Semisupervied Data Driven Word Sense Disambiguation for Resource-poor Languages

Pratibha Rani[†], **Vikram Pudi**[†], **Dipti M. Sharma**[§]
[†]Data Sciences and Analytics Center, [§]Language Technologies Research Center
Kohli Center on Intelligent Systems
International Institute of Information Technology, Hyderabad, India
pratibha_rani@research.iiit.ac.in,{vikram, dipti}@iiit.ac.in

Abstract

In this paper, we present a generic semi-supervised Word Sense Disambiguation (WSD) method. Currently, the existing WSD methods extensively use domain resources and linguistic knowledge. Our proposed method extracts *context based lists* from a small sense-tagged and untagged training data without using domain knowledge. Experiments on Hindi and Marathi Tourism and Health domains show that it gives good performance without using any language specific linguistic information except the sense IDs present in the sense-tagged training set and works well even with small training data by handling the data sparsity issue. Other advantages are that domain expertise is not needed for crafting and selecting features to build the WSD model and it can handle the problem of non availability of matching contexts in sense-tagged training set. It also finds sense IDs of those test words which are not present in sense-tagged training set but their associated sense IDs are present. This feature can help human annotators while preparing sense-tagged corpus for a language by suggesting them probable senses of unknown words. These properties make the method generic and especially suitable for resource-poor languages and it can be used for various languages without requiring a large sense-tagged corpus.

1 Introduction

Word Sense Disambiguation (WSD) is considered as one of the most challenging Natural Language Processing (NLP) task and is described as an AI-complete problem (Navigli, 2009; Mallery, 1988). This is a classification task which involves determining the correct meaning of each word in a sentence/phrase based on the neighboring context words. Humans are very good at judging meaning of words in different contexts but when it comes to automate this task, it becomes very tough. Design of automated WSD methods, both supervised and unsupervised, requires the intuitive knowledge transfer from humans to WSD algorithms via knowledge structures like WordNet (Fellbaum, 1998; Banerjee and Pedersen, 2002), machine readable dictionaries (Lesk, 1986) and sense-tagged training corpus (Navigli, 2009). Creation of such knowledge structures is a costly and time taking process which requires extensive amount of domain resources and linguistic expertise. Along with this, domain expertise is also needed to create and select hand crafted features and rules from the training data which are required in the automated methods. These requirements make it difficult to design a WSD algorithm for (6500+) (Nakov and Ng, 2009) "resource-poor" languages.

The existing literature on WSD methods report that the naive *Most Frequent Sense* (MFS) baseline obtained from a sense-tagged corpus is very hard to beat (Navigli, 2009; Bhingardive et al., 2015b). When (Preiss et al., 2009) tried to refine the selection of most frequent sense by using supplementary linguistic resources like POS tagger and Lemmatizer of the concerned language they found that performance of such a system is limited by the performance of used linguistic resources. This observation shows that for resource-poor languages use of other linguistic resources is not much beneficial in WSD task, since their performances are also dependent on the availability of tagged/knowledge corpus. This inspires us to explore methods for WSD which do not rely on other linguistic resources and can take advantage of contextual information about words and

Proc. of ICON-2017, Kolkata, India. December 2017 ©*2017 NLPAI*, pages 503–512

senses present in the sense-tagged and raw untagged training sets. Also, the challenges of requiring domain expertise and non availability of large sets of sense-tagged data motivated us to develop semi-supervised methods for WSD task. The semi-supervised methods can take advantage of raw untagged data and would require only a moderate or small amount of sense-tagged training data. In semi-supervised scenario, WSD method builds its disambiguation model from a corpus of untagged raw sentences and a set of sense-tagged sentences and is formally defined as:

Using (1) sense IDs set $\Gamma = \{SID_1, SID_2, \ldots, SID_n\}$, (2) **sense-tagged** sentences set $AD = \{St_1, St_2, \ldots St_N\}$, where, $St_i = \langle W_1/SID_i, W_2/SID_j \ldots W_n/SID_k$, W_i is a word and SID_i is a sense ID from Γ and (3) **raw untagged** sentences set $RD = \{RS_1, RS_2 \ldots RS_M\}$, where $RS_i = \langle W_1 W_2 \ldots W_m \rangle$, build a WSD model Θ which outputs the best sense ID sequence $\langle SID_1 SID_2 \ldots SID_l \rangle$ for an input sequence of words $\langle W_1 W_2 \ldots W_l \rangle$.

Here, we propose a semi-supervised WSD method which uses the concept of *context based list* (Rani et al., 2016) to build the WSD model from a set of sense-tagged and raw untagged training corpus. Our proposed method is also influenced by the *one sense per collocation* hypothesis of Yarowsky (1993) which tells that the sense of a word in a document is effectively determined by its *context* (Yarowsky, 1995). Our approach takes help of raw untagged data and expands the notions of context and *context based list* (Rani et al., 2016) to tackle the data sparsity issue. Our method does not require any preprocessing such as, stop/non-content word removal and feature generation and selection from the sense-tagged training corpus. It works without using any additional knowledge structure like dictionary etc., other than the small sense-tagged corpus and moderate sized raw untagged data. This is easily obtainable even for resource-poor languages.

The obtained results show that our method performs well even with very small sized sense-tagged training data for Hindi and Marathi languages and its performance is better than the *Random Baseline* (Navigli, 2009) which selects a random sense for each polysemous test word, comparable to the *Most Frequent Sense* (MFS) baseline that selects the most frequent sense available in the sense-tagged training corpus for each polysemous word and at par with the reported results on the used datasets (Bhingardive et al., 2015a; Bhingardive et al., 2013; Khapra et al., 2011a; Khapra et al., 2011b; Khapra et al., 2008).

Rest of the paper is organized as follows: Section 2 presents related work. Section 3 describes our proposed approach. Section 4 presents and discusses the results and Section 5 concludes the paper and mentions future work directions.

2 Related Work

Generally, all the existing WSD techniques can be categorized into one of the following approaches (Navigli, 2009; Pal and Saha, 2015): i) Knowledge based approach, which uses knowledge structures like, WordNet (Fellbaum, 1998; Banerjee and Pedersen, 2002) or machine readable dictionaries (Lesk, 1986), ii) Supervised approach, which uses machine learning (Kågebäck and Salomonsson, 2016) and statistical methods (Iacobacci et al., 2016) on manually created sense-tagged training corpus. It also requires domain expertise for creating and selecting features and rules to be used for preprocessing and transforming the training data into the form required for designing the algorithm (Navigli, 2009; Iacobacci et al., 2016), iii) Unsupervised approach, which uses large amount of raw untagged training corpus (Pedersen and Bruce, 1997; Lin, 1998) to find word clusters which discriminates the senses of the words in different clusters, or use multilingual parallel corpora (Ide et al., 2002; Bhingardive et al., 2013), a knowledge resource like WordNet (Patwardhan et al., 2007; Chen et al., 2009; Bhingardive et al., 2015b; Bhingardive et al., 2015a) or multilingual dictionary (Khapra et al., 2011a), and iv) Semi-supervised approach, that uses both sense-tagged and untagged data in different proportions with different methods like, co-training with multilingual parallel corpora (Yu et al., 2011), bootstrapping (Yarowsky, 1995; Khapra et al., 2011b), neural network (Taghipour and Ng, 2015; Yuan et al., 2016) and word sense induction (Baskaya and Jurgens, 2016).

All types of WSD algorithms require knowledge structures and resources like, WordNet (Fellbaum, 1998; Banerjee and Pedersen, 2002), machine readable dictionaries (Lesk, 1986), sense-tagged training corpus (Navigli, 2009), parallel corpora and lage untagged raw corpus. Creation

of such knowledge structures and resources is a costly and time taking process which requires extensive amount of domain resources and linguistic expertise. Due to this, for resource-poor languages, special methods are needed which can handle data sparsity issue present in sense-tagged training data and can work with small/moderate set of untagged corpus without requiring knowledge structures and linguistic resources.

To handle the WSD task related challenges of resource-poor languages some specific methods have been proposed. For Chinese language, Yang and Huang (2012) propose handling data sparsity issue by using synonyms for expansion of context, their first method regards synonyms as topic contextual feature to train Bayesian model and second method treats context words made up of synonyms as pseudo training data. Baskaya and Jurgens (2016) propose a Word Sense Induction and Disambiguation (WSID) (Agirre and Soroa, 2007) model in which they combine a small amount of sense-tagged data with information obtained from word sense induction (a fully unsupervised technique that automatically learns the different senses of a word based on how it is used). Yu et al. (2011), Khapra et al. (2011b), Khapra et al. (2011a) and Bhingardive et al. (2013) propose methods to use one language to help other language by means of multilingual parallel corpora, multilingual dictionary, translation and bilingual bootstrapping. Mancini et al. (2016) and Bhingardive et al. (2015a) propose to use word and sense embeddings derived from raw untagged data and WordNet. In this method a large raw corpus is needed to obtain word embeddings.

Bhingardive et al. (2015a), Bhingardive et al. (2013), Khapra et al. (2011a), Khapra et al. (2011b) and Khapra et al. (2008) have reported results on the same dataset which we have used in our experiments. The method used in Khapra et al. (2008) combines sense distributions and sense co-occurrences learned from corpora with semantic relations present in WordNet by specially selecting linguistic features from the sense-tagged data, WordNet, multilingual sense dictionary and a parallel corpus. Khapra et al. (2011b) uses bilingual bootstrapping in which, a model is first trained using the seed annotated data of one language and then it is used to annotate the untagged data of other language and vice versa using parameter projection. Then from both the languages untagged instances annotated with high confidence are added to their seed data and the above process is repeated. Khapra et al. (2011a) uses an unsupervised bilingual Expectation Maximization (EM) based approach requiring synset-aligned bilingual dictionary and in-domain corpora of the concerned language pairs to estimate sense distributions of words in one language based on the raw counts of the words in the aligned synset in the other language. Bhingardive et al. (2013) add use of context in this EM method (Khapra et al., 2011a) and approximate the co-occurrence counts using WordNet-based similarity measures. Bhingardive et al. (2015a) further extends this EM method by using distributional similarity obtained from Word Embeddings to approximate the co-occurrence counts.

3 Proposed Semi-supervised Word Sense Disambiguation Method

Since a context can occur in multiple places in the text, we utilize the contextual similarity property based on *one sense per collocation* hypothesis of Yarowsky (1993) to develop our semi-supervised WSD method. We build upon the concept of *context based list* (CBL) proposed by Rani et al. (2016) for POS-tagging. They call the list of words occurring in a particular context as CBL and use association rule mining (Agrawal et al., 1993) for obtaining effective context based POS tagging rules from the set of tagged and raw untagged training data. We extend their idea by supplementing CBL with the concepts of *extended context list*, *context based sense list* and *context based word list* (defined below) to handle the peculiar problems of WSD due to data sparsity like:

1. Non availability of matching contexts of a word in sense-tagged training set. Use of raw untagged data with concept of *extended context list* helps in dealing with this problem.

2. Non availability of words in sense-tagged training set. Use of raw untagged data with concept of *context based lists* helps in dealing with this problem.

3. Large imbalance in frequencies of senses associated with a word in training set. Defined threshold parameters and *context based lists* help in handling this problem.

Our notion of *context* is a word pair, we use the left and right immediate neighboring words of a

Algo_Present($SIDListSet$, $MWordTaggedListSet$, $MWordUntaggedListSet$, W_t, W_{tl}, W_{tr})

1. **If** test word W_t and its context (W_{tl}, W_{tr}) is present as trigram (W_{tl}, W_t, W_{tr}) in sense-tagged text collection **Then:**
2. Find the corresponding sense IDs of W_t from set $SIDListSet$ and **Return** the sense ID having highest W_t count
3. **Else:**
4. Find set $ExpandTestConPList$ of contexts similar to (W_{tl}, W_{tr}) by finding its Extended Context List using set $MWordTaggedListSet$
5. Find set $ProbTestSIDList$ of all available sense IDs of W_t with their counts from sense-tagged text collection
6. From set $ExpandTestConPList$ find the contexts which are present in sense-tagged text collection with W_t as trigram using set $MWordTaggedListSet$ and from these trigrams select those having highest **ExtContextCount** value in set $ExpandTestConPList$ to make set $maxProbConSet$
7. **For each** context (W_{ptl}, W_{ptr}) of set $maxProbConSet$ **:**
8. Find the sense IDs associated with (W_{ptl}, W_{ptr}) using the set $SIDListSet$ and filter out those which exist in $ProbTestSIDList$ to make set $FinalTestSIDList$
9. **If** $FinalTestSIDList$ is not empty **Then:**
10. **Return** the sense ID from $FinalTestSIDList$ having highest W_t count
11. **Else:**
12. **If** Context Based Word List of context (W_{tl}, W_{tr}) obtained from set $MWordUntaggedListSet$ contains test word W_t **Then:**
13. Find the sense IDs associated with (W_{tl}, W_{tr}) using set $SIDListSet$ and filter out those which exist in $ProbTestSIDList$ to make set $ConFinalTestSIDList$
14. **If** $ConFinalTestSIDList$ is not empty **Then:**
15. **Return** the sense ID from $ConFinalTestSIDList$ having highest W_t count
16. **Else:**
17. **Return** the sense ID from $ProbTestSIDList$ having highest W_t count
18. **Else:**
19. **Return** the sense ID from $ProbTestSIDList$ having highest W_t count

Algo 1: Algorithm to find Sense ID of words present in sense-tagged text collection.

word/sense ID in a sentence/phrase as its context. Formally, in a given trigram $(W_{i-1}\ W_i\ W_{i+1})$ of words, (W_{i-1}, W_{i+1}) word pair is called *context* of W_i. The preceding word W_{i-1} is called *left context* and succeeding word W_{i+1} is called *right context*. Note that, in a text collection there can be multiple contexts available for a word. We use these terms in defining following concepts used in our WSD method:

Single Sense Word List is a list of word instances (with associated single sense ID) which have only one sense ID associated with them in the sense-tagged text collection.

Context Based Word List is a list of word instances from a text collection sharing the same context. For a given context, (W_l, W_r), its *context based word list* is the list of all words W_m having (W_l, W_r) as one of their contexts in the text collection. This list allows to store multiple instances of a word.

Context Based Sense List is a list of sense ID instances from a sense-tagged text collection sharing the same context. For a given context, (W_l, W_r), its *context based sense list* is the list of sense IDs SID_m having (W_l, W_r) as one of their contexts in the sense-tagged text collection. This list can store multiple instances of a sense ID.

Extended Context List: For a given context, (W_l, W_r) of a word W_m, let $PreListSet$ be the set of words obtained from those context based word lists which have left context W_l in their word list and let, $PostListSet$ be the set of words obtained from those context based word lists which have right context W_r in their word list.

Algo_Absent($SIDListSet$, $MWordTaggedListSet$, $MWordUntaggedListSet$, W_t, W_{tl}, W_{tr})

1. For test word W_t find Extended Context List set $ExpandTestConTagList$ of contexts similar to its context (W_{tl}, W_{tr}) using set $MWordTaggedListSet$
2. From set $ExpandTestConTagList$ select context (W_{extl}, W_{extr}) with highest **ExtContextCount** value
3. Find Context Based Word List $TrainExConListTest$ of (W_{extl}, W_{extr}) from $MWordTaggedListSet$
4. **If** $ListSupport(TrainExConListTest) \geq Minsizethreshold$ **Then:**
5. Using $SIDListSet$ find set $ProbTagSenset$ of sense IDs associated with $TrainExConListTest$ having $UniqueSenseSupport \geq (ListSupport(TrainExConListTest) \times Percentagethreshold)$
6. From set $ProbTagSenset$ find and **Return** $Predsentest$ having highest value of $TotalSenseSupport$ and set $Found = True$
7. **If** $Found \neq True$ **Then:**
8. Find Context Based Word List $RawConListTest$ associated with (W_{tl}, W_{tr}) from $MWordUntaggedListSet$ in which W_t is present
9. Find Context Based Word List $TrainConListTest$ of (W_{tl}, W_{tr}) from $MWordTaggedListSet$
10. **If** $ListSupport(RawConListTest) \geq Minsizethreshold$ and $ListSupport(TrainConListTest) \geq Minsizethreshold$ and Number of matching words between $RawConListTest$ and $TrainConListTest$ $\geq$ (size of smaller list among two -1) **Then:**
11. Using $SIDListSet$ find set $ProbTrSenset$ of sense IDs associated with $TrainConListTest$ having $UniqueSenseSupport \geq (ListSupport(TrainConListTest) \times Percentagethreshold)$
12. From set $ProbTrSenset$ find and **Return** $Predsentest$ having highest value of $TotalSenseSupport$ and set $Found = True$
13. **If** $Found \neq True$ **Then:**
14. Find Extended Context List set $ExpandTestConUntagList$ of contexts similar to context (W_{tl}, W_{tr}) using set $MWordUntaggedListSet$
15. From set $ExpandTestConUntagList$ select context (W_{exutl}, W_{exutr}) with highest **ExtContextCount** value
16. Find Context Based Word List $TrainUtExConListTest$ of (W_{exutl}, W_{exutr}) from $MWordTaggedListSet$
17. **If** $ListSupport(TrainUtExConListTest) \geq Minsizethreshold$ **Then:**
18. Using $SIDListSet$ find set $ProbUtSenset$ of sense IDs associated with $TrainUtExConListTest$ having $(UniqueSenseSupport \geq (ListSupport(TrainUtExConListTest) \times Percentagethreshold))$
19. From set $ProbUtSenset$ find and **Return** $Predsentest$ having highest value of $TotalSenseSupport$ and set $Found = True$
20. **If** $Found \neq True$ **Then:**
21. **Return** $NOEXISTSEN$

Algo 2: Algorithm to find Sense ID of words NOT present in sense-tagged text collection.

Let, $FullExtendConListSet$ be the set of all contexts (W_{pre}, W_{post}) prepared by taking word W_{pre} from $PreListSet$ and word W_{post} from $PostListSet$. Then, *extended context list* is the list of all those contexts from $FullExtendConListSet$ which have W_m in their context based word list. This list contains contexts similar to the given context (W_l, W_r). There is a count value **ExtContextCount** associated with each context present in *extended context list* which shows how many word combinations from $PreListSet$ and $PostListSet$ generated that context.

For a list of words L, in which multiple instances of a word can be present, we define following parameters:

ListSupport(L) is defined as the number of unique words present in L.

UniqueSenseSupport of a particular sense ID, SID, is defined as the number of unique words of L which have SID associated with them in the sense-tagged text collection.

TotalSenseSupport of a particular sense ID, SID, is defined as the total number of words of L (includes repeated occurrences of a word with a sense ID) which have SID associated with them in the sense-tagged text collection.

Minsizethreshold parameter defines the minimum number of words required to be present in a Context Based Word List to consider it for finding sense of words not present in sense-tagged text collection.

Percentagethreshold parameter is used for calculating percentage of words supporting a particular sense ID in a list of words L.

Overview of our WSD method

In the training phase, using a sliding window of size three, we collect all the *context based word lists*, *context based sense lists*, *single sense word list*, word and sense counts from the sense-tagged and raw untagged text collection in a single iteration, taking care of the sentence boundaries. Then in testing phase, Algo 1 and Algo 2 are used to find sense IDs of test words according to their presence/absence in the sense-tagged training set. Algo 1 always provides an output for test words present in sense-tagged training set but Algo 2 returns $NOEXISTSEN$ when it is not able to find any valid sense ID for test words not present in sense-tagged training set.

Both the algorithms use directly available immediate context information and indirectly available extended context information from the sense-tagged and raw untagged text collection in a priority order to handle the issues of non availability of matching contexts and imbalance in sense frequencies associated with a word in sense-tagged training set. Information obtained from sense-tagged set is given higher priority. Algo 2 uses raw untagged set to handle issue of non availability of words in sense-tagged training set and takes help of the defined support and threshold parameters to make confident choice of sense ID. Due to these properties it is able to find sense IDs of those test words also which are not present in sense-tagged training set but their associated sense IDs are present. The detailed steps involved in our WSD method are given in Section 3.1.

3.1 Word Sense Disambiguation Method

Following steps are used in our WSD method:

1. Find *Single Sense Word List* from the sense-tagged text collection.
2. Find set $\mathbf{SIDListSet}$ of *Context Based Sense Lists* of sense IDs from sense-tagged text collection.
3. Find set $\mathbf{MWordTaggedListSet}$ of *Context Based Word Lists* of words from sense-tagged text collection.
4. Find set $\mathbf{MWordUntaggedListSet}$ of *Context Based Word Lists* of words from raw untagged text collection.
5. If test word, W_t, present in sense-tagged text collection and is also present in *Single Sense Word List* then output associated sense ID. Else, find its context (W_{tl}, W_{tr}) from test sentence and apply Algo 1.
6. If test word, W_t, is not present in sense-tagged text collection then find its context (W_{tl}, W_{tr}) from test sentence and apply Algo 2.

4 Results and Discussion

We have used publicly available Health and Tourism domain sense-tagged corpus of Hindi and Marathi languages created by IIT Mumbai[1] (Khapra et al., 2010) and Hindi language raw untagged Health and Tourism domain ILCI data (Jha, 2010). Table 2 gives the dataset details. Table 1 shows average 4-fold cross validation results obtained by our algorithm for polysemous test words which are not present in the sense-tagged training set. Table 3 presents the average 4-fold cross validation results obtained for polysemous test words along with *Random Baseline* and MFS baseline results.

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

The results are presented in terms of Precision, Recall and F-Score accuracy measures as defined below (Navigli, 2009):

$$Precision = \frac{\text{No. of correctly predicted test words}}{\text{Total No. of predicted test words}} \quad (1)$$

Here, Total No. of predicted test words = (Total No. of test words - Test words flagged $NOEXISTSEN$ by algorithm).

$$Recall = \frac{\text{No. of correctly predicted test words}}{\text{Total No. of test words}} \quad (2)$$

$$FScore = \frac{2 \times Precision \times Recall}{Precision + Recall} \quad (3)$$

Table 1: Average 4-fold cross validation results obtained by our algorithm for polysemous test words NOT present in the sense-tagged training corpus.

Dataset	Precision (%)	Recall (%)	FScore (%)
Hindi Tourism	28.93	22.90	25.56
Marathi Tourism	34.50	12.0	18.0
Hindi Health	31.65	25.41	28.19
Marathi Health	32.43	8.72	13.74

The results of Table 1 shows the advantage of our approach in terms of ability to find sense IDs of those test words also which are not present in the sense-tagged training set but their associated sense IDs are present. To the best of our knowledge, currently supervised and semi-supervised WSD methods do not handle words absent in the sense-tagged training corpus. The *Random Baseline* and MFS baseline methods also can't find sense IDs for words which are absent in the sense-tagged training set. This ability can be used as a tool to help human annotators by suggesting them probable senses of unknown words while preparing sense-tagged corpus for a language.

To study the effect of parameter values on our approach, we experimented with parameter values $Minsizethreshold = 3, 5, 10$ and $Percentagethreshold = 0.5, 0.8$ and observed that variation in obtained results is very less ($\pm 0.5\%$) which shows that our approach is not very sensitive towards parameter values in this range of values. Following parameter values generated best results for our approach presented in Tables 1, 3 and 5: 1) For Hindi Tourism, $Minsizethreshold = 5$ and $Percentagethreshold = 0.8$. 2) For Hindi Health, $Minsizethreshold = 3$ and $Percentagethreshold = 0.8$. 3) For Marathi Tourism, $Minsizethreshold = 3$ and $Percentagethreshold = 0.5$. 4) For Marathi Health, $Minsizethreshold = 3$ and $Percentagethreshold = 0.5$. Our approach uses both the sense-tagged and raw untagged datasets of each domain mentioned in Table 2. We have divided the original Marathi Health and Tourism datasets into two exclusive parts and used one part as raw untagged set and other as tagged set.

Table 3 shows that results of our approach are better than the *Random Baseline* results and very close to the MFS baseline results. We can't directly compare our results with the earlier reported results (see Table 4) on these dataset by Bhingardive et al. (2015a), Bhingardive et al. (2013), Khapra et al. (2011a), Khapra et al. (2011b) and Khapra et al. (2008) due to difference in dataset size and content.

By observing the difference between reported accuracies of approach used by Khapra et al. (2008) and the MFS baseline results reported by them we can conclude that our simple generic approach gives results close to MFS baseline without using any complex feature selection process (domain based and generic) and without requiring too many linguistic and domain resources. For Hindi Tourism, Marathi Tourism and Hindi Health domains our results are better than the results reported by Bhingardive et al. (2015a), Bhingardive et al. (2013), Khapra et al. (2011b) and Khapra et al. (2011a) without using huge raw untagged data and without using any linguistic and domain resources like WordNet, a large multilingual parallel corpus or a multilingual dictionary which are required by the other methods.

Table 5 presents results for experiments with sense-tagged set size smaller than 100×10^3 words and shows that for small training set sizes (less than 50×10^3 words), Recall of our algorithm is better than MFS and Precision and F-Scores are in close range. Hence, it is a good choice for resource-poor languages, especially for those languages for which resources are in development phase. These results and our other experiments show that as sense-tagged training data size increases performance of our method also improves.

Table 2: Statistics of sense tagged and raw untagged datasets.

Dataset	Total No. of Sentences	Total No. of Words	No. of unique Words	No. of unique Sense IDs	Total No. of Polysemous Words	No. of unique Polysemous Words
Hindi Tourism sense-tagged	15395	424836	33500	8088	243959	5015
Marathi Tourism sense-tagged	13914	305337	54780	6307	141019	6758
Hindi Health sense-tagged	8001	189677	13356	4405	108006	2321
Marathi Health sense-tagged	6344	119764	21720	3643	47451	2790
Hindi Tourism raw untagged	24999	424128	29368	-	-	-
Hindi Health raw untagged	24461	447330	21811	-	-	-
Marathi Tourism raw untagged	2011	35208	11104	-	-	-
Marathi Health raw untagged	577	13468	4156	-	-	-

Table 3: Average 4-fold cross validation results obtained for polysemous test words.

Dataset	Our Approach			Random Baseline			MFS		
	Precision (%)	Recall (%)	FScore (%)	Precision (%)	Recall (%)	FScore (%)	Precision (%)	Recall (%)	FScore (%)
Hindi Tourism	76.22	76.14	76.18	39.39	39.39	39.39	78.66	78.27	78.46
Marathi Tourism	64.80	64.03	64.41	45.61	45.61	45.61	66.0	64.80	65.39
Hindi Health	69.97	69.79	69.88	45.47	45.47	45.47	71.45	70.72	71.08
Marathi Health	60.11	59.12	59.61	48.01	48.01	48.01	60.93	59.58	60.24

Table 4: Average 4-fold cross validation F-Score (%) results obtained for polysemous test words of various datasets by our approach and other WSD algorithms.

Algorithms	Hindi Tourism	Marathi Tourism	Hindi Health	Marathi Health
Our Approach	**76.18**	**64.41**	**69.88**	59.61
Bhingardive et al. (2015a)	-	-	60.94	61.30
Bhingardive et al. (2013)	60.70	58.67	59.63	59.77
Khapra et al. (2011a)	53.87	55.20	54.64	58.72
Khapra et al. (2011b)	60.67	61.90	57.99	**64.97**
Khapra et al. (2008)	**74.10**	**74.40**	**74.20**	**78.70**

To study the effect of raw untagged data size, for a particular size sense-tagged training set we varied the raw untagged data size in the range of 2×10^3 to maximum possible for that dataset and observed that as raw untagged data size increases the number of correctly predicted test words not existing in sense-tagged training set also increases which adds to the overall performance of our approach.

5 Conclusions and Future Work

In this paper, we proposed a generic semi-supervised method for Word Sense Disambiguation (WSD) task which uses concept of context based lists and extended context lists. It makes the WSD model without using domain knowledge from a small set of sense-tagged corpus along with raw untagged text data as training data. It works well with small training data also and handles data sparsity issue. It does not require domain expertise for crafting and selecting features to be used in the algorithm and outputs senses of those test words also which are not present in sense-tagged training set but their associated senses are present. It is generic enough to be used for WSD task of various languages without requiring a large sense-tagged corpus and is especially suitable for resource-poor languages. Our exper-

Table 5: Results obtained for polysemous test words for various sense-tagged training set sizes ($\leq 100 \times 10^3$ words).

Dataset	No. of Polysemous Test words	Sense tagged set size	Our Approach				MFS		
			Untagged set size	Precision (%)	Recall (%)	FScore (%)	Precision (%)	Recall (%)	FScore (%)
Hindi Tourism	45721	36457	424128	72.33	**70.40**	71.35	75.38	69.96	72.57
		38377		73.24	**71.32**	72.27	76.87	71.06	73.85
		76436		74.31	73.50	73.90	76.87	73.81	75.31
Marathi Tourism	33316	21747	35208	62.22	**47.77**	54.04	62.85	47.27	53.96
		43251		62.06	**53.05**	57.20	62.73	52.56	57.20
		85296		63.06	**58.16**	60.51	63.79	57.89	60.70
Hindi Health	21648	16936	447330	51.89	**49.35**	50.59	56.99	47.23	51.65
		31144		52.49	**50.93**	51.7	56.26	50.24	53.08
		59035		59.93	59.11	59.52	63.45	60.18	61.78
Marathi Health	10340	7665	13468	77.96	**57.89**	64.44	78.51	57.88	66.64
		15678		73.21	**61.94**	67.12	73.43	61.52	66.95
		33753		70.44	65.62	67.94	71.48	66.06	68.67
		75379		64.66	63.09	63.87	65.37	63.36	64.35
		94411		64.40	63.44	63.92	65.78	64.45	65.11

iments on Tourism and Health domains of Hindi and Marathi languages show good performance without using any language specific linguistic information.

Future work would be to test it on other languages including English. Further exploration can be done to enhance the property of finding sense IDs of non existing words. We can also try to include more generic features in the algorithm to enhance performance.

References

Eneko Agirre and Aitor Soroa. 2007. Semeval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems. In *Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval '07)*, pages 7–12. Association for Computational Linguistics.

Rakesh Agrawal, Tomasz Imieliński, and Arun Swami. 1993. Mining Association Rules Between Sets of Items in Large Databases. In *SIGMOD'93*, pages 207–216.

Satanjeev Banerjee and Ted Pedersen. 2002. An adapted Lesk algorithm for word sense disambiguation using WordNet. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 136–145. Springer.

Osman Baskaya and David Jurgens. 2016. Semi-supervised Learning with Induced Word Senses for State of the Art Word Sense Disambiguation. *J. Artif. Int. Res.*, 55(1):1025–1058.

Sudha Bhingardive, Samiulla Shaikh, and Pushpak Bhattacharyya. 2013. Neighbors Help: Bilingual Unsupervised WSD Using Context. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL*, volume 2: Short Papers, pages 538–542.

Sudha Bhingardive, Dhirendra Singh, V Rudramurthy, and Pushpak Bhattacharyya. 2015a. Using Word Embeddings for Bilingual Unsupervised WSD. In *Proceedings of the 12th International Conference on Natural Language Processing (ICON 2015)*, pages 59–64.

Sudha Bhingardive, Dhirendra Singh, Rudra Murthy V, Hanumant Harichandra Redkar, and Pushpak Bhattacharyya. 2015b. Unsupervised Most Frequent Sense Detection using Word Embeddings. In *NAACL HLT 2015, The 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1238–1243.

Ping Chen, Wei Ding, Chris Bowes, and David Brown. 2009. A Fully Unsupervised Word Sense Disambiguation Method Using Dependency Knowledge. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 28–36.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for Word Sense Disambiguation: An Evaluation Study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL*, volume 1.

Nancy Ide, Tomaz Erjavec, and Dan Tufis. 2002. Sense Discrimination with Parallel Corpora. In *Proceedings of the ACL-02 Workshop on Word Sense Disambiguation: Recent Successes and Future Directions - Volume 8*, pages 61–66.

Girish Nath Jha. 2010. The TDIL Program and the Indian Langauge Corpora Intitiative (ILCI). In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*. European Language Resources Association.

Mikael Kågebäck and Hans Salomonsson. 2016. Word sense disambiguation using a bidirectional LSTM. *CoRR*, abs/1606.03568.

Mitesh Khapra, Pushpak Bhattacharyya, Shashank Chauhan, Soumya Nair, and Aditya Sharma. 2008. Domain specific iterative word sense disambiguation in a multilingual setting. In *Proceedings of International Conference on NLP (ICON 2008)*.

Mitesh M. Khapra, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All Words Domain Adapted WSD: Finding a Middle Ground Between Supervision and Unsupervision. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1532–1541. Association for Computational Linguistics.

Mitesh M Khapra, Salil Joshi, and Pushpak Bhattacharyya. 2011a. 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. Asian Federation of Natural Language Processing.

Mitesh M. Khapra, Salil Joshi, Arindam Chatterjee, and Pushpak Bhattacharyya. 2011b. Together We Can: Bilingual Bootstrapping for WSD. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies - Volume 1*, pages 561–569. Association for Computational Linguistics.

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*, pages 24–26. ACM.

Dekang Lin. 1998. Automatic retrieval and clustering of similar words. In *Proceedings of the 17th international conference on Computational linguistics-Volume 2*, pages 768–774.

John C Mallery. 1988. Thinking about foreign policy: Finding an appropriate role for artificially intelligent computers. In *Master's thesis, MIT Political Science Department*.

Massimiliano Mancini, José Camacho-Collados, Ignacio Iacobacci, and Roberto Navigli. 2016. Embedding Words and Senses Together via Joint Knowledge-Enhanced Training. *CoRR*, abs/1612.02703.

Preslav Nakov and Hwee Tou Ng. 2009. Improved Statistical Machine Translation for Resource-poor Languages Using Related Resource-rich Languages. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 3*, pages 1358–1367. Association for Computational Linguistics.

Roberto Navigli. 2009. Word Sense Disambiguation: A Survey. *ACM Computing Survey*, 41(2):10:1–10:69.

Alok Ranjan Pal and Diganta Saha. 2015. Word sense disambiguation: A Survey. *CoRR*, abs/1508.01346.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. 2007. UMND1: Unsupervised Word Sense Disambiguation Using Contextual Semantic Relatedness. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 390–393.

Ted Pedersen and Rebecca Bruce. 1997. Distinguishing Word Senses in Untagged Text. In *eprint arXiv: cmp-lg/9706008*.

Judita Preiss, Jon Dehdari, Josh King, and Dennis Mehay. 2009. Refining the Most Frequent Sense Baseline. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions*, DEW '09, pages 10–18. Association for Computational Linguistics.

Pratibha Rani, Vikram Pudi, and Dipti Misra Sharma. 2016. A semi-supervised associative classification method for POS tagging. *International Journal of Data Science and Analytics*, 1(2):123–136.

Kaveh Taghipour and Hwee Tou Ng. 2015. Semi-Supervised Word Sense Disambiguation Using Word Embeddings in General and Specific Domains. In *HLT-NAACL*, pages 314–323.

Zhizhuo Yang and Heyan Huang. 2012. Chinese word sense disambiguation based on context expansion. In *COLING (Posters)*, pages 1401–1408.

David Yarowsky. 1993. One sense per collocation. In *Proceedings of the workshop on Human Language Technology*, pages 266–271.

David Yarowsky. 1995. Unsupervised Word Sense Disambiguation Rivaling Supervised Methods. In *ACL*, pages 189–196.

Mo Yu, Shu Wang, Conghui Zhu, and Tiejun Zhao. 2011. Semi-supervised learning for word sense disambiguation using parallel corpora. In *2011 Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)*, volume 3, pages 1490–1494. IEEE.

Dayu Yuan, Julian Richardson, Ryan Doherty, Colin Evans, and Eric Altendorf. 2016. Semi-supervised word sense disambiguation with neural models. *arXiv preprint arXiv:1603.07012*.