It Takes Two to Tango: A Bilingual Unsupervised Approach for Estimating Sense Distributions Using Expectation Maximization

Mitesh M. Khapra Salil Joshi Pushpak Bhattacharyya
Department Of Computer Science and Engineering,
IIT Bombay,
Powai,
Mumbai, 400076.
{miteshk,salilj,pb}@cse.iitb.ac.in

Abstract

Several bilingual WSD algorithms which exploit translation correspondences between parallel corpora have been proposed. However, the availability of such parallel corpora itself is a tall task for some of the resource constrained languages of the world. We propose an ***unsupervised*** bilingual EM based algorithm which relies on the counts of translations to estimate sense distributions. ***No parallel or sense annotated corpora are needed***. The algorithm relies on a synset-aligned bilingual dictionary and in-domain corpora from the two languages. A symmetric generalized Expectation Maximization formulation is used wherein the sense distributions of words in one language are estimated based on the raw counts of the words in the aligned synset in the target language. The overall performance of our algorithm when tested on 4 language-domain pairs is better than current state-of-the-art knowledge based and bilingual unsupervised approaches.

1 Introduction

Word Sense Disambiguation (WSD) is one of the central and most widely investigated problems in Natural Language Processing (NLP). A wide variety of approaches ranging from supervised to unsupervised algorithms have been proposed. Of these, supervised approaches (Ng and Lee, 1996; Lee et al., 2004) which rely on sense annotated corpora have proven to be more successful, and they clearly outperform knowledge based and unsupervised approaches (Lesk, 1986; Walker and Amsler, 1986; Agirre and Rigau, 1996; Rada, 2005; Agirre and Soroa, 2009; McCarthy et al., 2004). However, creation of sense annotated corpora has always remained a costly proposition, especially for some of the resource deprived languages.

In this context, *"Disambiguation by Translation"* is a popular paradigm which tries to obviate the need for sense annotated corpora without compromising on accuracy. Such algorithms rely on the frequently made observation that a word in a given source language tends to have different translations in a target language depending on its sense. Given a sentence-and-word-aligned parallel corpus, these different translations in the target language can serve as automatically acquired sense labels for the source word. Although these algorithms (*e.g.*, (Diab and Resnik, 2002; Ng et al., 2003)) give high accuracies, the requirement of a significant amount of bilingual parallel corpora may be an unreasonable demand for many language pairs (perhaps more unreasonable than collecting sense annotated corpora itself).

Recent work by Khapra et al. (2009) has shown that, within a domain, it is possible to leverage the annotation work done for WSD on one language (L_2) for the purpose of another language (L_1), by projecting parameters learned from wordnet and sense annotated corpus of L_2 to L_1. This method does not require a parallel corpus. However, it requires sense marked corpus for one of the two languages. In this work, we focus on scenarios where no sense marked corpus is available in either language. Our method requires only untagged in-domain corpora from the two languages. Given such bilingual in-domain corpora (non-parallel) the counts of different translations appearing in the other language can be used to estimate the sense distributions in one language.

For example, consider the word *facility* which has two senses, *viz.*, "a building used for a particular industry" and "a service (*e.g., gym/internet facility*)". Given a set of documents from the Sports domain, it is intuitive to expect that the sec-

Proceedings of the 5th International Joint Conference on Natural Language Processing, pages 695–704,
Chiang Mai, Thailand, November 8 – 13, 2011.

ond sense would be more prevalent. Similarly, if we are given a corpus of another language (say, Hindi) belonging to the same domain (*i.e.,* Sports) then we would expect to see more words which are manifestations of the second sense than the first sense. Thus, we can estimate the probabilities of different senses of the word *'facility'* by looking at the counts of its translations in different senses. In this case, the count of the translations belonging to the second sense would be more and hence this sense would emerge as the winner sense. However, the catch here is that the translations themselves might be ambiguous and hence simply relying on their counts would lead to errors. Hence, we propose a generalized Expectation Maximization based formulation where the counts get weighted by the sense probabilities estimated in the previous iteration.

The overall performance of our algorithm, when tested in an all-words scenario (as opposed to testing on specific target words) for two languages across two domains, is better than state-of-the-art knowledge based and bilingual unsupervised approaches. Further, when the evaluation is restricted to only those words which have different translations across senses, the overall performance of our algorithm is better than the wordnet first sense baseline for 2 out of the 4 language-domain pairs. Such words account for 82-83% of the total test words. This is appreciable as the wordnet first sense baseline is often *hard-to-beat* for an unsupervised approach even when restricted to specific domains. For example, in the SEMEVAL-2010 task on *"All Words WSD on a specific domain"* (Agirre et al., 2010), no unsupervised system was able to perform better than the wordnet first sense baseline.

The remainder of this paper is organized as follows. In section 2 we present related work. Section 3 describes the Synset aligned multilingual dictionary which lies at the heart of our work. In section 4 we discuss the EM formulation used for estimating sense distributions with the help of a motivating example. Section 5 presents the experimental setup. In section 6 we give the results followed by discussions in section 7. Section 8 concludes the paper.

2 Related Work

Monolingual approaches to Word Sense Disambiguation are abundant ranging from supervised, semi-supervised to unsupervised methods. Supervised approaches such as SVM (Lee et al., 2004) and k-NN (Ng and Lee, 1996) give high accuracies, but the requirement of large annotated corpora renders them unsuitable for resource scarce languages. On the other hand, Knowledge based approaches (Lesk, 1986; Walker and Amsler, 1986; Agirre and Rigau, 1996; Rada, 2005; Agirre and Soroa, 2009) which use wordnet, and unsupervised approaches (McCarthy et al., 2004) which use untagged corpus, are less demanding in terms of resources but fail to deliver good results. This situation underlines the need for high accuracy resource conscious approaches to WSD.

In this context, usupervised Word Sense Induction (WSI) methods (Jean, 2004; Klapaftis and Manandhar, 2008) which induce corpus senses by partitioning the co-occurrence graph of a target word have shown promise. One drawback of these approaches is that they require a large number of untagged instances (typically, collected from the web) for every target word to induce meaningful partitions in the co-occurrence graph. Collecting such target-word specific instances is a difficult proposition (especially in an all-words scenario) for resource constrained languages such as Hindi and Marathi which have very poor web presence. Further, in a bilingual setting where parameters need to be ported from one language to another, it is important to associate labels with the clusters induced from the graph partitions so that these clusters can be aligned across languages. This is a difficult proposition and does not fall under the purview of WSI. Hence, in this work we stick to dictionary defined senses as opposed to corpus induced senses.

Disambiguation by Translation (Gale et al., 1992; Dagan and Itai, 1994; Resnik and Yarowsky, 1999; Ide et al., 2001; Diab and Resnik, 2002; Ng et al., 2003; TufiŞ et al., 2004; Apidianaki, 2008) is another paradigm which attempts at reducing the need for annotated corpora, while ensuring high accuracy. The idea is to use the different target translations of a source word as automatically acquired sense labels. A severe drawback of these algorithms is the requirement of a significant amount of parallel corpora which may be difficult to obtain for many language pairs.

Li and Li (2004) proposed an approach based on bilingual bootstrapping which does not need parallel corpora and relies only on in-domain corpora

from two languages. However, their approach is semi-supervised in contrast to our approach which is unsupervised. Further, they focus on the more specific task of Word Translation Disambiguation (WTD) as opposed to our work which focuses on the broader task of WSD.

Kaji and Morimoto (2002) proposed an unsupervised bilingual approach which aligns statistically significant pairs of related words in language L_1 with their cross-lingual counterparts in language L_2 using a bilingual dictionary. This approach is based on two assumptions (i) words which are most significantly related to a target word provide clues about the sense of the target word and (ii) translations of these related words further reinforce the sense distinctions. The translations of related words thus act as cross-lingual clues for disambiguation. This algorithm when tested on 60 polysemous words (using English as L_1 and Japanese as L_2) delivered high accuracies (coverage=88.5% and precision=77.7%). However, when used in an all-words scenario on our dataset, this algorithm performed poorly (see section 6).

Our work focuses on a bilingual approach for estimating sense distributions and the only resources required for our work are in-domain corpora from two languages and a synset aligned multilingual dictionary which is described in the next section.

3 Synset Aligned Multilingual Dictionary

A novel and effective method of storage and use of dictionary in a multilingual setting was proposed by Mohanty et al. (2008). For the purpose of current discussion, we will refer to this multilingual dictionary framework as *MultiDict*. One important departure in this framework from the 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	{malechild, boy}	{लड़का *(ladkaa)*, बालक *(baalak)*, बच्चा *(bachchaa)*}	{मुलगा *(mulgaa)*, पोरगा *(porgaa)*, पोर *(por)*}

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, *English, Hindi and Marathi*). The pivot language to which other languages link is Hindi. This approach of creating wordnet for a new language by linking to the synsets of a pivot language - more popularly known as the expansion approach - has several advantages over creating a wordnet from scratch as discussed in Mohanty et al. (2008).

Note that every word in the Marathi synset is considered to be a translation of the corresponding words in the Hindi synset. Thus, the Marathi words *mulgaa*, *porgaa* and *por* are translations of the Hindi word *ladakaa* and so on. These synset-specific translations play a very important role in our work as explained in the next section.

4 Bilingual EM for estimating sense distributions

We first explain the intuition behind our approach and then derive the E and M steps of our algorithm with the help of an example.

4.1 Intuition

Our work relies on the key observation of Khapra et al. (2009) that within a domain, the co-occurrence counts of *(word, sense)* in one language can be used to estimate the sense distributions of their translations in another language. For example, consider two languages, say $L_1 = Hindi$ and $L_2 = Marathi$. Now, for a given word u in L_2 if a particular sense (say S_1) is more prevalent in a domain then a target language (L_1) corpus from the same domain will contain more words which are translations of sense S_1 as compared to words which are translations of other senses of this word. For example, the Marathi word *maan*, when used in the sense of *"body part (neck)"* gets translated in Hindi as *gardan* or *galaa* whereas when it is used in the sense of *"prestige"*, it gets translated as *aadar* or *izzat*. Now consider that corpora for the two languages are available from the Health domain. Since, in the Health domain, the *"body part (neck)"* sense is more prevalent we can expect the words *gardan* or *galaa* to be more prevalent in a Hindi Health corpus as compared to *aadar* or *izzat*. The probability of the dif-

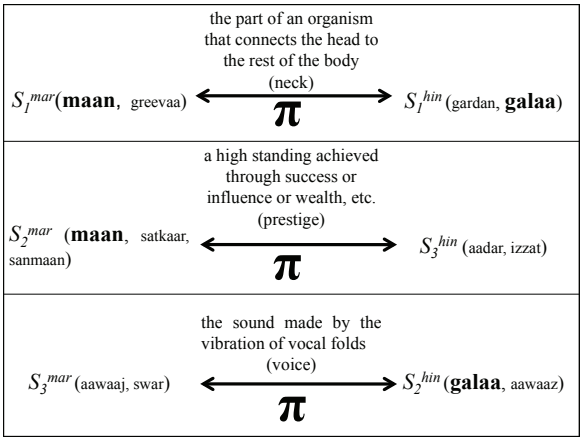

Figure 1: Alignment between different synsets of *maan* and *galaa*

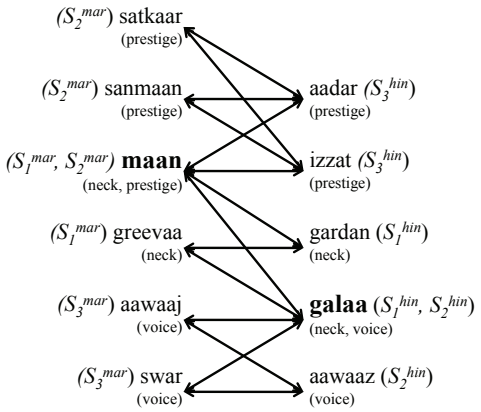

Figure 2: A bipartite graph of translation correspondences

ferent senses of *maan* can thus be estimated based on the counts of {*gardan, galaa*} and {*aadar, izzat*}. However, since the words {*gardan, galaa*} and {*aadar, izzat*} may themselves be ambiguous, their raw counts cannot be used directly for estimating the sense distributions of *maan*. Instead, these counts are refined iteratively using an EM algorithm as explained in the next subsection.

4.2 Derivation with illustration

With the basic intuition provided above, we can now start deriving the EM formulation for estimating sense distributions. For ease of understanding we present the derivation with the help of an illustration. We use the following notations,

- L_1 = first language (say, Hindi)
- L_2 = second language (say, Marathi)
- $synsets_L(word) = \{S^L | word \in S^L\}$ where, S^L denotes a synset in language L
- $words(S^L) = \{word | word \in S^L\}$
- $\pi_{L_2}(S^{L_1})$ = S^{L_2} $s.t.$ $Sense(S^{L_1}) = Sense(S^{L_2})$ *i.e.,* S^{L_1} & S^{L_2} represent the same concept in L_1 and L_2 respectively. The synsets S^{L_1} and $\pi_{L_2}(S^{L_1})$ will thus be aligned in the *MultiDict*.
- $translations_{L_2}(word, S^{L_1})$ = $words(\pi_{L_2}(S^{L_1}))$.The function $translations$ thus gives the translations of a $word \in S^{L_1}$ in the corresponding projected synset in L_2.

Now, consider the word $maan \in S_1^{mar}$ and the word $galaa \in S_1^{hin}$ where $\pi_{hin}(S_1^{mar})$ = S_1^{hin} and vice versa. Further, $galaa \in translations_{hin}(maan, S_1^{mar})$ and $maan \in translations_{mar}(galaa, S_1^{hin})$. The different synsets to which these words belong and the corresponding aligned synsets in the other language are shown in Figure 1. The complete set of translations of these words are shown in Figure 2. We are now interested in estimating $P(S_1^{mar}|maan)$ and $P(S_1^{hin}|galaa)$. Figures 1 and 2 should be referred to while reading the derivation below.

Using the basic definition of probability, we have,

$$P(S_1^{mar}|maan) = \frac{\#(S_1^{mar}, maan)}{\#(S_1^{mar}, maan) + \#(S_2^{mar}, maan)}$$

where,

$$\#(S_i^{mar}, maan) = \textit{no. of times maan appears with sense } S_i^{mar}$$

Following the approach of Khapra et al. (2009) we replace the counts of $(S_i^{mar}, maan)(i \in \{1,2\})$ by the collective counts of the translations in the aligned synsets. The rationale behind the above substitution is that if $v \in L_2$ is a translation of $u \in L_1$ in sense S then the co-occurrence count of (v, S) gives a good approximation for the co-occurrence count of (u, S). Thus,

$$P(S_1^{mar}|maan)$$
$$\approx \frac{\#(S_1^{hin}, gardan) + \#(S_1^{hin}, galaa)}{\#(S_1^{hin}, gardan) + \#(S_1^{hin}, galaa) + \#(S_3^{hin}, aadar) + \#(S_3^{hin}, izzat)}$$

where,

$S_1^{hin} = \pi_{hin}(S_1^{mar})$ $(see\ Figure\ 1)$

$S_3^{hin} = \pi_{hin}(S_2^{mar})$ $(see\ Figure\ 1)$

$(gardan, galaa) \in translations_{hin}(maan, S_1^{mar})$ $(see\ Figure\ 2)$

$(aadar, izzat) \in translations_{hin}(maan, S_2^{mar})$ $(see\ Figure\ 2)$

If we had a sense annotated corpus in Hindi then we could have easily estimated the above probability as shown by Khapra et al. (2009). We propose that even in the absence of such annotated corpus we can still estimate the sense distributions using the expected value of the terms in the above equation as shown below,

E-step

$$P(S_1^{mar}|maan) \approx \frac{P(S_1^{hin}|gardan)\cdot\#(gardan) + P(S_1^{hin}|galaa)\cdot\#(galaa)}{Z}$$

$$\begin{aligned}\text{where, } Z = {} & P(S_1^{hin}|gardan)\cdot\#(gardan) \\ & + P(S_1^{hin}|galaa)\cdot\#(galaa) \\ & + P(S_3^{hin}|aadar)\cdot\#(aadar) \\ & + P(S_3^{hin}|izzat)\cdot\#(izzat)\end{aligned}$$

The above equation takes care of the fact that the different translations of $maan$ would themselves be ambiguous and hence their raw counts (*e.g*, $\#(gardan)$, $\#(galaa)$, *etc.*) cannot be used directly for estimations. Instead, these counts are weighted with the appropriate probability to calculate the expected count (*e.g.*, $E[\#(S_1^{hin}, galaa)] = P(S_1^{hin}|galaa)\cdot\#(galaa)$). The parameters $P(S_1^{hin}|galaa)$, $P(S_1^{hin}|gardan)$, $P(S_3^{hin}|aadar)$ and $P(S_3^{hin}|izzat)$ above are unknown and can in turn be estimated using the counts of the corresponding translations of these words (see Figure 2) as shown below:

M-step

$$P(S_1^{hin}|galaa) \approx \frac{P(S_1^{mar}|maan)\cdot\#(maan) + P(S_1^{mar}|greeva)\cdot\#(greeva)}{Z}$$

$$\begin{aligned}Z = {} & P(S_1^{mar}|maan)\cdot\#(maan) \\ & + P(S_1^{mar}|greeva)\cdot\#(greeva) \\ & + P(S_3^{mar}|aawaaj)\cdot\#(aawaaj) \\ & + P(S_3^{mar}|swar)\cdot\#(swar) \\ & \text{where,} \\ & S_1^{mar} = \pi_{hin}(S_1^{hin})\ (see\ Figure\ 1) \\ & S_3^{mar} = \pi_{mar}(S_2^{hin})\ (see\ Figure\ 1) \\ & (maan, greeva) \in translations_{mar}(galaa, S_1^{hin})\ (see\ Figure\ 2) \\ & (aawaaj, swar) \in translations_{mar}(galaa, S_2^{hin})\ (see\ Figure\ 2)\end{aligned}$$

Similarly, the other parameters (*i.e.*, $P(S_1^{hin}|gardan)$, $P(S_3^{hin}|aadar)$ and $P(S_3^{hin}|izzat)$) can be estimated. The overall process of estimating sense distributions in the two languages can thus be considered to be a back-and-forth traversal over translation correspondences as shown in Figure 2. The two languages thus mutually help each other in estimating sense distributions. In general, for a word $u \in L_1$ and a word $v \in L_2$ the E and M steps can be written as shown below.

E-Step:

$$P(S_k^{L_1}|u) \approx \frac{\sum_v P(\pi_{L_2}(S_k^{L_1})|v)\cdot\#(v)}{\sum_{S_i^{L_1}}\sum_y P(\pi_{L_2}(S_i^{L_1})|y)\cdot\#(y)}$$

$$\begin{aligned}\text{where, } & S_k^{L_1}, S_i^{L_1} \in synsets_{L_1}(u) \\ & v \in translations_{L_2}(u, S_k^{L_1}) \\ & y \in translations_{L_2}(u, S_i^{L_1})\end{aligned}$$

M-Step:

$$P(S_j^{L_2}|v) \approx \frac{\sum_a P(\pi_{L_1}(S_j^{L_2})|a)\cdot\#(a)}{\sum_{S_i^{L_2}}\sum_b P(\pi_{L_1}(S_i^{L_2})|b)\cdot\#(b)}$$

$$\begin{aligned}\text{where, } & S_j^{L_2}, S_i^{L_2} \in synsets_{L_2}(v) \\ & a \in translations_{L_1}(v, S_j^{L_2}) \\ & b \in translations_{L_1}(v, S_i^{L_2})\end{aligned}$$

Note that the E and M steps are symmetrical except for the change in languages. Either of them could be the E-step, making the other as the M-step. Once the sense distributions have been estimated using the above EM algorithm, each word in the test corpus is disambiguated by assigning it the most frequent sense as learned from the sense distributions.

4.3 Problematic cases in estimating sense distributions using EM (non-progressiveness estimation)

Some words have the same translations in the target language across senses. For example, the word *samudra* in Marathi has two senses, *viz.*, S_1 = *a large water body* and S_2 = *a limitless quantity*, which is a metaphorical sense (*e.g.*, a *sea* of opportunities). The corresponding Hindi synsets contain the same word, *viz.*, *saagar*. In other words, *samudra* in Marathi gets translated as *saagar* in Hindi irrespective of its sense. Further, the back-translation of *saagar* in Marathi is *samudra* in both the senses. These words thus form a closed loop of translations. In such cases the algorithm

Category	Polysemous words		Monosemous words	
	Tourism	Health	Tourism	Health
Noun	62336	24089	35811	18923
Verb	6386	1401	3667	5109
Adjective	18949	8773	28998	12138
Adverb	4860	2527	13699	7152
All	92531	36790	82175	43322

Table 2: Polysemous and Monosemous words per category in each domain for Hindi

Category	Polysemous words		Monosemous words	
	Tourism	Health	Tourism	Health
Noun	45589	17482	27386	11383
Verb	7879	3120	2672	1500
Adjective	13107	4788	16725	6032
Adverb	4036	1727	5023	1874
All	70611	27117	51806	20789

Table 3: Polysemous and Monosemous words per category in each domain for Marathi

Category	Avg. degree of wordnet polysemy for polysemous words	
	Tourism	Health
Noun	3.02	3.17
Verb	5.05	6.58
Adjective	2.66	2.75
Adverb	2.52	2.57
All	3.09	3.23

Table 4: Average degree of wordnet polysemy per category in the 2 domains for Hindi

Category	Avg. degree of wordnet polysemy for polysemous words	
	Tourism	Health
Noun	3.06	3.18
Verb	4.96	5.18
Adjective	2.60	2.72
Adverb	2.44	2.45
All	3.14	3.29

Table 5: Average degree of wordnet polysemy per category in the 2 domains for Marathi

will not progress and get stuck with the initial values. It will thus fail to produce better estimates in successive iterations.

Further, for some language-specific words appearing in L_1 (or L_2), no projected synsets were available in L_2 (or L_1 respectively). As evident from the E and M steps, in the absence of such synsets, the algorithm will assign zero probabilities to all the senses of such words.

5 Experimental Setup

We used the publicly available dataset[1] described in Khapra et al. (2010) for all our experiments. The data was collected from two domains, *viz.*, Tourism and Health. The data for Tourism domain was collected by manually translating English documents downloaded from Indian Tourism websites into Hindi and Marathi. Similarly, English documents for Health domain were obtained from two doctors and were manually translated into Hindi and Marathi. The entire data was then manually annotated by three lexicographers adept in Hindi and Marathi. To calculate the inter-tagger agreement (ITA), we we got a small portion (around 5%) of the corpus annotated by two annotators[2]. The ITA on this small corpus was found to be around 85%.

Since ours is an unsupervised algorithm, we refer to the manually assigned sense labels only for evaluation and do not use them during training. The various statistics pertaining to the total number of words, number of words per POS category and average degree of polysemy are described in Tables 2 to 5. Although Tables 2 and 3 also report the number of monosemous words, we would like to clearly state that we do not include monosemous words while evaluating the performance of our algorithms (such words do not need any disambiguation).

We did a 2-fold cross validation of our algorithm using this corpus. The unsupervised parameter estimation was done using 1 fold and testing was done on the remaining fold. Each word in the test corpus is disambiguated by assigning it the most frequent sense as learned from the estimated sense distributions. Note that even though the corpora were parallel we have not used this property in any way in our experiments or EM formulation. In fact, the documents in the two languages were arbitrarily split into 2 folds so that the parallel documents do not fall in the same folds for the two languages. Further, we observed that whether the documents are split arbitrarily (such that parallel documents do not lie in the same fold) or carefully (such that parallel documents lie in the same fold) the overall F-scores remain comparable (within $\pm$ 0.5%). Also note that there was sufficient variety in our corpus as the Tourism documents were related to places from all over India. Similarly, the Health documents were related to a wide range of diseases from common cold to cancer.

6 Results

We report the results using following algorithms:

a. Wordnet first sense ***(WFS)***: The F-score obtained by selecting the first sense of every word. This is a typically reported baseline for supervised approaches as the WFS of a word in

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

[2] It is very expensive to get the entire corpus tagged by 2 annotators. Hence, we calculated the ITA based on the agreement between two lexicographers on a small portion of the corpus

Algorithm	Average				
	N	R	A	V	O
WFS	60.00	68.64	52.39	39.65	57.29
EM	53.35	56.95	51.39	29.98	51.26
PPR	56.17	0.00	38.94	29.74	48.88
RB	34.74	44.32	39.38	17.21	34.79
MI	10.97	3.89	10.07	5.63	9.97

Table 6: Average 2-fold cross validation results averaged over all Language-Domain pairs for all words

Algorithm	Average				
	N	R	A	V	O
WFS	60.86	65.00	52.64	42.00	57.70
EM	57.78	61.28	54.16	31.87	54.98
PPR	58.03	0.00	40.91	30.58	50.42
RB	34.17	43.37	39.21	15.64	34.13
MI	9.62	4.69	8.96	4.17	8.78

Table 7: Average 2-fold cross validation results averaged over all Language-Domain pairs for words which do not face the problem of non-progressiveness estimation

Algo-	HINDI-HEALTH					MARATHI-HEALTH					HINDI-TOURISM					MARATHI-TOURISM				
rithm	N	R	A	V	O	N	R	A	V	O	N	R	A	V	O	N	R	A	V	O
WFS	52.12	73.59	50.79	22.06	52.12	58.52	68.00	44.29	47.91	55.43	64.22	75.66	51.13	33.30	59.99	58.97	57.36	58.26	44.65	57.16
EM	50.87	54.30	55.05	5.87	50.43	56.78	54.96	50.33	41.93	53.81	54.02	57.88	49.88	20.09	51.07	52.44	58.35	51.52	37.57	50.95
PPR	44.82	0.00	40.56	20.66	41.22	54.88	0.00	38.04	39.94	48.32	58.44	0.00	36.44	24.06	50.11	59.55	0.00	41.8	31.92	51.49
RB	34.31	45.01	40.72	9.10	35.65	37.33	44.34	38.42	20.68	36.05	34.20	44.62	39.37	12.62	34.33	34.71	43.51	38.86	20.99	34.46
MI	12.73	6.5	11.13	5.65	11.69	9.78	4.65	10.16	5.16	9.01	11.07	2.11	9.41	3.27	9.77	10.37	4.09	10.28	7.73	9.72

Table 8: Average 2-fold cross validation results for each Language-Domain pair for all words

Algo-	HINDI-HEALTH					MARATHI-HEALTH					HINDI-TOURISM					MARATHI-TOURISM				
rithm	N	R	A	V	O	N	R	A	V	O	N	R	A	V	O	N	R	A	V	O
WFS	54.25	69.17	50.77	21.21	52.95	61.50	67.86	44.27	52.30	57.68	64.50	69.45	50.14	34.61	59.46	58.98	57.73	61.02	47.34	57.85
EM	56.39	56.54	57.35	4.70	54.64	62.58	58.99	53.78	45.24	58.72	57.47	65.22	51.09	20.70	53.87	57.09	61.19	56.78	40.01	55.20
PPR	47.07	0.00	40.50	18.80	42.68	57.98	0.00	39.79	44.11	51.27	59.65	0.00	37.88	23.10	51.12	61.54	0.00	46.30	33.29	53.19
RB	34.14	46.43	40.76	6.54	35.54	35.86	45.38	38.65	18.35	34.96	33.99	40.01	38.57	11.31	33.67	33.79	43.92	39.30	19.45	33.71
MI	11.96	8.23	10.20	3.73	10.98	9.17	5.42	9.69	2.97	8.38	9.08	2.66	7.91	2.52	8.13	9.32	4.33	9.43	5.93	8.64

Table 9: Average 2-fold cross validation results for each Language-Domain pairs for words which do not face the problem of non-progressiveness estimation

Hindi and Marathi wordnets is determined manually by a lexicographer based on his/her native speaker intuition.

b. Random Baseline ***(RB)***: The F-score obtained by selecting a random sense of every word. This is a typically reported baseline for unsupervised approaches.

c. Bilingual Expectation Maximization ***(EM)***: The F-score obtained by using our approach.

d. Personalized PageRank ***(PPR)***: The F-score obtained by using a state-of-the-art knowledge based approach (Agirre and Soroa, 2009).

e. Mutual Information ***(MI)***: The F-score obtained by using the bilingual unsupervised approach of Kaji and Morimoto (2002) which uses cross-lingual clues based on in-domain corpora and aligned synsets.

6.1 A note on other state-of-the-art approaches

The unsupervised algorithm by McCarthy et al. (2004) which uses in-domain corpora to estimate predominant senses would have been more appropriate for comparison with our approach as it is a corpus based approach as opposed to PPR which is a wordnet based approach. However, this approach requires a dependency parser to extract syntactic relations to construct a feature vector for identifying the nearest neighbors of a target word. Unfortunately, such parsers are not available for Hindi and Marathi and hence we could not compare our algorithm with this approach. Further, there are other unsupervised approaches (see section 2) which use corpus induced senses and/or parallel corpora. However, our work focuses on dictionary defines senses and does not need parallel corpora. Hence we did not find it appropriate to present a comparison with these algorithms.

6.2 Non-progressiveness estimation

We observed that around 17-18% of the total words in the corpus face the problem discussed in section 4.3. Hence we report 2 sets of results:

(i) only for those words which do not face the problem of non-progressiveness estimation.

(ii) for all words.

The first set of results thus covers 82-83% of the words in the corpus depending on the language and the domain.

All the results are summarized in Tables 6 to 9. Table 6 gives the overall average F-score for

all-words over all language domain pairs. Similarly, Table 7 gives the overall average F-score for only those words which do not face the problem of non-progressiveness estimation. Tables 8 and 9 give the average F-score for each language-domain pair for all words and for words which do not face the problem of non-progressiveness estimation respectively. In all tables, we report F-scores for each POS category (N:-nouns, R:-adverbs, A:-adjectives, V:-verbs, O:-all).

7 Discussions

We discuss the important observations made from Tables 6 to 9.

7.1 Performance on all words

The overall performance of our algorithm (see Table 6) is better than state-of-the-art knowledge based approach (PPR) by 3%, bilingual unsupervised approach (MI) by 41% and random baseline (RB) by 17%. These results are consistent across all language-domain pairs except for MARATHI-TOURISM where the performance of PPR is better than our algorithm by 0.5%. On an average the performance of PPR on nouns is better than our algorithm by 3%. However, in 2 out of the 4 language-domain pairs our algorithm does better on nouns than PPR (by 6% in HINDI-HEALTH and 2% in MARATHI-HEALTH - see Table 8). PPR gives an F-score of 0% for adverbs in all language-domain pairs because Hindi and Marathi wordnets do not have any synset relations defined for adverbs.

The performance of all the algorithms is less than the wordnet first sense baseline. As stated earlier, this is a hard baseline for unsupervised approaches (Agirre et al., 2010). Note that the wordnet first sense baseline is more like a supervised approach because the first sense of a word is either determined manually by a lexicographer or by using counts from a mixed domain sense marked corpus. This is a laborious and expensive task which is difficult to do for wordnets of resource deprived languages.

7.2 Performance on words not facing the problem of non-progressiveness estimation

When the performance is restricted to words which do not face the problem of non-progressiveness estimation our approach still does better than PPR, MI and random baseline (see Table 7). Here, the results are consistent across all language-domain pairs (see Table 9). In addition, for two language-domain pairs (*viz.*, MARATHI-HEALTH and HINDI-HEALTH) our algorithm does better than the wordnet first sense baseline. Even though the overall improvement over WFS is small (1-2%) it is still appreciable for an unsupervised approach. Note that none of the other approaches (PPR, MI) are able to perform better than WFS in any language-domain pair.

7.3 Poor performance on verbs

Amongst all the POS categories, the performance of our algorithm is lowest for verbs. We observed that there are two main reasons for this. Firstly, the polysemy of verbs is much higher than that of other POS categories (see Tables 4 & 5). This is a commonly observed problem for all algorithms. Secondly, we observed that many verbs have very fine senses because of which they tend to have overlapping sets of translations across senses. Even though they do not form a closed loop of translations they share many translations across senses. For example, the Hindi word *karna* has the same Marathi translation *karne* in 8 out of the 21 senses that it appears in. Due to these shared translations, the approach of "disambiguation by translation" does not have much scope in the case of such verbs.

8 Conclusions

We presented an unsupervised bilingual approach for estimating sense distributions of words. The algorithm does not require any parallel corpora and uses only in-domain corpora from the two languages. The sense distributions are estimated using a novel bilingual EM formulation by performing a back-and-forth traversal over translation correspondences in the two languages. The algorithm consistently beats the random baseline and state-of-the-art knowledge based and unsupervised approaches. Further, when tested on words which have different translations across senses, the algorithm gives slight improvement over the wordnet first sense baseline in 2 out of the 4 language-domain pairs.

As future work, we would like to test our algorithm on language pairs which belong to distant families so that the number of words having same translations across senses would be less.

References

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

Eneko Agirre and Aitor Soroa. 2009. Personalizing pagerank for word sense disambiguation. In *EACL*, pages 33–41.

Eneko Agirre, Oier López de Lacalle, Christiane Fellbaum, Shu-Kai Hsieh, Maurizio Tesconi, Monica Monachini, Piek Vossen, and Roxanne Segers. 2010. Semeval-2010 task 17: All-words word sense disambiguation on a specific domain. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 75–80, Uppsala, Sweden, July. Association for Computational Linguistics.

Marianna Apidianaki. 2008. Translation-oriented word sense induction based on parallel corpora. In *LREC*.

Ido Dagan and Alon Itai. 1994. Word sense disambiguation using a second language monolingual corpus. *Comput. Linguist.*, 20:563–596, December.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 255–262, Morristown, NJ, USA. Association for Computational Linguistics.

William Gale, Kenneth Church, and David Yarowsky. 1992. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*, 26:415–439. 10.1007/BF00136984.

Nancy Ide, Tomaž Erjavec, and Dan TufiŞ. 2001. Automatic sense tagging using parallel corpora. In *In Proceedings of the 6 th Natural Language Processing Pacific Rim Symposium*, pages 212–219.

Véronis Jean. 2004. Hyperlex: Lexical cartography for information retrieval. In *Computer Speech and Language*, pages 18(3):223–252.

Hiroyuki Kaji and Yasutsugu Morimoto. 2002. Unsupervised word sense disambiguation using bilingual comparable corpora. In *Proceedings of the 19th international conference on Computational linguistics - Volume 1*, COLING '02, pages 1–7, Stroudsburg, PA, USA. Association for Computational Linguistics.

Mitesh M. Khapra, Sapan Shah, Piyush Kedia, and Pushpak Bhattacharyya. 2009. Projecting parameters for multilingual word sense disambiguation. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 459–467, Singapore, August. Association for Computational Linguistics.

Mitesh Khapra, Saurabh Sohoney, Anup Kulkarni, and Pushpak Bhattacharyya. 2010. Value for money: Balancing annotation effort, lexicon building and accuracy for multilingual wsd. In *Proceedings of the 23rd International Conference on Computational Linguistics*.

Ioannis P. Klapaftis and Suresh Manandhar. 2008. Word sense induction using graphs of collocations. In *Proceeding of the 2008 conference on ECAI 2008: 18th European Conference on Artificial Intelligence*, pages 298–302, Amsterdam, The Netherlands, The Netherlands. IOS Press.

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

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

Hang Li and Cong Li. 2004. Word translation disambiguation using bilingual bootstrapping. *Comput. Linguist.*, 30:1–22, March.

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.

Rajat Mohanty, Pushpak Bhattacharyya, Prabhakar Pande, Shraddha Kalele, Mitesh Khapra, and Aditya Sharma. 2008. Synset based multilingual dictionary: Insights, applications and challenges. In *Global Wordnet Conference*.

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.

Hwee Tou Ng, Bin Wang, and Yee Seng Chan. 2003. Exploiting parallel texts for word sense disambiguation: an empirical study. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*, ACL '03, pages 455–462, Morristown, NJ, USA. Association for Computational Linguistics.

Mihalcea Rada. 2005. Large vocabulary unsupervised word sense disambiguation with graph-based algorithms for sequence data labeling. In *In Proceedings of the Joint Human Language Technology and Empirical Methods in Natural Language Processing Conference (HLT/EMNLP)*, pages 411–418.

Philip Resnik and David Yarowsky. 1999. Distinguishing systems and distinguishing senses: new evaluation methods for word sense disambiguation. *Nat. Lang. Eng.*, 5:113–133, June.

Dan TufiŞ, Radu Ion, and Nancy Ide. 2004. Fine-grained word sense disambiguation based on parallel corpora, word alignment, word clustering and aligned wordnets. In *Proceedings of the 20th international conference on Computational Linguistics*, COLING '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

D. Walker and R. Amsler. 1986. The use of machine readable dictionaries in sublanguage analysis. In *In Analyzing Language in Restricted Domains, Grishman and Kittredge (eds), LEA Press*, pages 69–83.