Projecting Parameters for Multilingual Word Sense Disambiguation

Mitesh M. Khapra Sapan Shah Piyush Kedia Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai – 400076,
Maharashtra, India.
{miteshk,sapan,charasi,pb}@cse.iitb.ac.in

Abstract

We report in this paper a way of doing Word Sense Disambiguation (WSD) that has its origin in multilingual MT and that is cognizant of the fact that *parallel corpora, wordnets* and *sense annotated corpora are scarce resources.* With respect to these resources, languages show different levels of readiness; *however a more resource fortunate language can help a less resource fortunate language*. Our WSD method can be applied to a language even when no sense tagged corpora for that language is available. This is achieved by ***projecting wordnet and corpus parameters*** from another language to the language in question. The approach is centered around a novel synset based multilingual dictionary and the empirical observation that within a domain the distribution of senses remains more or less invariant across languages. The effectiveness of our approach is verified by doing parameter projection and then running two different WSD algorithms. The accuracy values of approximately 75% (F1-score) for three languages in two different domains establish the fact that within a domain it is possible to circumvent the problem of scarcity of resources by projecting parameters like sense distributions, corpus-co-occurrences, conceptual distance, *etc.* from one language to another.

1 Introduction

Currently efforts are on in India to build large scale Machine Translation and Cross Lingual Search systems in consortia mode. These efforts are large, in the sense that 10-11 institutes and 6-7 languages spanning the length and breadth of the country are involved. The approach taken for translation is transfer based which needs to tackle the problem of word sense disambiguation (WSD) (Sergei *et. al*., 2003). Since 90s machine learning based approaches to WSD using sense marked corpora have gained ground (Eneko Agirre & Philip Edmonds, 2007). However, the creation of sense marked corpora has always remained a costly proposition. Statistical MT has obviated the need for elaborate resources for WSD, because WSD in SMT happens implicitly through parallel corpora (Brown *et. al*., 1993). But parallel corpora too are a very costly resource.

The above situation brings out the challenges involved in Indian language MT and CLIR. Lack of resources coupled with the multiplicity of Indian languages severely affects the performance of several NLP tasks. In the light of this, we focus on the problem of developing methodologies that *reuse resources*. The idea is to *do the annotation work for one language and find ways of using them for another language.*

Our work on WSD takes place in a multilingual setting involving *Hindi* (national language of India; 500 million speaker base), *Marathi* (20 million speaker base), *Bengali* (185 million speaker base) *and Tamil* (74 million speaker base). The wordnet of Hindi and sense marked corpora of Hindi are used *for all these languages.* Our methodology rests on a novel multilingual dictionary organization and on the idea of "parameter projection" from Hindi to the other languages. Also the domains of interest are *tourism* and *health.*

The roadmap of the paper is as follows. Section 2 describes related work. In section 3 we introduce the parameters essential for domain-specific WSD. Section 4 builds the case for parameter projection. Section 5 introduces the Multilingual Dictionary Framework which plays a key role in parameter projection. Section 6 is the core of the work, where we present parameter projection from one language to another. Section 7 describes two WSD algorithms which combine various parameters for do-

Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing, pages 459–467,
Singapore, 6-7 August 2009.

main-specific WSD. Experiments and results are presented in sections 8 and 9. Section 10 concludes the paper.

2 Related work

Knowledge based approaches to WSD such as Lesk's algorithm (Michael Lesk, 1986), Walker's algorithm (Walker D. & Amsler R., 1986), conceptual density (Agirre Eneko & German Rigau, 1996) and random walk algorithm (Mihalcea Rada, 2005) essentially do Machine Readable Dictionary lookup. However, these are fundamentally *overlap based* algorithms which suffer from overlap sparsity, dictionary definitions being generally small in length.

Supervised learning algorithms for WSD are mostly word specific classifiers, *e.g.*, WSD using SVM (Lee *et. al*., 2004), Exemplar based WSD (Ng Hwee T. & Hian B. Lee, 1996) and decision list based algorithm (Yarowsky, 1994). The requirement of a large training corpus renders these algorithms unsuitable for resource scarce languages.

Semi-supervised and unsupervised algorithms do not need large amount of annotated corpora, but are again word specific classifiers, *e.g.*, semi-supervised decision list algorithm (Yarowsky, 1995) and Hyperlex (Véronis Jean, 2004)). Hybrid approaches like WSD using Structural Semantic Interconnections (Roberto Navigli & Paolo Velardi, 2005) use combinations of more than one knowledge sources (wordnet as well as a small amount of tagged corpora). This allows them to capture important information encoded in wordnet (Fellbaum, 1998) as well as draw syntactic generalizations from minimally tagged corpora.

At this point we state that no single existing solution to WSD completely meets our requirements of **multilinguality, high domain accuracy** and **good performance in the face of not-so-large annotated corpora**.

3 Parameters for WSD

We discuss a number of parameters that play a crucial role in WSD. To appreciate this, consider the following example:

*The river flows through this region to meet the **sea.***

The word *sea* is ambiguous and has three senses as given in the Princeton Wordnet (PWN):

S1: (n) sea (a division of an ocean or a large body of salt water partially enclosed by land)
S2: (n) ocean, sea (anything apparently limitless in quantity or volume)
S3: (n) sea (turbulent water with swells of considerable size) "heavy seas"

Our first parameter is obtained from ***Domain specific sense distributions.*** In the above example, the first sense is more frequent in the tourism domain (verified from manually sense marked tourism corpora). Domain specific sense distribution information should be harnessed in the WSD task.

The second parameter arises from the ***dominance of senses in the domain.*** Senses are expressed by synsets, and we define a dominant sense as follows:

> A synset node in the wordnet hypernymy hierarchy is called *Dominant* if the synsets in the sub-tree below the synset are frequently occurring in the domain corpora.

A few dominant senses in the Tourism domain are *{place, country, city, area}*, *{body of water}*, *{flora, fauna}*, *{mode of transport}* and *{fine arts}*. In disambiguating a word, that sense which belongs to the sub-tree of a domain-specific dominant sense should be given a higher score than other senses. The value of this parameter (θ) is decided as follows:

θ = 1; if the candidate synset is a dominant synset

θ = 0.5; if the candidate synset belongs to the sub-tree of a dominant synset

θ = 0.001; if the candidate synset is neither a dominant synset nor belongs to the sub-tree of a dominant synset.

Our third parameter comes from ***Corpus co-occurrence.*** Co-occurring monosemous words as well as *already disambiguated words* in the context help in disambiguation. For example, the word *river* appearing in the context of *sea* is a monosemous word. The frequency of co-occurrence of *river* with the "water body" sense of *sea* is high in the tourism domain. Corpus co-occurrence is cal-

culated by considering the senses which occur in a window of 10 words around a sense.

Our fourth parameter is based on the ***semantic distance*** between any pair of synsets in terms of the shortest path length between two synsets in the wordnet graph. An edge in the shortest path can be any semantic relation from the wordnet relation repository (*e.g., hypernymy, hyponymy, meronymy, holonymy, troponymy etc.).*

For nouns we do something additional over and above the semantic distance. We take advantage of the deeper hierarchy of noun senses in the wordnet structure. This gives rise to our fifth and final parameter which arises out of the ***conceptual distance*** between a pair of senses. Conceptual distance between two synsets S_1 and S_2 is calculated using Equation (1), motivated by Agirre Eneko & German Rigau (1996).

$$\textit{Conceptual Distance (S1, S2)} = \frac{\text{Length of the path between (S1, S2) in terms of hypernymy hierarchy}}{\text{Height of the lowest common ancestor of S1 and S2 in the wordnet hierarchy}} \quad (1)$$

The conceptual distance is proportional to the path length between the synsets, as it should be. The distance is also inversely proportional to the height of the common ancestor of two sense nodes, because as the common ancestor becomes more and more general the conceptual relatedness tends to get vacuous (e.g., two nodes being related through *entity* which is the common ancestor of EVERYTHING, does not really say anything about the relatedness).

To summarize, our various parameters used for domain-specific WSD are:

Wordnet-dependent parameters

- *belongingness-to-dominant-concept*
- *conceptual-distance*
- s*emantic-distance*

Corpus-dependent parameters

- *sense distributions*
- *corpus co-occurrence*.

In section 7 we show how these parameters are used to come up with a scoring function for WSD.

4 Building a case for Parameter Projection

Wordnet-dependent parameters depend on the graph based structure of Wordnet whereas the ***Corpus-dependent parameters*** depend on various statistics learnt from a sense marked corpora. Both the tasks of (a) constructing a wordnet from scratch and (b) collecting sense marked corpora for multiple languages are tedious and expensive. An important question being addressed in this paper is: *whether the effort required in constructing semantic graphs for multiple wordnets and collecting sense marked corpora can be avoided?* Our findings seem to suggest that by ***projecting relations*** from the wordnet of a language and by ***projecting corpus*** statistics from the sense marked corpora of the language we can achieve this end. Before we proceed to discuss the way to realize parameter projection, we present a novel dictionary which facilitates this task.

5 Synset based multilingual dictionary

Parameter projection as described in section 4 rests on a novel and effective method of storage and use of dictionary in a multilingual setting proposed by Mohanty *et. al*. (2008). For the purpose of current discussion, we will call this multilingual dictionary framework *MultiDict*. One important departure from traditional dictionary is that **synsets are linked, and after that the words inside the synsets are linked.** The basic mapping is thus between synsets and thereafter between the words.

Concepts	**L1 (English)**	**L2 (Hindi)**	**L3 (Marathi)**
04321: a youthful male person	{male child, boy}	{लड़का ladkaa, बालक,baalak बच्चा bachchaa}	{मुलगाmulgaa , पोरगाporgaa , पोरpor }

Table 1: Multilingual Dictionary Framework

Table 1 shows the structure of MultiDict, with one example row standing for the concept of *boy*. The first column is the pivot describing a concept with a unique ID. The subsequent columns show the words expressing the concept in respective languages (in the example table above, *English, Hindi* and *Marathi*). Thus to express the concept *'04321: a youthful male person'*, there are two lexical elements in English, which constitute a *synset.* Correspondingly, the Hindi and Marathi synsets contain 3 words each.

It may be noted that the *central language* whose synsets the synsets of other languages link to is Hindi. This way of linking synsets- more popularly known as the *expansion* approach- has several advantages as discussed in (Mohanty *et. al.*, 2008). One advantage germane to the point of this paper is that the synsets in a particular column automatically inherit the various semantic relations of the Hindi wordnet (Dipak Narayan *et. al., 2000*), which saves the effort involved in reconstructing these relations for multiple languages.

After the synsets are linked, **cross linkages are set up** manually from the words of a synset to the words of a linked synset of the central language. The average number of such links per synset per language pair is approximately 3. These cross-linkages actually solve the problem of *lexical choice* in translating from text of one language to another.

Thus for the Marathi word मुलगा *{mulagaa}* denoting "*a youthful male person",* the correct lexical substitute from the corresponding Hindi synset is लड़का *{ladakaa}* (Figure 1). One might argue that any word within the synset could serve the purpose of translation. However, the exact lexical substitution has to respect native speaker acceptability.

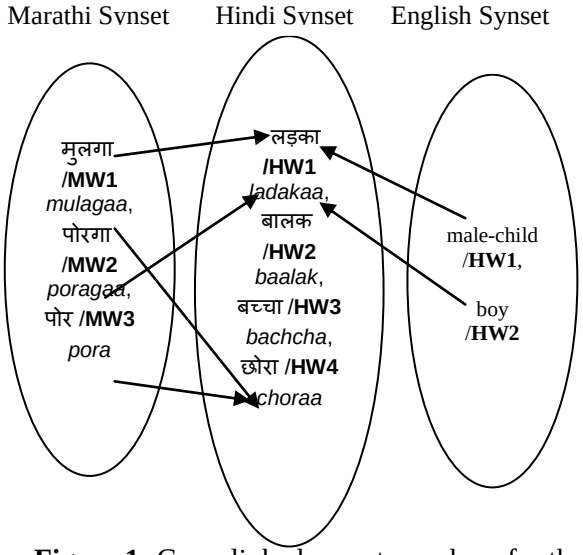

Figure 1: Cross linked synset members for the concept: *a youthful male person*

We put these cross linkages to another use, as described later.

Since it is the MultiDict which is at the heart of parameter projection, we would like to summarize the main points of this section. (1) By linking with the synsets of Hindi, the cost of building wordnets of other languages is partly reduced (semantic relations are inherited). The wordnet parameters of Hindi wordnet now become projectable to other languages. (2) By using the *cross linked words* in the synsets, corpus parameters become projectable (*vide* next section).

6 Parameter projection using MultDict

6.1 *P(Sense|Word)* parameter

Suppose a word (say, ***W***) in language L_1 (say, Marathi) has *k* senses. For each of these *k* senses we are interested in finding the parameter $P(S_i|W)$- which is the probability of sense S_i given the word *W* expressed as:

$$P(S_i|W) = \frac{\#(S_i, W)}{\sum_j \#(S_j, W)}$$

where '#' indicates 'count-of'. Consider the example of two senses of the Marathi word सागर *{saagar}, viz., sea* and *abundance* and the corresponding cross-linked words in Hindi (Figure 2 below):

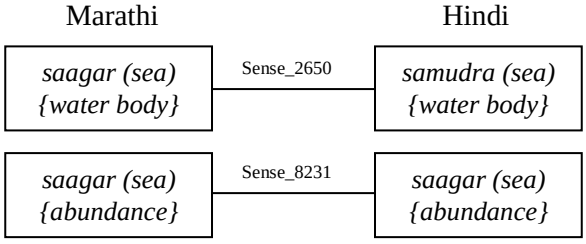

Figure 2: Two senses of the Marathi word सागर *(saagar), viz., {water body}* and *{abundance},* and the corresponding cross-linked words in Hindi[1].

The probability *P({water body}|saagar)* for Marathi is

$$\frac{\#(\{\boldsymbol{water\ body}\}, \boldsymbol{saagar})}{\#(\{\boldsymbol{water\ body}\}, \boldsymbol{saagar}) + \#(\{\boldsymbol{abundance}\}, \boldsymbol{saagar})}$$

We propose that this can be approximated by the counts from Hindi sense marked corpora by replacing *saagar* with the cross linked Hindi words *samudra* and *saagar*, as per Figure 2:

$$\frac{\#(\{\text{water body}\}, \text{samudra})}{\#(\{\text{water body}\}, \text{samudra}) + \#(\{\text{abundance}\}, \text{saagar})}$$

[1] Sense_8231 shows the same word *saagar* for both Marathi and Hindi. This is not uncommon, since Marathi and Hindi are sister languages.

Thus, the following formula is used for calculating the sense distributions of Marathi words using the sense marked Hindi corpus from the same domain:

$$P(S_i|W) = \frac{\#(S_i\ , cross\_linked\_hindi\_word)}{\sum_j\ \#(S_j\ , cross\_linked\_hindi\_word)} \qquad (2)$$

Note that we are not interested in the *exact* sense distribution of the words, but only in the *relative* sense distribution.

To prove that the projected relative distribution is faithful to the actual relative distribution of senses, we obtained the sense distribution statistics of a set of Marathi words from a sense tagged Marathi corpus (we call the sense marked corpora of a language its **self corpora**). These sense distribution statistics were compared with the statistics for these same words obtained by *projecting from* a sense tagged Hindi corpus using Equation (2). The results are summarized in Table 2.

Sr. No	Marathi Word	Synset	P(S|word) as learnt from sense tagged Marathi corpus	P(S|word) as projected from sense tagged Hindi corpus
1	किंमत (kimat)	{ worth }	0.684	0.714
		{ price }	0.315	0.285
2	रस्ता (rasta)	{ roadway }	0.164	0.209
		{road, route}	0.835	0.770
3	ठिकाण (thikan)	{ land site, place}	0.962	0.878
		{ home }	0.037	0.12
4	सागर (saagar)	{water body}	1.00	1.00
		{abundance}	0	0

Table 2: Comparison of the sense distributions of some Marathi words learnt from Marathi sense tagged corpus with those projected from Hindi sense tagged corpus.

The fourth row of Table 2 shows that whenever सागर *(saagar) (sea)* appears in the Marathi tourism corpus there is a 100% chance that it will appear in the *"water body"* sense and 0% chance that it will appear in the sense of *"abundance"*. Column 5 shows that the same probability values are obtained using projections from Hindi tourism corpus. Taking another example, the third row shows that whenever ठिकाण (*thikaan) (place, home)* appears in the Marathi tourism corpus there is a much higher chance of it appearing in the sense of "place" (96.2%) then in the sense of "home" (3.7%). Column 5 shows that the relative probabilities of the two senses remain the same even when using projections from Hindi tourism corpus (i.e. by using the corresponding cross-linked words in Hindi). To quantify these observations, we calculated the average KL divergence and Spearman's correlation co-efficient between the two distributions. The KL divergence is 0.766 and Spearman's correlation co-efficient is 0.299. Both these values indicate that there is a high degree of similarity between the distributions learnt using projection and those learnt from the self corpus.

6.2 Co-occurrence parameter

Similarly, within a domain, the statistics of co-occurrence of senses remain the same across languages. For example, the co-occurrence of the Marathi synsets {आकाश (akash) (sky), अंबर (ambar) (sky)} and {मेघ (megh) (cloud), अभ्र (abhra) (cloud)} in the Marathi corpus remains more or less same as (or proportional to) the co-occurrence between the corresponding Hindi synsets in the Hindi corpus.

Sr. No	Synset	Co-occurring Synset	P(co-occurrence) as learnt from sense tagged Marathi corpus	P(co-occurrence) as learnt from sense tagged Hindi corpus
1	{रोप, रोपटे} {small bush}	{झाड, वृक्ष, तरुवर, द्रुम, तरू, पादप} {tree}	0.125	0.125
2	{मेघ, अभ्र} {cloud}	{आकाश, आभाळ, अंबर} {sky}	0.167	0.154
3	{क्षेत्र, इलाक़ा, इलाका, भूखंड} {geographical area}	{यात्रा, सफ़र} {travel}	0.0019	0.0017

Table 3: Comparison of the corpus co-occurrence statistics learnt from Marathi and Hindi Tourism corpus.

Table 3 shows a few examples depicting similarity between co-occurrence statistics learnt from Marathi tourism corpus and Hindi tourism corpus. Note that we are talking about co-occurrence of synsets and not words. For example, the second row shows that the probability of co-occurrence of the synsets {cloud} and {sky} is almost same in the Marathi and Hindi corpus.

7 Our algorithms for WSD

We describe two algorithms to establish the usefulness of the idea of parameter projection. The first algorithm- called *iterative WSD (IWSD-)* is greedy, and the second based on PageRank algorithm is exhaustive. Both use scoring functions that make use of the parameters detailed in the previous sections.

7.1 Iterative WSD (IWSD)

We have been motivated by the Energy expression in Hopfield network (Hopfield, 1982) in formulating a scoring function for ranking the senses. Hopfield Network is a fully connected bidirectional symmetric network of bi-polar (0/1 or +1/-1) neurons. We consider the asynchronous Hopfield Network. At any instant, a randomly chosen neuron (a) examines the weighted sum of the input, (b) compares this value with a threshold and (c) gets to the state of 1 or 0, depending on whether the input is greater than or less than or equal to the threshold. The assembly of 0/1 states of individual neurons defines a state of the whole network. Each state has associated with it an energy, *E*, given by the following expression

$$E = -\theta_i V_i + \sum_{i=1}^{N} \sum_{j>i}^{N} W_{ij}\, V_i V_j \qquad (3)$$

where, *N* is the total number of neurons in the network, V_i and V_j are the activations of neurons *i* and *j* respectively and W_{ij} is the weight of the connection between neurons *i* and *j*. Energy is a fundamental property of Hopfield networks, providing the necessary machinery for discussing convergence, stability and such other considerations.

The energy expression as given above cleanly separates the influence of self-activations of neurons and that of interactions amongst neurons to the global macroscopic property of energy of the network. This fact has been the primary insight for equation (4) which was proposed to score the most appropriate synset in the given context. The correspondences are as follows:

Neuron	→	*Synset*
Self-activation	→	*Corpus Sense Distribution*
Weight of connection between two neurons	→	*Weight as a function of corpus co-occurrence and Wordnet distance measures between synsets*

$$S^* = \underset{i}{\operatorname{argmax}} \left(\theta_i * V_i + \sum_{j \in \mathrm{J}} W_{ij} * V_i * V_j \right) \quad (4)$$

where,

$\mathrm{J} = Set\ of\ disambiguated\ Words$

$\theta_i = BelongingnessToDominantConcept\ (S_i)$

$V_i = P(S_i \mid word)$

$W_{ij} = CorpusCooccurences\ (S_i, S_j) * 1/WNConceptualDistance(S_i, S_j) * 1/WNSemanticGraphDistance(S_i, S_j)$

The component $\theta_i * V_i$ of the energy due to the self activation of a neuron can be compared to the corpus specific sense of a word in a domain. The other component $w_{ij} * V_i * V_j$ coming from the interaction of activations can be compared to the score of a sense due to its interaction in the form of corpus co-occurrence, conceptual distance, and wordnet-based semantic distance with the senses of other words in the sentence. The first component thus captures the rather *static corpus sense*, whereas the second expression brings in the sentential context.

Algorithm 1: *performIterativeWSD(sentence)*

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 select that sense for a word which maximizes the score given by Equation (4)

Algorithm1: Iterative WSD

IWSD is clearly a greedy algorithm. It bases its decisions on already disambiguated words, and ignores words with higher degree of polysemy. For example, while disambiguating bisemous words, the algorithm uses only the monosemous words.

7.2 Modified PageRank algorithm

Rada Mihalcea (2005) proposed the idea of using PageRank algorithm to find the best combination of senses in a sense graph. The nodes in a sense graph correspond to the senses of all the words in a sentence and the edges depict the strength of interaction between senses. The score of each node in the graph is then calculated using the following recursive formula:

$$score(S_i) = (1-\mathrm{d}) + \mathrm{d} * \sum_{S_j \in In(S_i)} \frac{W_{ij}}{\sum_{S_k \in Out(S_i)} W_{jk}} * Score(S_j)$$

Instead of calculating W_{ij} based on the overlap between the definition of senses S_i and S_j as proposed by Rada Mihalcea (2005), we calculate the edge weights using the following formula:

$$W_{ij} = CorpusCooccurences(S_i, S_j) * 1/WNConceptualDistance(S_i, S_j) * 1/WNSemanticGraphDistance(S_i, S_j) * P(S_i \mid word_i) * P(S_j \mid word_j)$$

$d = damping\ factor\ (typically\ 0.85)$

This formula helps capture the edge weights in terms of the corpus bias as well as the interaction between the senses in the corpus and wordnet. It should be noted that this algorithm is *not greedy*. Unlike IWSD, this algorithm allows all the senses of all words to play a role in the disambiguation process.

8 Experimental Setup:

We tested our algorithm on tourism corpora in 3 languages (*viz.*, Marathi, Bengali and Tamil) and health corpora in 1 language (Marathi) using projections from Hindi. The corpora for both the domains were manually sense tagged. A 4-fold cross validation was done for all the languages in both the domains. The size of the corpus for each language is described in Table 4.

Language	**# of polysemous words (tokens)**	
	Tourism Domain	**Health Domain**
Hindi	50890	29631
Marathi	32694	8540
Bengali	9435	-
Tamil	17868	-

Table 4: Size of manually sense tagged corpora for different languages.

Table 5 shows the number of synsets in MultiDict for each language.

Language	**# of synsets in MultiDict**
Hindi	29833
Marathi	16600
Bengali	10732
Tamil	5727

Table 5: Number of synsets for each language

Algorithm	**Language**					
	Marathi			**Bengali**		
	P %	**R %**	**F %**	**P %**	**R %**	**F %**
IWSD (training on self corpora; no parameter projection)	81.29	80.42	80.85	81.62	78.75	79.94
IWSD (training on Hindi and reusing parameters for another language)	**73.45**	**70.33**	**71.86**	**79.83**	**79.65**	**79.79**
PageRank (training on self corpora; no parameter projection)	79.61	79.61	79.61	76.41	76.41	76.41
PageRank (training on Hindi and reusing parameters for another language)	**71.11**	**71.11**	**71.11**	**75.05**	**75.05**	**75.05**
Wordnet Baseline	58.07	58.07	58.07	52.25	52.25	52.25

Table 6: Precision, Recall and F-scores of IWSD, PageRank and Wordnet Baseline. Values are reported with and without parameter projection.

9 Results and Discussions

Table 6 shows the results of disambiguation (precision, recall and F-score). We give values for two algorithms in the tourism domain: IWSD and PageRank. In each case figures are given for both with and without parameter projection. The wordnet baseline figures too are presented for the sake of grounding the results.

Note the lines of numbers in bold, and compare them with the numbers in the preceding line. This shows the fall in accuracy value when one tries the parameter projection approach in place of self corpora. For example, consider the F-score as given by *IWSD* for Marathi. It degrades from about 81% to 72% in using parameter projection in place of self corpora. Still, the value is much more than the baseline, *viz.*, the wordnet first sense (a typically reported baseline).

Coming to PageRank for Marathi, the fall in accuracy is about 8%. Appendix A shows the corresponding figure for Tamil with IWSD as 10%. Appendix B reports the fall to be 11% for a different domain- Health- for Marathi (using *IWSD*).

In all these cases, even after degradation the performance is far above the wordnet baseline. This shows that one could trade accuracy with the cost of creating sense annotated corpora.

10 Conclusion and Future Work:

Based on our study for 3 languages and 2 domains, we conclude the following:

(i) Domain specific sense distributions- if obtainable- can be exploited to advantage.

(ii) Since sense distributions remain same across languages, it is possible to create a disambiguation engine that will work even in the absence of sense tagged corpus for some resource deprived language, provided (a) there are aligned and cross linked sense dictionaries for the language in question and another resource rich language, (b) the domain in which disambiguation needs to be performed for the resource deprived language is the same as the domain for which sense tagged corpora is available for the resource rich language.

(iii) Provided the accuracy reduction is not drastic, it may make sense to trade high accuracy for the effort in collecting sense marked corpora.

It would be interesting to test our algorithm on other domains and other languages to conclusively establish the effectiveness of parameter projection for multilingual WSD.

It would also be interesting to analyze the contribution of corpus and wordnet parameters independently.

References

Agirre Eneko & German Rigau. 1996. *Word sense disambiguation using conceptual density*. In Proceedings of the 16th International Conference on Computational Linguistics (COLING), Copenhagen, Denmark.

Dipak Narayan, Debasri Chakrabarti, Prabhakar Pande and P. Bhattacharyya. 2002. *An Experience in Building the Indo WordNet - a WordNet for Hindi*. First International Conference on Global WordNet, Mysore, India.

Eneko Agirre & Philip Edmonds. 2007. *Word Sense Disambiguation Algorithms and Applications*. Springer Publications.

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

Hindi Wordnet.
http://www.cfilt.iitb.ac.in/wordnet/webhwn/

J. J. Hopfield. April 1982. *"Neural networks and physical systems with emergent collective computational abilities"*, Proceedings of the National Academy of Sciences of the USA, vol. 79 no. 8 pp. 2554-2558.

Lee Yoong K., Hwee T. Ng & Tee K. Chia. 2004. *Supervised word sense disambiguation with support vector machines and multiple knowledge sources*. Proceedings of Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, 137-140.

Lin Dekang. 1997. *Using syntactic dependency as local context to resolve word sense ambiguity*. In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics (ACL), Madrid, 64-71.

Michael Lesk. 1986. *Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone*. In Proceedings of the 5th annual international conference on Systems documentation, Toronto, Ontario, Canada.

Mihalcea Rada. 2005. *Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling*. In Proceedings of the Joint Human Language Technology and Empiri-

cal Methods in Natural Language Processing Conference (HLT/EMNLP), Vancouver, Canada, 411-418.

Ng Hwee T. & Hian B. Lee. 1996. *Integrating multiple knowledge sources to disambiguate word senses: An exemplar-based approach*. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), Santa Cruz, U.S.A., 40-47.

Peter F. Brown and Vincent J.Della Pietra and Stephen A. Della Pietra and Robert. L. Mercer. 1993. *The Mathematics of Statistical Machine Translation: Parameter Estimation*. Computational Linguistics Vol 19, 263-311.

Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra and Aditya Sharma. 2008. *Synset Based Multilingual Dictionary: Insights, Applications and Challenges*. Global Wordnet Conference, Szeged, Hungary, January 22-25.

Resnik Philip. 1997. *Selectional preference and sense disambiguation*. In Proceedings of ACL Workshop on Tagging Text with Lexical Semantics, Why, What and How? Washington, U.S.A., 52-57.

Roberto Navigli, Paolo Velardi. 2005. *Structural Semantic Interconnections: A Knowledge-Based Approach to Word Sense Disambiguation*. IEEE Transactions On Pattern Analysis and Machine Intelligence.

Sergei Nirenburg, Harold Somers, and Yorick Wilks. 2003. *Readings in Machine Translation.* Cambridge, MA: MIT Press.

Véronis Jean. 2004. *HyperLex: Lexical cartography for information retrieval*. Computer Speech & Language, 18(3):223-252.

Walker D. and Amsler R. 1986. *The Use of Machine Readable Dictionaries in Sublanguage Analysis*. In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press, pp. 69-83.

Yarowsky David. 1994. *Decision lists for lexical ambiguity resolution: Application to accent restoration in Spanish and French*. In Proceedings of the 32nd Annual Meeting of the association for Computational Linguistics (ACL), Las Cruces, U.S.A., 88-95.

Yarowsky David. 1995. *Unsupervised word sense disambiguation rivaling supervised methods*. In Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics (ACL), Cambridge, MA, 189-196.

Appendix A: Results for Tamil (Tourism Domain)

Algorithm	P %	R %	F %
IWSD (training on Tamil)	89.50	88.18	88.83
IWSD (training on Hindi and reusing for Tamil)	84.60	73.79	78.82
Wordnet Baseline	65.62	65.62	65.62

Table 7: Tamil Tourism corpus using parameters projected from Hindi

Appendix B: Results for Marathi (Health Domain)

Algorithm Words	P %	R %	F %
IWSD (training on Marathi)	84.28	81.25	82.74
IWSD (training on Hindi and reusing for Marathi)	75.96	67.75	71.62
Wordnet Baseline	60.32	60.32	60.32

Table 8: Marathi Health corpus parameters projected from Hindi