Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting

Mitesh M. Khapra
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
Maharashtra, India
miteshk@cse.iitb.ac.in

Pushpak Bhattacharyya
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
Maharashtra, India
pb@cse.iitb.ac.in

Shashank Chauhan
Dharamsinh Desai University
Nadiad-387001
Gujarat, India
shashank1.iitm@gmail.com

Soumya Nair
Dharamsinh Desai University
Nadiad-387001
Gujarat, India
soumya.iitm@gmail.com

Aditya Sharma
Department of Computer Science and Engineering
Indian Institute of Technology, Bombay
Powai, Mumbai 400076
Maharashtra, India
adityas@cse.iitb.ac.in

Abstract

We report here our work on Domain Specific Iterative Word Sense Disambiguation (WSD) for nouns, adjectives and adverbs in the trilingual setting of *English*, *Hindi*[1] and *Marathi*[2]. The methodology proposed relies on dominant senses of words in specified domains. Starting from monosemous words we iteratively disambiguate bi, tri and polysemous words. We combine corpus biases for senses along with information in wordnet graph structure to arrive at the sense decisions. To the best of our knowledge, this is the first attempt at a large scale multilingual WSD involving Indian languages and English. The accuracy values of approximately 65% (F1-score) for all the three languages compares well with the state of the art.

1 Introduction

In a significant development in NLP R & D in India, large consortia projects have been initiated in the areas of Cross Lingual Search, English to Indian Language Machine Translation and Indian Language to Indian Language Machine Translation. A multilingual wordnet-synset-based dictionary forms the heart of these large scale activities, with multilingual word sense disambiguation (WSD) forming a critical component of the system. The domains in focus for these projects are *Tourism* and *Health.*

1.1. Multilingual Cross Linked Dictionary

A novel and effective method of storage and usage of dictionary in a multilingual framework was proposed (Rajat Mohanty *et al.*, 2008). Table 1 shows the structure of the multilingual dictionary.

[1] Hindi is the official national language of India. The language and its close cousin Urdu are spoken by approximately 500 million people in the world.
[2] Marathi is the official language of Maharashtra, a state in Western India. The language has close to 20 million speakers in the world.

Proceedings of ICON-2008: 6th International Conference on Natural Language Processing
Macmillan Publishers, India. Also accessible from http://ltrc.iiit.ac.in/proceedings/ICON-2008

Concepts	L1 (English)	L2 (Hindi)	L3 (Marathi)
Concept ID: Concept description	(W_1, W_2, W_3, W_4)	(W_1, W_2, W_3, W_4, W_5 W_6, W_7, W_8)	(W_1, W_2, W_3, W_4, W_5 W_6, W_7, W_8, W_9, W_{10})
02038: a typical star that is the source of light and heat for the planets in the solar system	(sun)	(सूर्य (soorya), सूरज (sooraj), भानु (bhaanu), दिवाकर (divaakar), भास्कर (bhaaskar), प्रभाकर (prabhaakar), दिनकर (dinkar), रवि (ravi), आदित्य (aaditya), दिनेश (dinesh), सविता (savitaa), पुष्कर (pushkar), मिहिर (mihir), अंशुमान (anshuman), अंशुमाली (anshumaalii))	(सूर्य (soorya), भानु(bhaanu), दिवाकर(divaakar), भास्कर (bhaaskar), प्रभाकर(prabhaakar), दिनकर(dinkar), मित्र (mitra), मिहिर(mihir), रवि (ravi), दिनेश (dinesh), अर्क (ark), सविता (savitaa))
04321: a youthful male person	(male_child, boy)	(लड़का (ladkaa), बालक (baalak), बच्चा (bachchaa), छोकड़ा (chokdaa), छोरा (choraa), छोकरा (chokraa), लौंडा (laundaa))	(मुलगा (mulgaa), पोरगा (porgaa), पोर (por), पोरगे (porge))

Table 1: Proposed multilingual dictionary model

Given a row, the first column is the pivot for *n* number of languages describing a concept. Each concept is assigned a unique ID. The columns (2-4) show the appropriate words expressing the concepts in respective languages. To express the concept *'04321: a youthful male person'*, there are two lexical elements in English, which constitute a *synset.* There are seven words in Hindi which form the Hindi synset, and four words in Marathi which constitute the Marathi synset. The members of a particular synset are arranged in the order of their frequency. The proposed model thus defines an *M X N* matrix as the multilingual dictionary, where each row is for a concept and each column for a particular language.

The proposed framework entails in it the problem of WSD and Lexical Choice. The former requires a correct row to be identified given the source language word. The latter demands that appropriate word is chosen from the mapped synset (as illustrated in Figure 1), once the correct row has been identified.

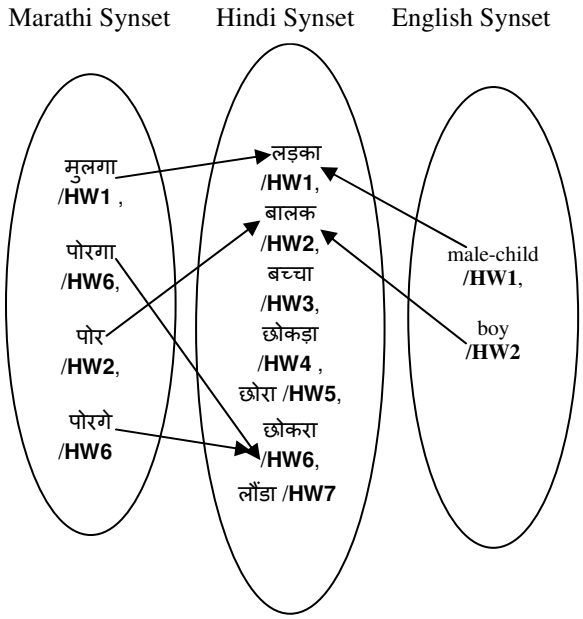

Figure 1: Illustration of aligned synset members for the concept: a youthful male person

The roadmap of the paper is as follows. Section 2 is on literature survey. Section 3 describes the features used in the WSD algorithm- a critical part to understand the rest of the paper. Section 4 gives the actual algorithm. Section 5 discusses the effort at achieving language independence. Experiments and results are presented in section 6. Section 7 concludes the paper.

2 Literature Survey

Major WSD approaches proposed till date can be broadly classified as *Knowledge Based Approaches and Machine Learning Based Approaches*.

Knowledge based approaches such as WSD using Selectional Preferences (Resnik Philip, 1997), Lesk's algorithm (Michael Lesk. 1986), Walker's algorithm (Walker D. & Amsler R., 1986), WSD using conceptual density (Agirre Eneko & German Rigau, 1996) and WSD using Random Walk Algorithm (Mihalcea Rada, 2005) are easy

to implement as they require a simple lookup of a knowledge resource like a Machine Readable Dictionary. Further, they do not require any corpus-tagged or untagged-, since no training is involved. However, these algorithms suffer from poor accuracies because of their complete dependence on dictionary defined senses which do not provide enough surface cues about the selectional preferences of different senses of a word (For example, we would expect the words "cigarette" and "ash" to co-occur as they are semantically related. However if we read the dictionary definitions of these words we find that neither has a reference to the other). Overlap based algorithms typically suffer from sparse overlap, as dictionary definitions are generally small in length. Another knowledge based approach proposed by Agirre Eneko & German Rigau (1996) is to use the conceptual distance between the senses of the context words and the sense of the target word as a measure for disambiguation. They proposed a formula for conceptual distance which is inversely proportional to the length of the path between two synsets in the wordnet (Fellbaum, C. 1998) graph and directly proportional to the depth of the two synsets in the wordnet hierarchy.

The study of machine learning based algorithms (supervised as well as unsupervised) suggested that extracting "sense definitions" or "usage patterns" from corpora helps in improving the accuracy of WSD. However, most supervised algorithms which perform very well are not general purpose WSD systems, but word specific classifiers (for example, WSD using SVM (Lee et al. 2004), Exemplar based WSD (Ng Hwee T. & Hian B. Lee. 1996) and Yarowsky's (1994) decision list algorithm). Further, some of these algorithms are not able to distinguish between the finer senses of a word. Finally, the requirement of a large training corpus renders these algorithms unsuitable for resource poor languages of which Indian languages are examples.

The study of semi-supervised and unsupervised machine learning algorithms suggests that they are capable of performing at par with supervised algorithms (David Yarowsky, 1995). The fact that these algorithms can work with very little or no tagged data makes them suitable for languages like Hindi. But here again it is difficult to build general purpose broad coverage models. Most semi-supervised and unsupervised algorithms which give very good performance are word specific classifiers (for example, Yarowsky's (1995) semi-supervised decision list algorithm and Hyperlex (Véronis Jean, 2004)). It was further observed that models (for example, Lin's algorithm (Lin Dekang, 1997)) that exploit syntactic dependencies between words are able to perform large scale disambiguation (*i.e.*, they act as generic classifiers) and at the same time give reasonably good accuracies.

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 as well as draw syntactic generalizations from minimally tagged corpora. *These methods seem to be the most suitable in building general purpose broad coverage classifiers*. This observation has been the motivation for our work.

1 Domain Specific Language Independent Iterative WSD:

Our primary goal has been to develop an algorithm to perform WSD *within a domain.* We combine sense distributions and sense co-occurrences learnt from corpora with semantic relations in wordnet to develop a robust WSD engine.

3.1. Features used for WSD

(i) Domain Specific Sense Distributions: Domain-specific most frequent senses of words are identified from sense tagged corpora. These statistics are then used as input for WSD. As an example, let us consider the sense distributions for **सुविधा** {suvidhaa} *(convenience)* which is a frequently occurring word in tourism corpus.

सुविधा {suvidhaa} *(convenience)*
Most Frequent Sense in Hindi Wordnet (Dipak Narayan *et al.*, 2002)
(http://www.cfilt.iitb.ac.in/wordnet/webhwn/):
Sense ID: 3530
Category: NOUN
Gloss:
वह स्थिति जिस में कोई काम करने में कुछ कठिनता या अड़चन न हो:~:{vaha sthiti jis mein koi kaam karne mein kuch kathinta yaa aDchan na ho}

(that state which in any work do in any difficulty or problem no <vaux>)
(a state in which there is no difficulty or problem in completing any work)
Synset-Members:
सुविधा {suvidhaa} *(convenience)*, सुभीता {subhiita} *(convenience)*, सुगमता {sugamtaa} *(convenience)*

Most Frequent Sense in the Domain:
Sense ID: 28213
Category: NOUN
Gloss:
वह सेवा जो एक संस्था या कोई उपकरण आपको देता है
{vaha sevaa jo ek sansthaa yaa upkaraN aapko detaa hain}
(that service which one institution or instrument you gives <vaux>)
(that service which is provided by an institution or an instrument)
Synset-Members:
सुविधा {suvidhaa} *(facility)*

As seen in the above example, for some words the domain specific frequent sense is different from the most frequent sense listed in wordnet. For some other words the domain specific frequent sense may be the same as the most frequent sense listed in wordnet. However, in either case learning this statistics from the corpus will be beneficial, as it will only improve the results of our disambiguation algorithm (by creating a bias towards the domain specific most frequent sense).

It was further observed that within a domain words tend to be monosemic. This observation was based on a statistical analysis of the Tourism and Health corpora for Hindi and Marathi (*vide* figure 2.a and figure 2.b)

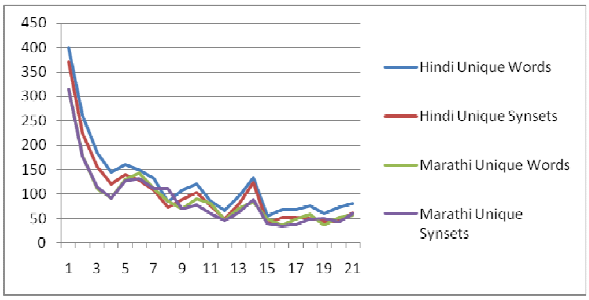

Figure 2.a: No. of Unique Words V/s No. of Documents and No. of Unique Synsets V/s No. of Documents for Hindi and Marathi Health corpus.

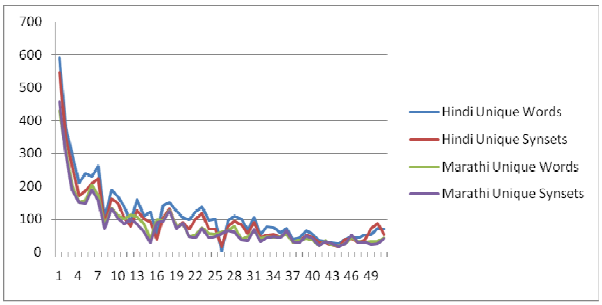

Figure 2.b: No. of Unique Words V/s No. of Documents and No. of Unique Synsets V/s No. of Documents for Hindi and Marathi Tourism corpus.

As we see more and more documents from the same domain, the number of new words as well as the number of new synsets encountered in each new document decreases. This shows that words in the same domain tend to appear in the same sense again and again.

Of interest are the sharp spikes in the graph, *e.g.*, for document 9. Further analysis of these documents showed that the surge of synsets was because of a change in the sub-domain of the document. Document 9 describes a tourist location which had a pro-war history. Hence there were a lot of references to words from military domain like *cavalry, infantry, weapons, etc.* which were otherwise not observed in the tourism domain. Apart from a few such anomalies the behavior is same for both the languages in both the domains.

We also calculated the average degree of polysemy of the words within the domain by counting the number of different senses of a word appearing in the domain corpora (around 8000 sentences were manually sense tagged by lexicographers). These figures were compared with the average degree of polysemy of the same words according to the number of senses listed in the wordnet. The results are summarized in Table 2.a and Table 2.b.

Domain	**No. of Unique Words**	
	Hindi	Marathi
Tourism	5976	4280
Health	2603	1962

Table 2.a: No. of unique words in the Tourism and Health corpus for Hindi and Marathi

Domain	**Average degree of polysemy calculated from corpus**		**Average degree of polysemy calculated from wordnet**	
	Hindi	Marathi	Hindi	Marathi
Tourism	1.20	1.12	2.21	1.84
Health	1.13	1.08	2.38	2.01

Table 2.b: Average degree of polysemy calculated from corpus and Wordnet

These observations vindicate the fact that the domain distribution of senses can in general be very different from the general distribution.

(ii) Dominant Concepts within a domain: We define *Dominant Concepts* as follows:

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

When we have to *choose* between two candidate synsets of a word, we give a higher weightage to the sense which belongs to the hierarchy of dominant concepts.

Dominant concepts obtained for representative "Health" and "Tourism" corpora are listed in Table 3 below.

Tourism	**Health**
{place, country, city, area}	{doctor, nurse}
{flora, fauna}	{patient}
{mode of transport}	{disease}
{fine arts}	{treatment}

Table 3: Dominant concepts from the Tourism and Health corpus

To illustrate the use of dominant concept, for a word like "सागर" {saagar} (sea), which has two senses, our algorithm will give a higher weightage to Sense 2650, since it occurs in the sub-tree of the domain specific dominant concept { क्षेत्र (kshetra) *(area)*}. See Figure 3 below.

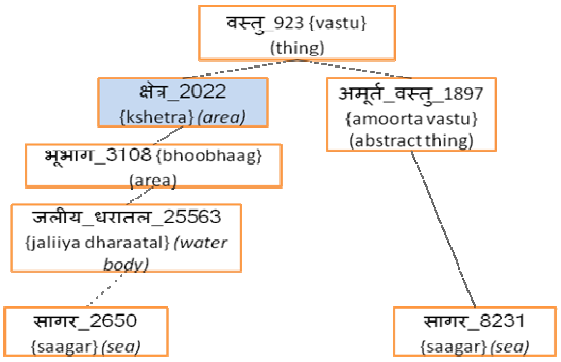

Figure 3: Hierarchy of the 2 senses of the word "सागर" {saagar} (sea)

सागर {saagar} *(sea)*
Sense ID: 28322
Category: NOUN
Gloss:
खारे पानी की वह विशाल राशि जो पृथ्वी के स्थल भाग को चारों ओर से घेरे हुए है
{khaare paanii kii vaha vishaal raashi jo prathvi ke sthal bhaag ko chaaron oar se ghere hue hain}
(salty water of that huge collection that earth of land part of four-erg from surrounded <vaux>)
(That huge expanse of salty water that surrounds land from all sides)

सागर {saagar} *(sea in metaphorical sense)*
Sense ID: 8231
Category: NOUN
Gloss:
किसी विषय के ज्ञान या गुण आदि का बहुत बड़ा आगार
{kisii vishay ke gyaan yaa guN aadi kaa bahut baDa aagaar}
(some type of knowledge or quality etc. of very big collection)
(Knowledge or quality of any type which is apparently limitless in quantity or volume)

(iii) Corpus co-occurrence frequency of senses: A common feature used by several WSD algorithms is to find the frequencies of **words** co-occurring with a particular sense of the target word (also known as the "Bag of Words" approach). We made a slight modification to this heuristic and concentrated on the **senses** which co-occur with a particular sense of the target word. This feature is expected to be better than "Bag of Words" approach. For example, the synset {हॉटल (hoTal) *(hotel)*} has a high co-occurrence with the synset {भोजन (bhojan) *(food)*, खाना (khaana) *(food)*} but the co-occurrence of individual words {हॉटल (hoTal) *(hotel)*} and भोजन (bhojan) *(food)* or {हॉटल (hoTal) *(hotel)*} and खाना (khaana) *(food)* is less than the co-occurrence of the two synsets. The same is true for synsets like {समय (samay) *(time)*} and {अच्छा (acchaa) *(good)*, बढ़िया (badhiyaa) *(good)*, ठीक (theek) *(alright)*} where the *co-occurrence between the synsets is higher than the co-occurrence between the individual words.*

(iv) Conceptual distance between senses: Equation (1) below defines *Conceptual Distance* be-

tween a pair of synsets, motivated by (Agirre Eneko & German Rigau, 1996)

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

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

We found several instances in the corpus where the conceptual distances proved to be effective in disambiguation. For example, if the word "नदी" {nadii} *(river)* (which is monosemic) appeared in the context of the polysemous word "सागर" {saagar} *(sea)* as in the sentence:

"आधुनिक नहर से भिन्न, प्राचीन नहरें लाल **सागर** को नील **नदी_4430** से जोड़ती थीं"

{aadhunik nahar se bhinna, praachiin naharein laal saagar ko niil nadii se jodtii thii}
(modern canal from different, ancient canals Red sea of Nile river to connect <vaux>)
(Unlike modern canals, ancient canals connected Red sea to the Nile river)

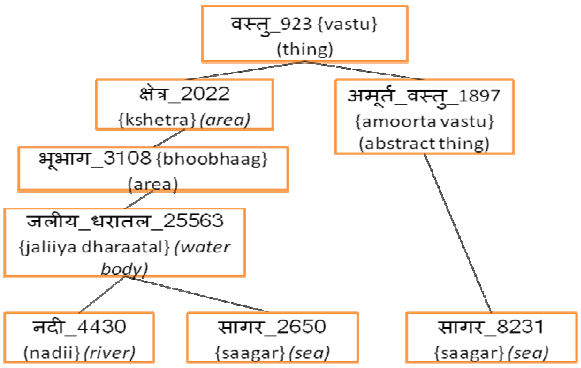

Figure 4: Conceptual distance between the disambiguated word "नदी" {nadii} *(river)* and the two senses of the word "सागर" {saagar} *(sea)*

In the above example, the disambiguated sense of "नदी" {nadii} *(river)* can be used to choose between the two senses of the word "सागर" {saagar} *(sea)* as shown in Figure 4.

Based on the conceptual distance formula in equation (1), we can say that the conceptual distance between the synsets 2650 and 4430 is less than the distance between the synsets 8231 and 4430. Hence, the synset 2650 should be given a higher rank as compared to the synset 8231.

(v) Semantic Graph Distance: Semantic Graph distance is defined as the shortest path length between two synset nodes in the wordnet graph. An edge on this shortest path can be **any semantic relation** (as opposed to conceptual distance where the path consists of only the hyponymy-hypernymy relations). We thus exploited the semantic interconnections between synsets as captured by the graph-like structure of wordnet. For example, wordnet captures the semantic relation (MODIFIES_NOUN) between the synset {स्वस्थ (swastha) (healthy)}:1831 and the synset {जंतु (jantu) (organism)}:748 as well as the semantic relation (HYPONYMY) between the synset {आदमी (aadmii) (man)}:3389 and the synset {जंतु (jantu) (organism)}:748. If we represent the synsets as nodes and the relations as edges, we get a graph as shown in Figure 5. We can now infer the relation between the synsets {स्वस्थ (swastha) (healthy)}:1831 and {आदमी (aadmii) (man)}:3389 which are not directly connected, but a path exists between them in the semantic graph. The semantic relatedness of the synsets {स्वस्थ (swastha) (healthy)}:1831 and {आदमी (aadmii) (man)}:3389 would be inversely proportional to the length of the path between them and can be used as a score for performing WSD.

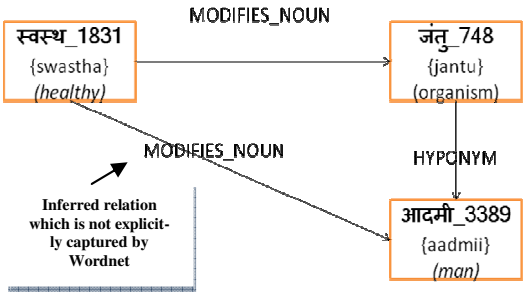

Figure 5: Semantic relations inferred from a semantic graph (wordnet)

2 Proposed Algorithm:

Ours is an iterative method. In the first iteration all the monosemic words are marked (these marked senses act as the seed input for the algorithm). In the next iteration bisemic words are disambiguated followed by trisemic words and so on. Disambiguating words in the order of their degree of polysemy ensures that more and more disambiguated words are available as input at every stage, as we move towards more and more ambiguous words. Thus, unlike most other WSD algorithms, this algorithm does not use ambiguous words as clues for disambiguating other words. **At each stage, the input to the algorithm consists of a set of disambiguated words**. The candidate synset which maximizes Equation (2) (which combines all the features described in section 3.1) is selected as the most appropriate synset at each stage:

Algorithm 1: *performIterativeWSD(sentence)*

1. Tag all monosemic words in the sentence.
2. Iteratively disambiguate the remaining words in the sentence in the order of their degree of polysemy.
3. At each stage select that synset for a word which maximizes the following score:

$$\underset{S \in candidateSenses}{argmax} \Big[P(S \mid word) * BelongingnessToDominantConcept(S) * \sum_{w \in disambiguatedWords} CorpusCooccurrence(S, S_w) * \sum_{w \in disambiguatedWords} 1/WNConceptualDistance(S, S_w) * \sum_{w \in disambiguatedWords} 1/WNSemanticGraphDistance(S, S_w) \Big] \quad (2)$$

Algorithm1: Iterative WSD

We note that:

- $P(S \mid word)$ helps bias the score towards the domain-specific most frequent sense of the word.
- *BelongingnessToDominantConcept*(S_w) helps bias the score towards synsets belonging to domain specific dominant concepts.
- *CorpusCooccurrence*(S, S_w) captures selectional preferences from a corpus (typically not captured by wordnets).
- *WNConceptualDistance*(S, S_w) captures conceptual density of nouns.
- *WNSemanticGraphDistance*(S, S_w) captures semantic relations between senses as stored in the wordnet

(*Monosemic words are used only as the seed input for the algorithm and are not included while calculating the precision and recall of the algorithm.)*

3 Towards Language Independence:

An interesting idea we investigated is how the features described in section 3.1 can be learnt from the sense tagged corpus and the wordnet of one language L_1 and **reused** to perform WSD for sentences of language L_2. This has been made possible by the use of the multilingual dictionary framework described in section 1.1.

Domain Specific Sense Distributions: Consider the example of two senses of the Marathi word *अखेर {akher} (end)* and the corresponding cross-linked words in Hindi (figure 6 below):

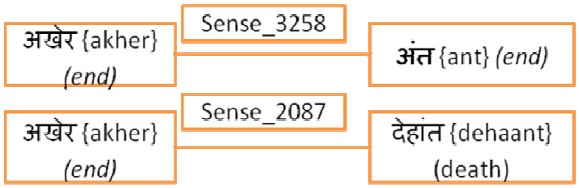

Figure 6: Two senses of the Marathi word "अखेर" {akher} (end, death) **and the corresponding cross-linked words in Hindi**

Based on the above cross-linkages we can say that the number of instances of the word "**अखेर**" {akher} (end) having sense 3258 in the Marathi corpus would be proportional to the number of instances

of the word अंत {ant} (end) having sense 3258 in the Hindi corpus. Thus the propability of the word "**अखेर**" {akher} (end) having the sense 3258 can be calculated as,

$$Pr_{\text{Marathi}}(\text{Sense}_{3258} \mid \text{अखेर } \{\text{akher}\}\ (\text{end})) \ \alpha$$

$$\frac{\text{No. of occurrences of (अंत } \{\text{ant}\}\ (\text{end}), 3258) \text{ in Hindi tagged corpus}}{\text{(No. of occurrences of (अंत } \{\text{ant}\}\ (\text{end}),\ 3258) \text{ in Hindi tagged corpus + No. of occurrences of (देहांत } \{\text{dehaant}\}\ (\text{death}),\ 2087) \text{ in Hindi tagged corpus)}} \quad (3)$$

In general, the following formula can be used for calculating sense distributions of Marathi words using parallel sense marked Hindi corpus.

$$Pr_{\text{Marathi}}(\text{Sense}_i \mid \text{Marathi\_word}) \ \alpha$$

$$\frac{\text{No. of occurrences of (cross\_linked\_hindi\_word, Sense}_i\text{) in Hindi tagged corpus}}{\sum_{S_i \in \text{all senses}} \text{No. of occurrences of (cross\_linked\_hindi\_word, } S_i\text{) in Hindi tagged corpus}} \quad (4)$$

Note that we are not interested in the *exact* sense distribution of the words, but only in the *relative* distribution, so that the score calculated using Equation (2) can be biased towards domain specific frequent senses. Hence, the above formula is sufficient for our purpose as long as it maintains the relative rank of the different senses of the word.

To prove that the above formula indeed serves the purpose, we learnt the statistics for some Marathi words from a sense tagged Marathi corpus and compared the statistics with the sense distributions learnt for these same words from a parallel sense tagged Hindi corpus using the above formula. The results are summarized in Table 4.

Sr. No	Marathi Word	Synset	P(S|word) as learnt from sense tagged Marathi corpus	P(S|word) as learnt from parallel sense tagged Hindi corpus
1	गोड {goD} *(sweet)*	{गोड (goD) *(sweet)*, सुरेल (surel) *(sweet)* } – sounds sweet	0.063	0.056
		{गोड (goaD) *(sweet)*, मधुर (madhur) *(sweet)* } – tastes sweet	0.937	0.944
2	मान {maan} *(neck, respect)*	{मान (maan) *(neck)* , ग्रीवा (griiva) (neck)} – neck	0.4	0.36
		{आब (aab) (respect), मान (maan) (respect)} – respect	0.6	0.64
3	आवड {aavaD} (liking, hobby)	{पसंती (pasan-ti) *(liking)*, आवड (aavaD) (liking)} – liking	0.24	0.21
		{आवड (aavaD) *(hobby)*, शौक (shauk) *(hob-by)*} – hobby	0.76	0.79
4	उत्तर {uttar} (north, answer)	{उत्तर (uttar) (north)} – north	0.94	0.98
		{उत्तर (uttar) *(answer)*, जबाब (jabaab) *(answer)*} – answer	0.06	0.02

Table 4: Comparison of the sense distributions of some Marathi words learnt from Marathi sense tagged corpus with those learnt from parallel Hindi sense tagged corpus.

It is clear that the relative rank of the senses for a particular word is maintained, independent of whether the *P(.)* values are from the Marathi corpus or from the parallel Hindi corpus.

Dominant Concepts within a domain: We found that concepts like *{place, country, city, area}*, *{flora, fauna}*, *{mode of transport}* and *{fine arts}*, which are dominant in Hindi tourism corpus, are dominant in Marathi tourism corpus too. Further, the Multilingual Dictionary Framework (section 1.1) ensures that the synset ids remain the same across languages. Hence, the dominant synset ids learnt for one language can be used as dominant synset ids for other languages also.

Corpus co-occurrence frequency of senses: The co-occurrence of senses should remain the same

across languages. For example, the co-occurrence of the Hindi synsets {होटल (hoTal) *(hotel)*} and {भोजन (bhojan) *(food),* खाना (khaana) *(food)*} in the Hindi corpus should be the same as (or proportional to) the co-occurrence between the corresponding Marathi synsets {हॉटल (hoTal) *(hotel)*} and {जेवण (jevaN) *(food)*, भोजन (bhojan) *(food)*} in the Marathi corpus.

Conceptual distance between senses: In the Multilingual Dictionary Framework, the hypernymy hierarchies for all languages are borrowed from the Hindi Wordnet (as the synset ids remain the same across languages). Since the conceptual distance depends only on the Hypernymy hierarchical structure of the wordnet, it very often is the same across languages for highly common synsets. Thus, revisiting the example illustrated in Figure 4, the conceptual distance between the synsets 2650 {नदी (nadii) *(river)*} and 4430 {सागर (saagar) *(sea*)} are same in Hindi and Marathi.

Semantic Graph Distance: As argued in case of conceptual distance, semantic graph distance also tends to remain same for common synsets across languages. Revisiting the example illustrated in Figure 5, the semantic graph distance between the synsets {स्वस्थ (swastha) *(healthy)*}:1831 and {आदमी (aadmii) *(man)*}:3389 would be the same in Hindi and Marathi.

4 Experiments:

We tested our algorithm on tourism corpus for 3 languages (*viz.,* Hindi, Marathi and English) and health corpus for 2 languages (*viz.*, Hindi and Marathi). We used two different parameter settings. In one case we consider the sense distributions (*i.e., P(S|word)*) learnt from a corpus only if the number of instances of the word in the corpus is greater than a certain threshold *(t: we used t=30).*

In the second case we consider the sense distributions for all the words irrespective of the number of instances of the word in the corpus *(i.e., t=0)*). The sole purpose of choosing two different values of the threshold is to highlight the effect of the *P(S|word)* factor in disambiguation.

Lowering the threshold brings in the less frequently occurring words. For such words the only hope of disambiguation is through the *P(S|word)* factor.

A 4-fold cross validation was done for all the languages in both the domains. The results of our algorithms were compared with the wordnet baseline (*i.e.*, selecting the first sense from wordnet) as well as the corpus baseline (*i.e.*, selecting the most frequent sense from the corpus). We first describe the different parameter settings used and then discuss the results of our experiments:

6.1. Results:

Tables 5.a to 5.g show a summary of the results of our experiments.

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	38649	71.2	62.1	66.4
Iterative WSD *(t = 0)*	38649	74.7	73.4	74.1
Wordnet Baseline	38649	61.1	61.1	61.1
Corpus Baseline	38649	79.9	75.6	77.7

Table 5.a: Average 4-fold cross validation results for Hindi Tourism corpus

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	24823	60.2	36.0	45.1
Iterative WSD *(t = 0)*	24823	67.5	62.0	64.6
Wordnet Baseline	24823	60.7	60.7	60.7
Corpus Baseline	24823	72.7	63.7	67.9

Table 5.b: Average 4-fold cross validation results for English Tourism corpus

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	17762	71.5	61.0	65.8
Iterative WSD *(t = 0)*	17762	75.1	73.7	74.4
Wordnet Baseline	17762	51.5	51.5	51.5
Corpus Baseline	17762	81.4	77.2	79.3

Table 5.c: Average 4-fold cross validation results for Marathi Tourism corpus

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	17746	69.6	56.8	62.5
Iterative WSD *(t = 0)*	17746	71.7	66.7	69.1
Wordnet Baseline	17746	51.4	51.4	51.4
Corpus Baseline	17746	76.3	65.7	70.6

Table 5.d: Average 4-fold cross validation results for Marathi Tourism corpus using features learnt from Hindi Tourism corpus.

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	10532	70.1	47.5	56.6
Iterative WSD *(t = 0)*	10532	76.4	72.1	74.2
Wordnet Baseline	10532	57.8	57.8	57.8
Corpus Baseline	10532	78.1	70.7	74.2

Table 5.e: Average 4-fold cross validation results for Hindi Health corpus

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	6145	75.6	55.0	63.6
Iterative WSD (*t = 0*)	6145	80.6	76.9	78.7
Wordnet Baseline	6145	57.6	57.6	57.6
Corpus Baseline	6145	83.9	77.1	80.4

Table 5.f: Average 4-fold cross validation results for Marathi Health corpus

Algorithm	Words	P %	R %	F %
Iterative WSD (*t = 30*)	6137	72.4	42.9	53.9
Iterative WSD (*t = 0*)	6137	77.3	65.1	70.7
Wordnet Baseline	6137	56.6	56.6	56.6
Corpus Baseline	6137	79.6	61.4	69.3

Table 5.g: Average 4-fold cross validation results for Marathi Health corpus using features learnt from Hindi Health corpus.

6.2. Observations:

It was observed that better results are obtained when *(t = 0)*, *i.e.*, when the sense distributions learnt from the sense tagged corpus are used for all the words. This shows that domain specific sense distributions play a very important role as they are significantly different from the sense distributions listed in wordnet. An interesting thing to note is that the results are consistent for all the languages tested in both the domains and are significantly better than the baseline. It should be noted that simply selecting the most frequent sense from the corpus performs better than our algorithm. This can be attributed to the fact that our test data is very small (1000-2500 sentences) and hence almost all the words in the test data were seen in the training data (5000-7000 sentences). If the test data is large (as would be the case when the system is deployed) then the most frequent corpus sense will not be available for unknown words. In such cases, our algorithm will still be able to perform disambiguation by relying on the other four terms in the formula (*i.e., BelongingnessToDominantConcept(S_w),CorpusCo occurrence(S, S_w), WNConceptualDistance(S, S_w) and WNSemanticGraphDistance(S, S_w)*).

As mentioned earlier, one of the main objectives of this work was to develop a disambiguation scheme which works even in the absence of sense tagged corpus for some resource poor language (say L_1), provided the corresponding parallel sense tagged corpus is available for another language (say L_2). The case in point was Hindi (L_2) and Marathi (L_1). We used Hindi sense tagged corpora for feature learning of Marathi corpora. The results obtained for Marathi show that our scheme is able to achieve this language independence to a great extent. The results are significantly better when compared to the baseline, but are not as good as those obtained when the engine is trained on Marathi sense tagged corpus.

This brings us to the issue of the trade-off between higher accuracy and efforts needed for collecting sense tagged corpus. Considering that the results are reasonably good for all POS categories across both the domains, we can sacrifice some accuracy in favor of reduced cost of sense tagged corpora.

7. 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 poor language, provided (a) there are aligned and cross linked sense dictionaries for the language in question and another resource rich language, (b) there are parallel corpora for the two languages and (c) the corpora for the other language is sense tagged.

(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 significance of domain specific sense distributions in WSD.

We have tested our algorithm only as a standalone application. We would like to integrate it with an existing Machine Translation System or a Cross-Lingual Information Retrieval System and test its effectiveness in enhancing the performance of these systems.

It would also be interesting to study the effect of the errors existing in wordnets (such as incorrect hypernymy-hyponymy links or missing hypernymy-hyponymy links) on the performance of our algorithm. Due to the iterative nature of the algorithm it is possible that the noisy predictions in the earlier stages could lead to more errors in the subsequent iterations. The effect of all such errors on the performance of the algorithm needs to be studied.

Acknowledgements

This project was partly funded by ***The Department of Information Technology, Government of India***.

We would like to thank Sonal Pathade and Geetanjali Rane for sense marking the Tourism and Health corpora.

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.

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

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

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 Empirical 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.

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.

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.