	308	463
All	4413	7763	11539

Table 5: Number of unique polysemous words per category in each domain.

[3] http://wordnetweb.princeton.edu/perl/webwn

The data is currently being enhanced by manually sense marking more words from each domain and will be soon freely available[4] for research purposes.

6 Results

We tested the 3 algorithms described in section 4 using SemCor, Tourism and Health domain corpora. We did a 2-fold cross validation for supervised adaptation and report the average performance over the two folds. Since the knowledge based and unsupervised methods do not need any training data we simply test it on the entire corpus from the two domains.

6.1 Knowledge Based approach

The results obtained by applying the Personalized PageRank (PPR) method to Tourism and Health data are summarized in Table 6. We also report the Wordnet first sense baseline (WFS).

Domain	Algorithm	P(%)	R(%)	F(%)
Tourism	PPR	53.1	53.1	53.1
	WFS	62.5	62.5	62.5
Health	PPR	51.1	51.1	51.1
	WFS	65.5	65.5	65.5

Table 6: Comparing the performance of Personalized PageRank (PPR) with Wordnet First Sense Baseline (WFS)

6.2 Unsupervised approach

The predominant sense for each word in the two domains was calculated using the method described in section 4.2. McCarthy et al. (2004) reported that the best results were obtained using $k = 50$ neighbors and the Wordnet Similarity ***jcn*** measure (Jiang and Conrath, 1997). Following them, we used $k = 50$ and observed that the best results for nouns and verbs were obtained using the ***jcn*** measure and the best results for adjectives and adverbs were obtained using the ***lesk*** measure (Banerjee and Pedersen, 2002). Accordingly, we used ***jcn*** for nouns and verbs and ***lesk*** for adjectives and adverbs. Each target word in the test set is then disambiguated by simply assigning it its predominant sense obtained using the above method. We tested this approach only on Tourism domain due to unavailability of large untagged Health corpus which is needed for constructing the thesaurus. The results are summarized in Table 7.

Domain	Algorithm	P(%)	R(%)	F(%)
Tourism	PPR	51.85	49.32	50.55
	WFS	62.50	62.50	62.50

Table 7: Comparing the performance of unsupervised approach with Wordnet First Sense Baseline (WFS)

6.3 Supervised adaptation

We report results in the **source setting**, **target setting** and **adaptation setting** as described earlier using the following four combinations for source and target data:

1. ***SemCor to Tourism (SC→T)*** where SemCor is used as the source domain and Tourism as the target (test) domain.
2. ***SemCor to Health (SC→H)*** where SemCor is used as the source domain and Health as the target (test) domain.
3. ***Tourism to Health (T→H)*** where Tourism is used as the source domain and Health as the target (test) domain.
4. ***Health to Tourism (H→T)*** where Health is used as the source domain and Tourism as the target (test) domain.

In each case, the target domain data was divided into two folds. One fold was set aside for testing and the other for injecting data in the ***adaptation setting***. We increased the size of the injected target examples from 1000 to 14000 words in increments of 1000. We then repeated the same experiment by reversing the role of the two folds.

Figures 1, 2, 3 and 4 show the graphs of the average F-score over the 2-folds for SC→T, SC→H, T→H and H→T respectively. The x-axis represents the amount of training data (in words) injected from the target domain and the y-axis represents the F-score. The different curves in each graph are as follows:

a. *only_random* : This curve plots the performance obtained using x randomly selected sense tagged words from the target domain and zero sense tagged words from the source domain (x was varied from 1000 to 14000 words in increments of 1000).

[4] http://www.cfilt.iitb.ac.in/wsd/annotated_corpus

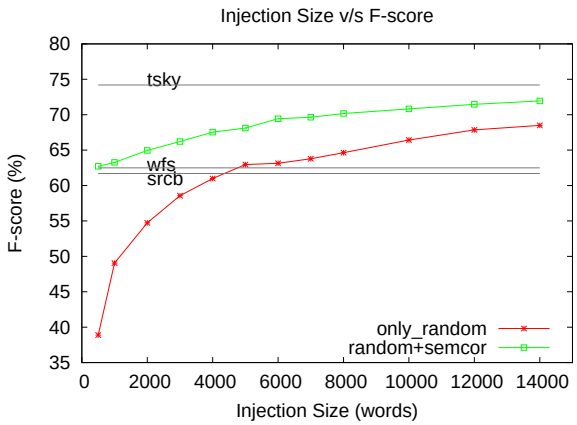

Figure 1: Supervised adaptation from SemCor to Tourism using injections

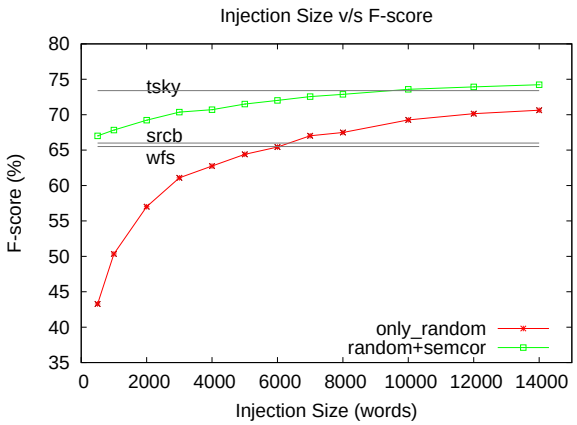

Figure 2: Supervised adaptation from SemCor to Health using injections

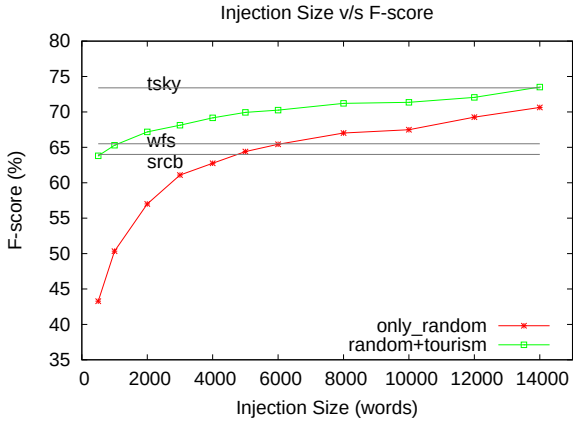

Figure 3: Supervised adaptation from Tourism to Health using injections

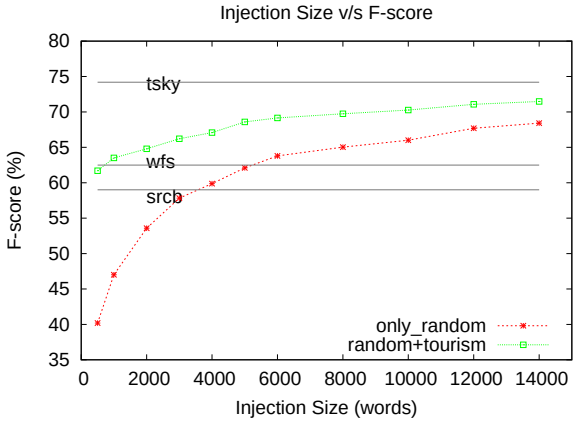

Figure 4: Supervised adaptation from Health to Tourism using injections

b. *random+source* : This curve plots the performance obtained by mixing x randomly selected sense tagged words from the target domain with the entire training data from the source domain (again x was varied from 1000 to 14000 words in increments of 1000).

c. *source_baseline* ***(srcb)*** : This represents the F-score obtained by training on the source data alone without mixing any examples from the target domain.

d. *wordnet_first_sense* ***(wfs)*** : This represents the F-score obtained by selecting the first sense from Wordnet, a typically reported baseline.

e. *target_skyline* ***(tsky)*** : This represents the average 2-fold F-score obtained by training on one entire fold of the target data itself (*Health*: 15320 polysemous words; *Tourism*: 47242 polysemous words) and testing on the other fold.

These graphs along with other results are discussed in the next section.

7 Discussions

We discuss the performance of the three approaches.

7.1 Knowledge Based and Unsupervised approaches

It is apparent from Tables 6 and 7 that knowledge based and unsupervised approaches do not perform well when compared to the Wordnet first sense (which is freely available and hence can be used for disambiguation). Further, we observe that the performance of these approaches is even less than the *source_baseline* (*i.e.*, the case when training data from a source domain is applied as it is to a target domain - without using any injections). These observations bring out the weaknesses of these approaches when used in an all-words setting and clearly indicate that they come nowhere close to replacing a supervised system.

7.2 Supervised adaptation

1. The F-score obtained by training on SemCor (mixed-domain corpus) and testing on the two target domains without using any injections (***srcb***) – F-score of 61.7% on Tourism and F-score of 65.5% on Health – is comparable to the best result reported on the SEMEVAL datasets (65.02%, where both training and testing happens on a mixed-domain corpus (Snyder and Palmer, 2004)). This is in contrast to previous studies (Escudero et al., 2000; Agirre and Martinez, 2004) which suggest that instead of adapting from a generic/mixed domain to a specific domain, it is better to completely ignore the generic examples and use hand-tagged data from the target domain itself. The main reason for the contrasting results is that the earlier work focused only on a handful of target words whereas we focus on all words appearing in the corpus. So, while the behavior of a few target words would change drastically when the domain changes, a majority of the words will exhibit the same behavior (*i.e.*, same predominant sense) even when the domain changes. We agree that the overall performance is still lower than that obtained by training on the domain-specific corpora. However, it is still better than the performance of unsupervised and knowledge based approaches which tilts the scale in favor of supervised approaches even when only mixed domain sense marked corpora is available.

2. Adding injections from the target domain improves the performance. As the amount of injection increases the performance approaches the skyline, and in the case of SC→H and T→H it even crosses the skyline performance showing that combining the source and target data can give better performance than using the target data alone. This is consistent with the domain adaptation results reported by Agirre and de Lacalle (2009) on a specific set of target words.

3. The performance of *random+source* is always better than *only_random* indicating that the data from the source domain does help to improve performance. A detailed analysis showed that the gain obtained by using the source data is attributable to reducing recall errors by increasing the coverage of seen words.

4. Adapting from one specific domain (*Tourism or Health*) to another specific domain (*Health or Tourism*) gives the same performance as that obtained by adapting from a mixed-domain (*SemCor*) to a specific domain (*Tourism*, *Health*). This is an interesting observation as it suggests that as long as data from one domain is available it is easy to build a WSD engine that works for other domains by injecting a small amount of data from these domains.

To verify that the results are consistent, we randomly selected 5 different sets of injections from fold-1 and tested the performance on fold-2. We then repeated the same experiment by reversing the roles of the two folds. The results were indeed consistent irrespective of the set of injections used. Due to lack of space we have not included the results for these 5 different sets of injections.

7.3 Quantifying the trade-off between performance and corpus size

To correctly quantify the benefit of adding injections from the target domain, we calculated the amount of target data (*peak_size*) that is needed to reach the skyline F-score (*peak_F*) in the absence of any data from the source domain. The *peak_size* was found to be 35000 (Tourism) and 14000 (Health) corresponding to *peak_F* values of 74.2% (Tourism) and 73.4% (Health). We then plotted a graph (Figure 5) to capture the relation between the size of injections (expressed as a percentage of the *peak_size*) and the F-score (expressed as a percentage of the *peak_F*).

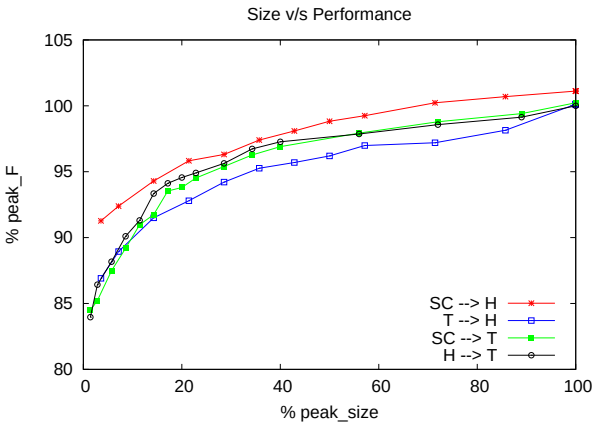

Figure 5: Trade-off between performance and corpus size

We observe that by mixing only 20-40% of the *peak_size* with the source domain we can obtain up to 95% of the performance obtained by using the

entire target data (*peak_size*). In absolute terms, the size of the injections is only 7000-9000 polysemous words which is a very small price to pay considering the performance benefits.

8 Does the choice of injections matter?

An obvious question which arises at this point is *"Why were the words selected at random?"* or *"Can selection of words using some active learning strategy yield better results than a random selection?"* An answer to this question requires a more thorough understanding of the *sense-behavior* exhibited by words across domains. In any scenario involving a shift from domain D_1 to domain D_2, we will always encounter words belonging to the following 4 categories:

a. W_{D_1} : This class includes words which are encountered only in the source domain D_1 and do not appear in the target domain D_2. Since we are interested in adapting to the target domain and since these words do not appear in the target domain, it is quite obvious that they are **not important** for the problem of domain adaptation.

b. W_{D_2} : This class includes words which are encountered only in the target domain D_2 and do not appear in the source domain D_1. Again, it is quite obvious that these words are **important** for the problem of domain adaptation. They fall in the category of unseen words and need handling from that point of view.

c. $W_{D1D2_{conformists}}$: This class includes words which are encountered in both the domains and exhibit the same predominant sense in both the domains. Correct identification of these words is **important** so that we can use the predominant sense learned from D_1 for disambiguating instances of these words appearing in D_2.

d. $W_{D1D2_{non-conformists}}$: This class includes words which are encountered in both the domains but their predominant sense in the target domain D_2 **does not conform** to the predominant sense learned from the source domain D_1. Correct identification of these words is **important** so that we can ignore the predominant senses learned from D_1 while disambiguating instances of these words appearing in D_2.

Table 8 summarizes the percentage of words that fall in each category in each of the three adaptation scenarios. The fact that nearly 50-60% of the words fall in the "conformist" category once again makes a strong case for reusing sense tagged data from one domain to another domain.

Category	SC→T	SC→H	T→H
$\mathbf{W_{D_2}}$	7.14%	5.45%	13.61%
Conformists	49.54%	60.43%	54.31%
Non-Conformists	43.30%	34.11%	32.06%

Table 8: Percentage of Words belonging to each category in the three settings.

The above characterization suggests that an *ideal* domain adaptation strategy should focus on injecting W_{D_2} and $W_{D1D2_{non-conformists}}$ as these would yield maximum benefits if injected into the training data. While it is easy to identify the W_{D_2} words, *"identifying non-conformists"* is a hard problem which itself requires some type of WSD[5]. However, just to prove that a *random* injection strategy does as good as an *ideal* strategy we assume the presence of an *oracle* which identifies the $W_{D1D2_{non-conformists}}$. We then augment the training data with 5-8 instances for W_{D_2} and $W_{D1D2_{non-conformists}}$ words thus identified. We observed that adding more than 5-8 instances per word does not improve the performance. This is due to the "one sense per domain" phenomenon – seeing only a few instances of a word is sufficient to identify the predominant sense of the word. Further, to ensure a better overall performance, the instances of the most frequent words are injected first followed by less frequent words till we exhaust the total size of the injections (500, 1000, 2000 and so on). We observed that there was a 75-80% overlap between the words selected by random strategy and oracle strategy. This is because oracle selects the most frequent words which also have a high chance of getting selected when a random sampling is done.

Figures 6, 7, 8 and 9 compare the performance of the two strategies. We see that the random strategy does as well as the oracle strategy thereby supporting our claim that *if we have sense marked corpus from one domain then simply injecting ANY small amount of data from the target domain will*

[5]Note that the unsupervised predominant sense acquisition method of McCarthy et al. (2007) implicitly identifies conformists and non-conformists

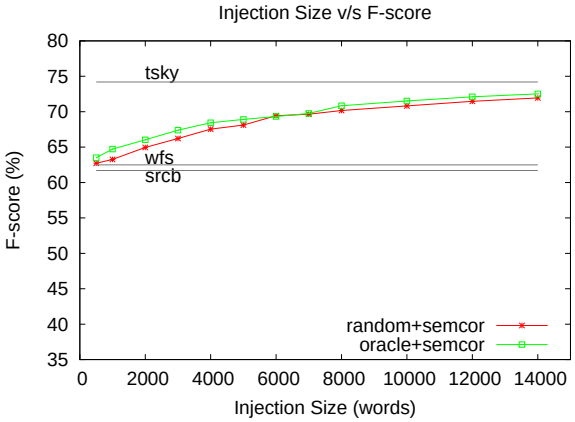

Figure 6: Comparing random strategy with oracle based ideal strategy for SemCor to Tourism adaptation

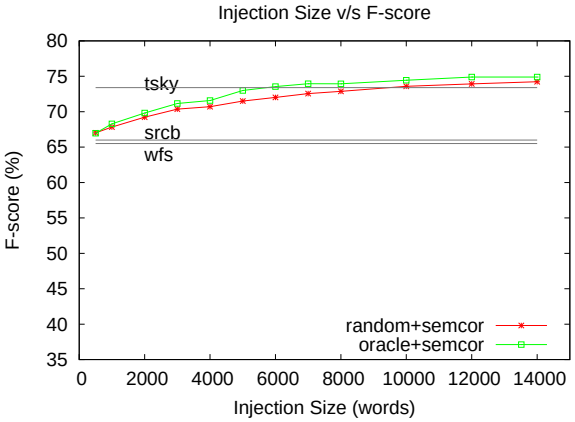

Figure 7: Comparing random strategy with oracle based ideal strategy for SemCor to Health adaptation

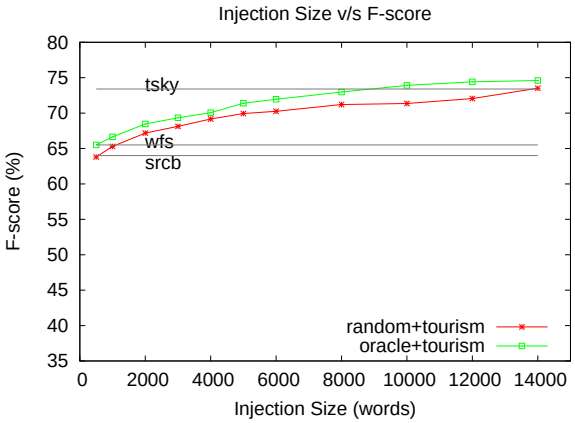

Figure 8: Comparing random strategy with oracle based ideal strategy for Tourism to Health adaptation

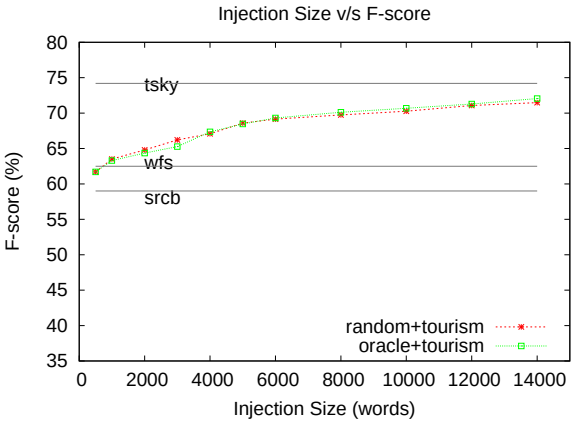

Figure 9: Comparing random strategy with oracle based ideal strategy for Health to Tourism adaptation

do the job.

9 Conclusion and Future Work

Based on our study of WSD in 3 domain adaptation scenarios, we make the following conclusions:

1. Supervised adaptation by mixing small amount of data (7000-9000 words) from the target domain with the source domain gives nearly the same performance (F-score of around 70% in all the 4 adaptation scenarios) as that obtained by training on the entire target domain data.

2. Unsupervised and knowledge based approaches which use distributional similarity and Wordnet based similarity measures do not compare well with the Wordnet first sense baseline performance and do not come anywhere close to the performance of supervised adaptation.

3. Supervised adaptation from a mixed domain to a specific domain gives the same performance as that from one specific domain (Tourism) to another specific domain (Health).

4. Supervised adaptation is not sensitive to the type of data being injected. This is an interesting finding with the following implication: as long as one has sense marked corpus - be it from a mixed or specific domain - simply injecting ANY small amount of data from the target domain suffices to beget good accuracy.

As future work, we would like to test our work on the Environment domain data which was released as part of the SEMEVAL 2010 shared task on “All-words Word Sense Disambiguation on a Specific Domain”.

References

Eneko Agirre and Oier Lopez de Lacalle. 2009. Supervised domain adaption for wsd. In *EACL '09: Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 42–50, Morristown, NJ, USA. Association for Computational Linguistics.

Eneko Agirre and David Martinez. 2004. The effect of bias on an automatically-built word sense corpus. In *Proceedings of the 4rd International Conference on Languages Resources and Evaluations (LREC)*.

Eneko Agirre, Oier Lopez de Lacalle, Christiane Fellbaum, Andrea Marchetti, Antonio Toral, and Piek Vossen. 2009a. Semeval-2010 task 17: all-words word sense disambiguation on a specific domain. In *DEW '09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, pages 123–128, Morristown, NJ, USA. Association for Computational Linguistics.

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

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted lesk algorithm for word sense disambiguation using wordnet. In *CICLing '02: Proceedings of the Third International Conference on Computational Linguistics and Intelligent Text Processing*, pages 136–145, London, UK. Springer-Verlag.

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

Gerard Escudero, Lluís Màrquez, and German Rigau. 2000. An empirical study of the domain dependence of supervised word sense disambiguation systems. In *Proceedings of the 2000 Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora*, pages 172–180, Morristown, NJ, USA. Association for Computational Linguistics.

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

J.J. Jiang and D.W. Conrath. 1997. Semantic similarity based on corpus statistics and lexical taxonomy. In *Proc. of the Int'l. Conf. on Research in Computational Linguistics*, pages 19–33.

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)*.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *IN PROCEEDINGS OF THE 41ST ANNUAL MEETING OF THE ASSOCIATION FOR COMPUTATIONAL LINGUISTICS*, pages 423–430.

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. *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 sense: an exemplar-based approach. In *Proceedings of the 34th annual meeting on Association for Computational Linguistics*, pages 40–47, Morristown, NJ, USA. Association for Computational Linguistics.

Siddharth Patwardhan and Ted Pedersen. 2003. The cpan wordnet::similarity package. http://search .cpan.org/ sid/wordnet-similarity/.

Benjamin Snyder and Martha Palmer. 2004. The english all-words task. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

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