Word Sense Disambiguation Using IndoWordNet

Sudha Bhingardive
Department of Computer Science and Engineering,
Indian Institute of Technology-Bombay,
Powai, Mumbai, India.
Email: sudha@cse.iitb.ac.in

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

Abstract

Word Sense Disambiguation (WSD) is considered as one of the toughest problem in the field of Natural Language Processing. IndoWordNet is a linked structure of WordNets of major Indian languages. Recently, several IndoWordNet based WSD approaches have been proposed and implemented for Indian languages. In this chapter, we present the usage of various other features of IndoWordNet in performing WSD. Here, we have used features like linked WordNets, semantic and lexical relations, *etc*. We have followed two unsupervised approaches, *viz*., (1) use of IndoWordNet in bilingual WSD for finding the sense distribution with the help of Expectation Maximization algorithm, (2) use of IndoWordNet in WSD for finding the most frequent sense using word and sense embeddings. Both these approaches justifies the importance of IndoWordNet for word sense disambiguation for Indian languages, as the results are found to be promising and can beat the baselines.

Keywords: IndoWordNet, WordNet, Word Sense Disambiguation, WSD, Bilingual WSD, Unsupervised WSD, Most Frequent Sense, MFS

1. Introduction

1.1. What is Word Sense Disambiguation?

Word Sense Disambiguation or WSD is the task of identifying the correct meaning of a word in a given context. The necessary condition for a word to be disambiguated is that it should have multiple senses. Generally, in order to disambiguate a given word, we should have a context in which the word has been used and knowledge about the word, otherwise it becomes difficult to get the exact meaning of a word. Also, if the concept of a sense is not well defined, then it becomes very elusive task for WSD. The senses of a word differ from dictionary to dictionary. Some of them are coarse, while other provides a fine-grained distinction between possible senses. This may be the reason why there does not exist any WSD classifier which can give an accuracy of 100%, not even human experts can agree on the sense of some words during manual disambiguation tasks.

Following is the example in Hindi which explains the WSD.

S_1: राम बगीचे के पौधों को काटता है।
(*Ram bagiiche ke paudhon ko kaatataa hai*)
(Ram cuts plants in the garden)

S_2: कुत्ते ने महिला को काटा।
(*kutte ne mahilaa ko kaataa*)
(dog bites a woman)

In sentences S_1 and S_2, the word काटना (*kaatanaa*) has two different senses. In S_1, the correct sense of काटना (*kaatanaa*) is 'to cut', as it appears with context words बगीचा (*baagiichaa*, garden) and पौधा (*paudhaa*, plant). However, in S_2, the correct sense of काटना (*kaatanaa*) is 'to bite', as it appears with the context word कुत्ता (*kutta*, dog).

1.2. Variants of Word Sense Disambiguation

The Word Sense Disambiguation is broadly categorized into two types:

- **Target Word WSD**:
  The target WSD system disambiguates a restricted set of target words, usually one per sentence. Here, supervised approaches are generally used for this purpose where a tagged corpus is used to train the model. This trained model is then used to disambiguate the words in the target document.

- **All Word WSD**:
  The all word WSD system disambiguates all open-class words in the target document. Here, knowledge based or unsupervised methods are usually used for this setting. This is because; the supervised approach faces the problem of data sparseness. In supervised settings, it is not always possible to have a large tagged corpus for training in order to improve the coverage. Hence, unsupervised methods are preferred in the case of all word WSD.

In this chapter, we will first describe various WSD approaches and how IndoWordNet is helpful for WSD in Indian languages. Then, the glimpse of existing WSD approaches which uses IndoWordNet is given. Further, we will discuss our unsupervised approaches for WSD. These approaches make use of IndoWordNet (a) for context based bilingual WSD and (b) for detecting the most frequent sense of a word.

2. Approaches for Word Sense Disambiguation

Over the years, many WSD approaches have been proposed. These are often classified according to the main source of knowledge used in sense differentiation. (a) Approaches that make use of annotated corpora for the training purpose or as seed data in a bootstrapping process, are termed as *supervised* and *semi-supervised*, (b) Approaches that rely completely on external information and are usually performed directly on raw corpora, are termed as *unsupervised*, and (c) Approaches that rely primarily on dictionaries, thesauri, and lexical

knowledge bases, without using any corpus evidence, are termed as *dictionary-based* or *knowledge-based*.

2.1. Supervised WSD Approaches

Supervised methods (Lee et al., 2004; Ng and Lee, 1996; Agirre et al., 2009; Giuliano et al., 2009) formulate WSD as a classification problem: the senses of a word represent classes, and a classifier assigns a class to each new instance of a word. Accordingly, almost any classifier from the machine learning literature can be applied. In addition to a dictionary, these algorithms need at least one annotated corpus, where each appearance of a word is tagged with the correct sense.

2.2. Unsupervised WSD Approaches

Creating annotated corpus for all language-domain pairs is impracticable looking at the amount of time and money required. Hence, unsupervised WSD approaches attracts most of the researchers (Dagan et al., 1991; Schütze, 1998; Diab and Resnik, 2002; Kaji and Morimoto, 2002; Specia et al., 2005; Lefever and Hoste, 2010; Khapra et al., 2011). Unsupervised methods have the potential to overcome the new knowledge acquisition bottleneck and have achieved good results. These methods are able to induce word senses from training text by clustering word occurrences, and then classifying new occurrences into the induced clusters/senses.

2.3. Knowledge Based WSD Approaches

WSD heavily depends on knowledge. This knowledge must be in the machine readable format. There are various structures designed for this purpose, they are known as lexical resources. Lexical resources are of diverse types. For example, tagged and untagged corpora, machine-readable dictionaries, thesauri, glossaries, ontologies, *etc*. The main use of lexical resources in WSD is to associate senses with words. Here, selectional restrictions, overlap of definition text, and semantic similarity measures are used for knowledge based WSD (Lesk, 1986; Mihalcea, 2006; Banerjee et al., 2006; Agirre et al., 2009; Jimeno-Yepes et al., 2010).

3. IndoWordNet for Word Sense Disambiguation

IndoWordNet[1] (Bhattacharyya, 2010) is a linked lexical knowledge base of WordNets of major Indian languages. It consists of synsets, semantic and lexical relations, ontological details, *etc*. It is mainly developed for the purpose of Word Sense Disambiguation in Indian languages. However, it can be used for various other Natural Language Processing (NLP) applications like Machine Translation, Information Retrieval, Sentiment Analysis, Text Entailment, *etc*.

3.1. IndoWordNet as Sense Repository for WSD

IndoWordNet is mainly used as a sense repository for Indian languages. Here, for each word, senses are provided according to its Part-of-Speech (POS) categories, *viz.*, nouns, verbs, adjectives and adverbs. The senses of words are chosen from this sense repository for creating the gold standard sense-annotated corpus. A sense-annotated corpus is created by human

[1] www.cfilt.iitb.ac.in/indowordnet/

experts by manually annotating each occurrence of the target word or all content words in a text. This sense-annotated corpus is used for supervised WSD approaches.

3.2. IndoWordNet as Input Features for WSD

Various IndoWordNet features can be used for WSD, they are described below:

- **Semantic Relations:**
  Semantic relations exist between synsets. These relations are very helpful for disambiguating a target word in a given context. Some of these relations are stated below:
  - **Hypernymy and Hyponymy:** This relation captures *is-a-kind-of* relationship between synsets.
    Example: आम (*aam*, mango) is a kind of फल (*fal,* fruit). So, आम (*aam*, mango) is the hyponymy of फल (*fal,* fruit) and फल (*fal,* fruit) is the hypernymy of आम (*aam*, mango).
  - **Meronymy and Holonymy**: This relation expresses *a-part-of* relationship and its inverse.
    Example: पत्ता (*patta,* leaf) is the meronym of पेड़ (*ped,* tree) and पेड़ (*ped* ,tree) is the holonym of पत्ता (*patta,* leaf).
  - **Entailment:** It is a semantic relationship between two verbs. A verb *X* entails a verb *Y*, if the meaning of *Y* follows logically and is strictly included in the meaning of *X*. This relation is unidirectional.
    Example: खर्राटे मारना (*kharrate maaranaa,* snoring) entails सोना (sonaa, sleeping), but सोना (sonaa, sleeping) does not entail खर्राटे मारना (*kharrate maaranaa,* snoring).
  - **Troponymy:** It is a semantic relation between two verbs when one is a specific 'manner' elaboration of another.
    Example: दहाड़ना (*dahaadanaa,* to roar) is the troponym of बोलना (*bolanaa,* to speak).
- **Lexical Relations:** Lexical relations exist between words. These relations are also helpful for disambiguating a target word in a particular context. Some of these relations are stated below.
  - **Synonymy:** This is the relationship between words in a synset. This relation is symmetric, reflexive and transitive.
    Example: In synset {हाथ, हस्त, कर, पाणि} ({*haath, hasth, kara, paaNi*}, hand), words हाथ, हस्त, कर and पाणि are related through synonymy relation.

- **Antonymy:** It is a lexical relation indicating 'opposites'.
  Example: पतला (*patalaa*, thin) is an antonym of मोटा (*motaa*, fat) and vice versa.

- **Linked structure:** IndoWordNet being a linked structure of Indian language WordNets, its cross linkages across WordNets are helpful for bilingual WSD.

3.3. Existing IndoWordNet based Approaches

Earlier, several WSD approaches by Sinha et al., 2006; Khapra et al., 2008; Mishra et al., 2009, Khapra et al., 2011; Singh et al., 2012; Singh et al., 2013; Bhingardive et al., 2013; Jain et al., 2015; Bhingardive et al., 2015 have been proposed which make use of IndoWordNet as a lexical resource.

4. Our IndoWordNet based WSD Approaches

4.1. Unsupervised Context Based Bilingual WSD

Recently, Bhingardive et al., (2013) published a paper on usage of IndoWordNet for unsupervised context based bilingual WSD approach. This uses Expectation Maximization (EM) algorithm for estimating sense distributions. It builds on the framework of Khapra et al., (2011). So, let us first understand the basic EM based approach by Khapra et al., (2011).

4.1.1. Basic EM Based WSD Approach

This approach relies on the key idea that, within a domain, the co-occurrence count of (word, sense) in one language can be used to estimate the sense distribution of their translations in another language. For example the word *maan* in Marathi with sense 'neck' is translated to Hindi as *galaa* or *gardan* and with sense 'respect' as *aadar* or *izzat*. Hence the probability of different senses of *maan* can be estimated by counts of {*galaa, gardan*} and {*aadar, izzat*}. But in Hindi, the word *galaa* has two meanings *viz*., neck and voice. Because the word *galaa* is itself ambiguous, the raw count of *galaa* cannot directly help in estimating the sense distribution of *maan*.

The approach needs in-domain corpora from two languages as opposed to supervised approaches which need annotated corpora. It uses EM algorithm for estimating sense distributions in comparable corpora. Every polysemous word is disambiguated using the raw counts of its translations in different senses. This approach uses a synset aligned multilingual dictionary (Mohanty, 2008) for finding the translations. This dictionary links synsets from different languages with respect to sense. All the synsets with same sense are aligned in the same row against its sense. In this dictionary, synsets are linked, and after that the words inside the synsets are also linked. For example, for the concept of 'boy', Hindi synset {*ladakaa, balak, bachhaa*} and Marathi synset {*mulagaa, poragaa, por*} are linked as seen in figure 1. The Marathi word '*mulagaa'* is linked to the Hindi word '*ladakaa*' which is its exact lexical substitution.

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)*}

Figure 1: Synset aligned multilingual dictionary

Algorithm:

Suppose words *u* in language L_1 and *v* in language L_2 are translations of each other and their senses are required. The EM based formulation is as follows:

E-Step:

$$p(S^{L_1}|u) = \frac{\sum_v p\left(\pi_{L_2}(S^{L_1}|v).\#(v)\right.}{\sum_{S_i^{L^1}} \sum_x p\left(\pi_{L_2}(S_i^{L_1}|x).\#(x)\right.}$$

where, $S_i^{L_1} \in synsets_{L_1}(u)$

$v \in crosslinks_{L_2}(u, S^{L_1})$

$x \in crosslinks_{L_2}\left(u, S_i^{L_1}\right)$

M-Step:

$$(S^{L_2}|v) = \frac{\sum_u p\left(\pi_{L_1}(S^{L_2}|u).\#(u)\right.}{\sum_{S_i^{L_2}} \sum_y p\left(\pi_{L_1}(S_i^{L_2}|y).\#(y)\right.}$$

where, $S_i^{L_2} \in synsets_{L_2}(v)$

$u \in crosslinks_{L_1}(u, S^{L_2})$

$y \in crosslinks_{L_1}\left(u, S_i^{L_2}\right)$

Here,

- '#' indicates the raw count
- $crosslinks_{L_1}(u, S^{L_2})$ is the set of possible translations of the word u from language L_1 to L_2 in the sense S^{L_2}
- $\pi_{L_2}(S^{L_1})$ means the linked synset of the sense S^{L_1} in L_2

E and M steps are symmetric except for the change in language. In both the steps, they estimate sense distribution in one language using raw counts of translations in another language.

But this approach has following limitations:

- **Poor performance on verbs**: This approach gives poor performance on verbs (25%-38%).

- **Same sense throughout the corpus**: Every occurrence of a word is tagged with the single sense found by the algorithm, throughout the corpus.
- **Closed loop of translations**: This formulation does not work for some common words which have the same translations in all senses. For example, the verb '*karna*' in Hindi has two different senses in the corpus *viz.,* '*to do'* (S_1) and '*to make*' (S_2). In both these senses, it gets translated as 'karne' in Marathi. The word 'karne' also back translates to 'karna' in Hindi through both its senses. In this case, the formulation works out as follows:

  The probabilities are initialized uniformly. Hence, $p(S_1|karana) = p(S_2|karana) = 0.5$ Now, in first iteration the sense of 'karne' will be estimated as follows (E-step):

$$p(S_1|karane) = \frac{p(S_1|karana) * \#(karana)}{\#(karana)} = 0.5$$

$$p(S_2|karane) = \frac{p(S_2|karana) * \#(karana)}{\#(karana)} = 0.5$$

  Similarly, in M-step, we will get $p(S_1|karana) = p(S_2|karana) = 0.5$. Eventually, it will end up with initial probabilities and no strong decision can be made.

.

To address these problems we introduced contextual clues in their formulation by using semantic relatedness. Our modified approach overcomes all the mentioned limitations.

4.1.2. Modified Bilingual EM Approach Using WordNet Similarity

We, Bhingardive et al., (2013) introduced context in the basic EM formulation stated earlier and treat context as a bag of words. We assume that each word in the context influences the sense of the target word independently. Hence,

$$p(S|w, C) = \prod_{c_i \in C} p(S|w, c_i)$$

where, *w* is the target word, *S* is one of the candidate synsets of *w*, *C* is the set of words in context (sentence in our case) and C_i is one of the context words.

Suppose we would have sense tagged data, $p(S|w, c)$ could have been computed as:

$$p(S|w, c) = \frac{\#(S, w, c)}{\#(w, c)}$$

But since the sense tagged corpus is not available, we cannot find $\#(S, w, c)$ from the corpus directly. However, we can estimate it using the comparable corpus in other language. Here, we assume that given a word and its context word in language L_1, the sense distribution in L_1 will be same as that in L_2 given the translation of a word and the translation of its context word in L_2. But these translations can be ambiguous, hence we can use Expectation Maximization approach similar to Khapra et al., (2011) as follows:

E-Step:

$$p(S^{L_1}|u,a) = \frac{\sum_{v,b} p\left(\pi_{L_2}(S^{L_1}|v,b\right).\sigma(v,b)}{\sum_{S_i^{L_1}} \sum_{x,b} p\left(\pi_{L_2}(S_i^{L_1}|x,b\right).\sigma(x,b)}$$

where, $S_i^{L_1} \in \text{synsets}_{L_1}(u)$
$a \in \text{context}(u)$
$v \in crosslinks_{L_2}(u, S^{L_1})$
$b \in \text{crosslinks}_{L_2}(a)$
$x \in crosslinks_{L_2}(u, S_i^{L_1})$

$crosslinks_{L_1}(a)$ is the set of all possible translations of the word *a* from L_1 to L_2 in all its senses. $\sigma(v, b)$ is the semantic relatedness between the senses of *v* and senses of *b*. Since, *v* and *b* go over all possible translations of *u* and *a* respectively. $\sigma(v, b)$ has the effect of indirectly capturing the semantic similarity between the senses of *u* and *a*. A symmetric formulation in the M-step below takes the computation back from language L_2 to language L_1. The semantic relatedness comes as an additional weighing factor, capturing context, in the probablistic score.

M-Step:

$$p(S^{L_2}|v,b) = \frac{\sum_{u,a} p\left(\pi_{L_1}(S^{L_2}|u,a\right).\sigma(u,a)}{\sum_{S_i^{L_2}} \sum_{y,a} p\left(\pi_{L_1}(S_i^{L_2}|y,a\right).\sigma(y,a)}$$

where, $S_i^{L_2} \in \text{synsets}_{L_2}(v)$
$b \in \text{context}(v)$
$u \in crosslinks_{L_2}(v, S^{L_2})$
$a \in \text{crosslinks}_{L_1}(b)$
$y \in crosslinks_{L_1}(v, S_i^{L_2})$

$\sigma(u, a)$ is the semantic relatedness between the senses of *u* and senses of *a* and contributes to the score like $\sigma(v, b)$. Note how the computation moves back and forth between L_1 and L_2 considering translations of both target words and their context words.

In the above formulation, we could have considered the term $\#(word, context_{word})$ (i.e., the co-occurrence count of the translations of the word and the context word) instead of $\sigma\,(\text{word}, \text{context}_{\text{word}})$ but it is very unlikely that every translation of a word will co-occur with every translation of its context word considerable number of times. This term may make sense only if we have arbitrarily large comparable corpus in the other language.

The semantic relatedness is computed by taking the inverse of the length of the shortest path among two senses in the WordNet graph Pedersen et al., (2005). All the semantic relations (including cross-part-of-speech links) viz., *hypernymy*, *hyponymy*, *meronymy*, *entailment*, *attribute etc*., are used for computing the semantic relatedness. Sense scores thus obtained are used to disambiguate all words in the corpus. We consider all the content words from the context for disambiguation of a word. The winner sense is the one with the highest probability.

4.1.2.1. Experiments and Results

We used freely available in-domain comparable corpora[2] in Hindi and Marathi languages. These corpora are available for health and tourism domains. The dataset is same as that used in Khapra et al., (2011) in order to compare the performance.

Algorithm	HIN-HEALTH					MAR-HEALTH				
	NOUN	ADV	ADJ	VERB	Overall	NOUN	ADV	ADJ	VERB	Overall
EM-C	59.82	67.80	56.66	**60.38**	59.63	62.90	62.54	53.63	**52.49**	59.77
EM	60.68	67.48	55.54	**25.29**	58.16	63.88	58.88	55.71	**35.60**	58.03
WFS	53.49	73.24	55.16	**38.64**	54.46	59.35	67.32	38.12	**34.91**	52.57
RB	32.52	45.08	35.42	**17.93**	33.31	33.83	38.76	37.68	**18.49**	32.45

Table 1 Comparison (F-Score) of EM-C and EM for Health domain

Algorithm	HIN-TOURISM					MAR-TOURISM				
	NOUN	ADV	ADJ	VERB	Overall	NOUN	ADV	ADJ	VERB	Overall
EM-C	62.78	65.10	54.67	**55.24**	60.70	59.08	63.66	58.02	**55.23**	58.67
EM	61.16	62.31	56.02	**31.85**	57.92	59.66	62.15	58.42	**38.33**	56.90
WFS	63.98	75.94	52.72	**36.29**	60.22	61.95	62.39	48.29	**46.56**	57.47
RB	32.46	42.56	36.35	**18.29**	32.68	33.93	39.30	37.49	**15.99**	32.65

Table 2 Comparison (F-Score) of EM-C and EM for Tourism domain

Table 1 and Table 2 compare the performance of the following two approaches:

- EM-C (EM with Context): Our modified approach
- EM: Basic EM based approach by Khapra et al., (2011)
- WFS: Wordnet First Sense baseline.
- RB: Random baseline.

Results clearly show that EM-C outperforms EM especially in case of verbs in all language-domain pairs. In health domain, verb accuracy is increased by 35% for Hindi and 17% for Marathi, while in tourism domain; it is increased by 23% for Hindi and 17% for Marathi. The overall accuracy is increased by (1.8 - 2.8%) for health domain and (1.5 - 1.7%) for tourism domain. Since there is less number of verbs, the improved accuracy is not directly reflected in the overall performance.

4.1.2.2. Error Analysis and Phenomena Study

Our approach tags most of the instances of a word depending on its context as opposed to basic EM approach. For example, consider the following sentence from the tourism domain:

वह पत्ते खेल रहे थे |

(*vaha patte khel rahe the*)

(They were playing cards/leaves)

Here, the word पत्ते (plural form of पत्ता) has two senses viz., *leaf* and *playing_card*. In tourism domain, the *leaf* sense is more dominant. Hence, basic EM will tag पत्ता with *leaf* sense. But its true sense is *playing_card*. The true sense is captured only if context is considered. Here,

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

the word खेलना (to play) (root form of खेल) endorses the *playing_card* sense of the word खेलना. This phenomenon is captured by our approach through semantic relatedness. But there are certain cases where our algorithm fails. For example, consider the following sentence:

वह पेड़ के निचे पत्ते खेल रहे थे|
(*vaha ped ke niche patte khel rahe the*)
(They were playing cards/leaves below the tree)

Here, two strong context words पेड़ (tree) and खेलना (play) are influencing the sense of the word पत्ता. Semantic relatedness between पेड़ (tree) and पत्ता (leaf) is more than that of खेलना (play) and पत्ता (playing_card). Hence, the *leaf* sense is assigned to पत्ता. This problem occurred because we considered the context as a bag of words. This problem can be solved by considering the semantic structure of the sentence. In this example, the word पत्ता (leaf/playing_card) is the subject of the verb खेलना (to play) while पेड़ (tree) is not even in the same clause with पत्ता (leaf/playing_card). Thus we could consider खेलना (to play) as the stronger clue for its disambiguation.

Our formulation solves the problems of 'inhibited progress due to lack of translation diversity' and 'uniform sense assignment, irrespective of context' that the previous EM based formulation of Khapra et al. suffers from. More importantly our accuracy on verbs is much higher and more than the state of the art, to the best of our knowledge.

Here, we saw how the cross-linked structure of IndoWordNet is helpful for finding the translations in bilingual WSD. Let us see our another approach which uses various semantic relations from IndoWordNet for creating the sense embeddings and then using for finding the most frequent sense of a word.

4.2. Unsupervised MFS Detection for WSD

This approach proposed by Bhingardive et al., (2015), needs only untagged corpora. Here, features from Hindi WordNet (which is a part of IndoWordNet) are used for detecting the Most Frequent Sense **(**MFS) of a word. Using a large amount of untagged corpora, we first train word embeddings. Then sense embeddings are created using various semantic features from the IndoWordNet. We compare word embeddings of a word with sense embeddings to get the most frequent sense. Approach can be easily ported to various domains and across languages.

4.2.1. Most Frequent Sense (MFS)

The MFS baseline is often hard to beat for any WSD system and it is considered as the strongest baseline in WSD (Agirre, 2007). It has been observed that supervised WSD approaches generally outperform the MFS baseline, whereas unsupervised WSD approaches fail to beat this baseline. The MFS baseline can be easily created if we have a large amount of sense annotated corpora. The frequencies of word senses are obtained from the available sense annotated corpora. Creating such a costly resource for all languages is infeasible, looking at the amount of time and money required. Hence, unsupervised approaches have received widespread attention as they do not use any sense annotated corpora.

McCarthy et al. (2007) proposed an unsupervised approach for finding the predominant sense using an automatic thesaurus. They used WordNet similarity for identifying the predominant sense. Their approach outperforms the SemCor baseline for words with SemCor frequency below five. Buitelaar et al. (2001) presented the knowledge based approach for ranking GermaNet synsets on specific domains. Lapata et al. (2004) worked on detecting the predominant sense of verbs where verb senses are taken from the Levin classes. Our approach is similar to that of McCarthy et al. (2007) as we are also learning predominant senses from the untagged text. Our approach is also unsupervised for detecting the most frequent sense for Hindi language.

4.2.2. Word Embeddings

Word Embeddings have recently gained popularity among Natural Language Processing community (Bengio, 2003; Collobert, 2011). They are based on Distributional Hypothesis which works under the assumption that similar words occur in similar contexts (Harris, 1954). Word Embeddings represent each word with a low-dimensional real valued vector with similar words occurring closer in that space.

In our approach, we use the word embedding of a given word and compare it with all its sense embeddings to find the most frequent sense of that word. Sense embeddings are created using the IndoWordNet based features in the light of the extended Lesk algorithm (Banerjee 2003) as described later in this paper.

4.2.3. Training of Word Embeddings

Word embeddings for Hindi have been trained using *word2vec*[3] tool (Mikolov 2013). This tool provides two broad techniques for creating word embeddings: Continuous Bag of Words (CBOW) and Skip-gram model. The CBOW model predicts the current word based on the surrounding context, whereas, the Skip-gram model tries to maximize the probability of a word based on other words in the same sentence (Mikolov 2013). Word embeddings for Hindi have been trained on Bojar's (2014) corpus. This corpus contains 44 million sentences. Here, the Skip-gram model is used for obtaining word embeddings. The dimensions are set as 200 and the window size as 7 (i.e. *w=7*). We used the test of similarity to establish the correctness of these word embeddings. We observed that given a word and its embedding, the list of words ranked by similarity score had at the top of the list those words which were actually similar to the given word.

4.2.4. Sense Embeddings Creation

Sense embeddings are similar to word embeddings which are low dimensional real valued vectors. Sense embeddings are obtained by taking the average of word embeddings of each word in the sense-bag. The sense-bag for each sense of a word is obtained by extracting the context words from the Hindi WordNet (a part of IndoWordNet) such as synset members (S), content words in the gloss (G), content words in the example sentence (E), synset members of the hypernymy-hyponymy synsets (HS), content words in the gloss of the hypernymy-hyponymy synsets (HG) and content words in the example sentence of the hypernymy-hyponymy synsets (HE). We consider word embeddings of all words in the sense-bag as a cluster of points and choose the sense embedding as the centroid of this cluster.

[3] https://code.google.com/p/word2vec/

Consider a word *w* with *k* senses $w_{S_1}, w_{S_2}, \ldots, w_{S_k}$ taken from the Hindi WordNet. Sense embeddings are created using the following formula,

$$vec(w_{S_i}) = \frac{\sum_{x \in SB(w_{S_i})} vec(x)}{N}$$

where, N is the number of words present in the sense-bag $SB(w_{S_i})$ and $SB(w_{S_i})$ is the sense-bag for the sense w_{S_i} which is given as,

$$SB(w_{S_i} = \{ x | x \in Features(w_{S_i})\}$$

$where, Features(w_{S_i})$ includes the Hindi WordNet based features for w_{S_i} which are mentioned earlier in this section.

As we can see in Figure 2, consider the sense-bag created for the senses of a word काटना (*kaatanaa*). Here, the word काटना (*kaatanaa*) has three senses, S_1 : *to bite*, S_2 :*to cut*, and S_3 :*to spend or to pass time.* The corresponding word embeddings of all words in the sense-bag will act as a cluster as shown in the Figure. Here, there are three clusters with centroids C_1, C_2, C_3 which corresponds to the three sense embeddings of the word काटना (*kaatanaa*).

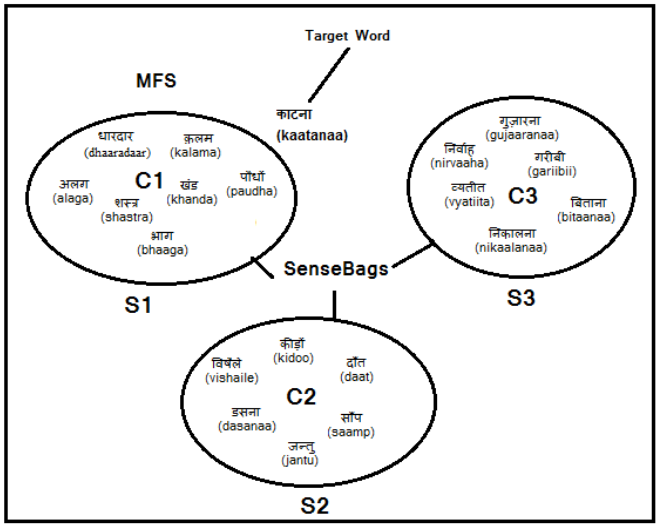

Figure 2 Most Frequent Sense (MFS) detection using Word Embeddings and Sense Embeddings

4.2.5. Most Frequent Sense Identification:

For a given word *w*, we obtain its word embedding and sense embeddings as discussed earlier. We treat the most frequent sense identification problem as finding the closest cluster centroid (i.e. sense embedding) with respect to a given word. We use the cosine similarity as the similarity measure. The most frequent sense is obtained by using the following formulation,

$$MFS_w = argmax_{w_{S_i}} cos(vec(w), vec(w_{S_i}))$$

where, $vec(w)$ is the word embedding for word w, w_{S_i} is the i[th] sense of word w and $vec(w_{S_i})$ is the sense embedding for w_{S_i}. As seen in figure 2, the word embedding of the word काटना (*kaatanaa*) is closer to the centroid C1 as compared to the centroids C_2 and C_3. Therefore, the MFS of the word काटना (*kaatanaa*) is chosen as S1: *to cut*.

4.2.6. Experiments

We performed several experiments to compare the accuracy of UMFS-WE for Hindi WSD. The experiments are restricted to only polysemous nouns. A newspaper sense-tagged dataset of around 80,000 polysemous noun entries was used. This is an in-house data. To compare the performance of UMFS-WE approach, we used the WFS baseline. In the WFS baseline, the first sense in the WordNet is used for WSD. For Hindi, the WFS is manually determined by a lexicographer based on his/her intuition. Results on the newspaper dataset are given in Table 3. The UMFS-WE approach achieves F-1 of 62% for the newspaper dataset.

System	P	R	F-Score
UMFS-WE	62.43	61.58	62.00
WFS	61.73	59.31	60.49

Table 3 Results of Hindi WSD on the newspaper dataset

We have performed several tests using various combinations of WordNet based features for Hindi WSD, as shown in Table 4. We study its impact on the performance of the system for WSD and present a detailed analysis below.

WordNet Features	P	R	F-Score
S	51.73	38.13	43.89
S+G	53.31	52.39	52.85
S+G+E	56.61	55.84	56.22
S+G+E+HS	59.53	58.72	59.12
S+G+E+HG	60.57	59.75	60.16
S+G+E+HE	60.12	59.3	59.71
S+G+E+HS+HG	57.59	56.81	57.19
S+G+E+HS+HE	58.93	58.13	58.52
S+G+E+HG+HE	**62.43**	**61.58**	**62.00**
S+G+E+HS+HG+HE	58.56	57.76	58.16
S+G+HS+HG	0.0	0.0	0.0

Table 4 UMFS-WE accuracy on Hindi WSD with various WordNet features

Our approach, UMFS-WE achieves better performance for Hindi WSD as compared to the WFS baseline. We used various Hindi WordNet based features for comparing results. It is observed that synset members alone are not sufficient for identifying the most frequent sense. This is because some of synsets have a very small number of synset members. Synset members along with gloss members improve results as gloss members are more direct in defining the sense. The other reason is to bring down the impact of topic drift which may have occurred because of polysemous synset members. Similarly, it is observed that adding hypernym/hyponym gloss members gives better performance compared to hypernym/hyponym

synset members. Example sentence members also provide additional information in determining the MFS of a word, which further improves the results. On the whole, we achieve the best performance when S, G, E, HG and HE features are used together. This is shown in Table 4.

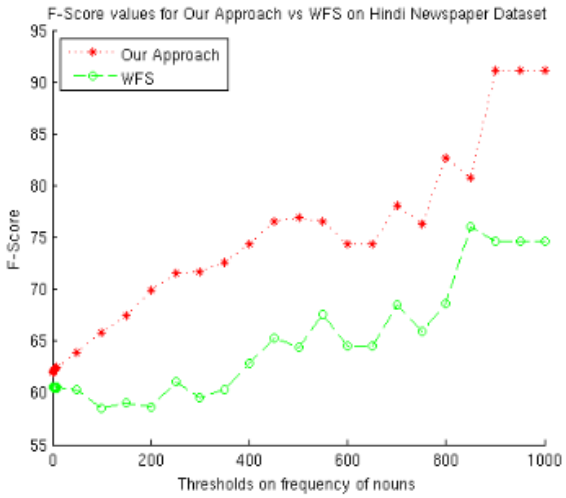

Figure 3 UMFS-WE accuracy on Hindi WSD for words with various frequency thresholds in Newspaper dataset

Also, we have calculated the F-1 score for increasing thresholds on the frequency of nouns appearing in the corpus. This is depicted in figure 3. Here, in the plot, it is clearly shown that, as the frequency of nouns in the corpus increases our approach outperforms baselines.

As opposed to baselines, our approach gives a feasible way to extract predominant senses in an unsupervised setup. Our approach is domain independent so that it can be very easily adapted to a domain specific corpus. To get the domain specific word embeddings, we simply have to run the *word2vec* program on the domain specific corpus. The domain specific word embeddings can be used to get the MFS for the domain of interest. Our approach is language independent. However, due to time and space constraints we have performed our experiments on only Hindi and English languages.

5. Summary

In this chapter, we have highlighted the role of IndoWordNet for performing Word Sense Disambiguation for Indian languages. IndoWordNet is used as a sense repository which consists of unique concepts, its semantic relations, lexical relations between words, *etc.* We have presented two major unsupervised approaches for WSD which use IndoWordNet sense repository.

In the first approach, a context based bilingual WSD is used where two languages help each other in performing WSD. This is done using the linked properties of IndoWordNet. This approach relies on a key idea that, within a domain, a sense distribution and co-occurrence sense distribution remains same across languages. Here, we have used EM based algorithm for finding the sense distribution using the linked WordNets. This approach outperformed the basic bilingual EM based WSD, especially for verbs.

In the second approach, the most frequent sense is detected by exploiting the usage of word embeddings and sense embeddings. The sense embeddings are created using various semantic features of WordNet *viz,* gloss, example sentences, synonyms, hypernyms, *etc.* This approach compares word embeddings with sense embeddings to obtain the most frequent sense. We have tested this approach on Hindi WSD and results are found be very impressive. This proves that the word embeddings capture the most frequent sense of words.

Hence, we can say that the unsupervised approaches are better alternatives as they do not require any sense annotated corpora whose creation needs lots of manual efforts. The two approaches which are described above are found to be very useful invention for the NLP researches and can be used or extended further for their research purpose, in future.

Bibliography

Agirre, E., and de Lacalle, O.L., (2009). *Supervised domain adaption for wsd*. In Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics (pp. 42-50). Association for Computational Linguistics.

Agirre E., and Edmonds, P., (2007). *Word Sense Disambiguation: Algorithms and Applications.*

Banerjee, S. and Pedersen, T., (2003). *Extended gloss overlaps as a measure of semantic relatedness*. In IJCAI, volume 3, pages 805–810.

Bengio, Y., Ducharme, R., Vincent, P., and Janvin, C., (2003). *A Neural Probabilistic Language Model*. J. Mach. Learn. Res., issn = 1532-4435, pp 1137-1155.

Bhattacharyya, P. (2010). *IndoWordNet*. Lexical Resources Engineering Conference. Malta, May.

Bhingardive, S., Bhattacharyya, P., & Shaikh, S., (2013). *Neighbors Help: Bilingual Unsupervised WSD Using Context*. Association for Computational Linguistics, Sofia, Bulgaria.

Bhingardive, S., Singh D., Rudramurty V., Redkar, H.H., Bhattacharyya, P., (2015). *Unsupervised Most Frequent Sense Detection using Word Embeddings*. NAACL, Denver, Colorado.

Bojar, O., Vojtˇech, D., Pavel, R., Pavel, S., V´ıt, S., Aleˇs, T., and Daniel, Z., (2014). *HindEnCorp - Hindi-English and Hindi-only Corpus for Machine Translation*. Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14).

Buitelaar, P., and Sacaleanu, B., (2001). *Ranking and selecting synsets by domain relevance*. Proceedings of WordNet and Other Lexical Resources, NAACL 2001 Workshop.

Chen, X., Liu, Z., and Sun, M., (2014). *A Unified Model for Word Sense Representation and Disambiguation*. Proceedings of ACL 2014.

Collobert, R., Weston, J., Bottou, L., Karlen, M., Kavukcuoglu, K., Kuksa, P., (2011). *Natural language processing (almost) from Scratch.* CoRR, http://arxiv.org/abs/1103.0398.

Dagan, I., Itai, A., & Schwall, U., (1991). *Two Languages Are More Informative Than One*. In Douglas E. Appelt, editor, ACL, pages 130–137. ACL

Diab, M., and Resnik, P., (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.

Giuliano, C., Gliozzo, A.M., and Strapparava, C., (2009). *Kernel methods for minimally supervised wsd*. Computational Linguistics, 35(4), pp.513-528.

Harris, Z., (1968). *Mathematical Structures of Language*. Wiley, New York.

Kaji, H., & Morimoto, Y., (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.

Jain, A., and Lobiyal, D.K., (2015). *Unsupervised Hindi word sense disambiguation based on network agglomeration*. In Computing for Sustainable Global Development (INDIACom), 2015 2nd International Conference on (pp. 195-200). IEEE.

Jimeno-Yepes, A.J. and Aronson, A.R., (2010). *Knowledge-based biomedical word sense disambiguation: comparison of approaches.* BMC bioinformatics, 11(1), p.1.

Khapra, M., Bhattacharyya, P., Chauhan, S., Nair, S., and Sharma, A., (2008). *Domain specific iterative word sense disambiguation in a multilingual setting*. In Proceedings of International Conference on NLP (ICON 2008), Pune, India.

Khapra, M., Joshi, S., and Bhattacharyya, P., (2011). *It takes two to tango: A bilingual unsupervised approach for estimating sense distributions using expectation maximization*. In Proceedings of 5th International Joint Conference on Natural Language Processing, pages 695–704, Chiang Mai, Thailand, November. Asian Federation of Natural Language Processing.

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.

Lapata, M., and Brew C., (2004). *Verb class disambiguation using informative priors*. Computational Linguistics, 30(1):45-75.

Lefever, E., and Hoste. V., (2010). *Semeval-2010 task 3: cross-lingual word sense disambiguation*. In Katrin Erk and Carlo Strapparava, editors, SemEval 2010: 5th International workshop on Semantic Evaluation: proceedings of the workshop, pages 15–20. Association for Computational Linguistics (ACL).

Lesk, M., (1986). *Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone*. Proceedings of the 1986 ACM SIGDOC Conference, Toronto, Canada, 24–26.

McCarthy, D., Koeling, R., Weeds, J., and Carroll J., (2007). *Unsupervised Acquisition of Predominant Word Senses*. Computational Linguistics, 33 (4) pp 553-590.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint arXiv:1301.3781.

Mihalcea, R., (2006). *Knowledge-based methods for WSD*. Word Sense Disambiguation: Algorithms and Applications, pp.107-131.

Mishra, N., Yadav, S., and Siddiqui, T. J., (2009). *An unsupervised approach to Hindi word sense disambiguation*. In Proceedings of the First International Conference on Intelligent Human Computer Interaction (pp. 327-335). Springer India.

Mohanty, R., Bhattacharyya, P., Pande, P., Kalele, Shraddha, Khapra, M., and Sharma, A., (2008). *Synset based multilingual dictionary: Insights, applications and challenges*. In Global Wordnet Conference.

Ng, H.T. & Lee, H.B., (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.

Pedersen, T., Banerjee, S., and Patwardhan, S., (2005). *Maximizing Semantic Relatedness to Perform Word Sense Disambiguation.* Research Report UMSI 2005/25, University of Minnesota Supercomputing Institute, March.

Schütze, H., (1998). *Automatic word sense discrimination.* Computational Linguistics, 24(1): 97–123.

Specia, L., Nunes, M. G., and Stevenson, M., (2005). *Exploiting parallel texts to produce a multilingual sense tagged corpus for word sense disambiguation*. In In Proceedings of RANLP-05, Borovets, pages 525–531.

Singh, S., and Siddiqui, T.J., (2012). *Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation*. In Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on (pp. 1-5). IEEE.

Singh, S., Singh, V.K., and Siddiqui, T.J., (2013). *Hindi Word Sense Disambiguation Using Semantic Relatedness Measure*. In Multi-disciplinary Trends in Artificial Intelligence (pp. 247-256). Springer Berlin Heidelberg.

Sinha, M., Reddy, M., and Bhattacharyya, P., (2006). *An approach towards construction and application of multilingual indo-wordnet*. In 3rd Global Wordnet Conference (GWC 06), Jeju Island, Korea.

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