Using Word Embeddings for Bilingual Unsupervised WSD

Sudha Bhingardive Dhirendra Singh Rudramurthy V Pushpak Bhattacharyya
Department of Computer Science and Engineering,
Indian Institute of Technology Bombay.
{sudha,dhirendra,rudra,pb}@cse.iitb.ac.in

Abstract

Unsupervised Word Sense Disambiguation (WSD) is one of the challenging problems in natural language processing. Recently, an unsupervised bilingual WSD approach has been proposed. This approach uses context aware EM formulation for estimating the sense distribution by using the co-occurrence counts of cross-linked words in comparable corpora. WordNet-based similarity measures are used for approximating the co-occurrence counts. In this paper, we explore the feasibility of the use of Word Embeddings for approximating these counts, which is an extension to the existing approach. We evaluated our approach for Hindi-Marathi language pair, on Health domain. On using the combination of Word Embeddings and WordNet-based similarity measures, we observed 8.5% and 2.5% improvement in the F-score of verbs and adjectives respectively for Marathi and 7% improvement in the F-score of adjectives for Hindi. The experiments show that the combination of Word Embeddings and WordNet-based similarity measures is a reasonable approximation for the bilingual WSD.

1 Introduction

One of the well known research area in the field of Natural Language Processing (NLP) is the word sense disambiguation. Over the years, various WSD algorithms are proposed. These algorithms come under two broad categories, *viz.,* Knowledge based and Machine Learning based. Knowledge based approaches rely on various lexical knowledge resources like machine readable dictionaries, ontologies, WordNets *etc*. Machine learning based approaches are further classified as supervised, semi-supervised and unsupervised. Supervised WSD approaches (Lee et al., 2004; Ng and Lee, 1996) always perform better because of the availability of the sense-annotated data. However, the cost of creation of the sense-annotated data limits their applicability to only a few resource rich languages. On the other hand, semi-supervised approaches (Yarowsky, 1995; Khapra et al., 2010) provide a fine balance in terms of resource requirements and accuracy, but they still rely on some amount of sense-annotated data. Therefore, despite of the less accuracy, much focus is given for unsupervised WSD algorithms (Diab and Resnik, 2002; Kaji and Morimoto, 2002; Mihalcea et al., 2004; Jean, 2004; Khapra et al., 2011). These algorithms do not need any sense-annotated data for the disambiguation. Moreover, they make use of lexical knowledge resources or comparable/parallel corpora for training the algorithm (Kaji and Morimoto, 2002; Diab and Resnik, 2002; Specia et al., 2005; Lefever and Hoste, 2010; Khapra et al., 2011).

Khapra et al. (2011) have shown that how two resource deprived languages can help each other in WSD without using any sense-annotated data in either of the languages. Here, the intuition is that, the sense distribution remains same across languages when the comparable corpora is provided. They used the Expectation Maximization (EM) based formulation for estimating the sense distribution of words. Further, Bhingardive et al. (2013) extended this approach and hypothesized that, the co-occurrence sense distribution also remains same across languages, given the comparable corpora. Since, the co-occurrence counts require a large corpora, they approximate the co-occurrence counts using WordNet-based similarity measures. An improvement of 17% - 35% in the F-Score of verbs was observed while the F-Score was comparable for other POS categories.

In this paper, we propose to explore the use of

D S Sharma, R Sangal and E Sherly. Proc. of the 12th Intl. Conference on Natural Language Processing, pages 59–64,
Trivandrum, India. December 2015.

distributional similarity measures as an approximation to the co-occurrence counts in Bhingardive et al. (2013) approach. We used the cosine distance between the word embeddings of the words as a similarity measure. These word embeddings are obtained from a large monolingual corpus.

The roadmap of the paper is as follows. Section 2 covers the background work on Bilingual EM Our extension of the Bilingual EM using distributional similarity is explained in Section 3. Section 4 gives detail about the experimental setup. Results are presented in section 5. Section 6 covers discussion on the results. Related work is given in section 7. Conclusion and future work are presented in section 8.

2 Bilingual EM using WordNet-based Similarity

Bhingardive et al. (2013) extended the bilingual EM approach (Khapra et al., 2011) and observed that adding contextual information further helps in the disambiguation process. Original bilingual EM approach estimates the sense distribution in one language by using the raw counts of the cross-linked words from the other language using EM algorithm. Bhingardive et al. (2013) modified this approach by replacing the raw counts of the words with the co-occurrence counts of the target word and the context words. They approximated the co-occurrence counts by using WordNet based similarity measures to avoid the data sparsity. The modified EM formulation with context information is as follows:

E-Step:

$$P(S_k^{L_1}|u,a) = \frac{\sum_{v,b} P(\pi_{L_2}(S_k^{L_1})|v,b) \cdot simi(v,b)}{\sum_{S_i^{L_1}} \sum_{x,b} P(\pi_{L_2}(S_i^{L_1})|x,b) \cdot simi(x,b)}$$

$$\begin{aligned} \textit{where}, S_i^{L_1}, S_k^{L_1} &\in synsets_{L_1}(u) \\ a &\in context(u) \\ v &\in crosslinks_{L_2}(u, S_k^{L_1}) \\ b &\in crosslinks_{L_2}(a) \\ x &\in crosslinks_{L_2}(u, S_i^{L_1}) \end{aligned}$$

Here, u is the target word to be disambiguated, a is the context word, $\pi_{L_2}(S_k^{L_1})$ means the linked synset of the sense S_k^{L1} in L_2. $simi(x,b)$ is the WordNet based similarity over all senses of words x and b. $crosslinks_{L_2}(u, S_j^{L_1})$ is the set of possible translations of the word 'u' from language L_1 to L_2 in the sense $S_j^{L_1}$. $crosslinks_{L_2}(a)$ is the set of all possible translations of the word 'a' from L_1 to L_2 in all its senses.

M-Step:

$$P(S_j^{L_2}|v,b) = \frac{\sum_{u,a} P(\pi_{L_1}(S_j^{L_2})|u,a) \cdot simi(u,a)}{\sum_{S_i^{L_2}} \sum_{y,b} P(\pi_{L_1}(S_i^{L_2})|y,a) \cdot simi(y,a)}$$

$$\begin{aligned} \textit{where}, S_i^{L_2}, S_j^{L_2} &\in synsets_{L_2}(v) \\ b &\in context(v) \\ u &\in crosslinks_{L_1}(v, S_j^{L_2}) \\ a &\in crosslinks_{L_1}(b) \\ y &\in crosslinks_{L_1}(v, S_i^{L_2}) \end{aligned}$$

where, $simi(y,a)$ is the WordNet based similarity over all senses of words y and a.

Here, given a target word and its context in one language, the probability of various senses of the target word is calculated in that particular context using their cross-links information from other language. Synset aligned multilingual dictionary (Mohanty et al., 2008) is used to find the cross-links of the target word and its context words in other language. The probability of the sense of the target word given its context is estimated by the modified EM formulation as mentioned earlier. The maximum similarity over all senses of the target word is chosen as the sense of the target word. In this way, given the bilingual comparable corpora and the synset aligned dictionary, context aware EM formulation is used to estimate the sense distributions in both the languages.

3 Our approach: Bilingual EM using Distributional Similarity

Continuous word embeddings have recently gained popularity in various NLP tasks like POS Tagging, Named Entity Recognition, Semantic Role Labeling, Sentiment Analysis, *etc*. (Collobert et al., 2011; Tang et al., 2014). Word embeddings have shown to capture the syntactic and semantic information about a word. In our approach, we look forward to use these word embeddings for the bilingual WSD and compare the results with the existing approaches.

WSD Algorithm	HIN-HEALTH				
	NOUN	ADV	ADJ	VERB	Overall
EM-C-DistSimi+WnSimi	59.32	68.98	63.18	60.02	**60.94**
EM-C-DistSimi	59.59	69.20	**63.87**	55.73	61.09
EM-C-WnSimi	59.82	67.80	56.66	**60.38**	59.63
EM	**60.68**	67.48	55.54	25.29	58.16
WFS	53.49	**73.24**	55.16	38.64	54.46
RB	32.52	45.08	35.42	17.93	33.31

Table 1: Comparison(F-Score) of our approach (EM-C-DistSimi-WnSimi and EM-C-DistSimi) with other WSD algorithms on Hindi-Health domain

WSD Algorithm	MAR-HEALTH				
	NOUN	ADV	ADJ	VERB	Overall
EM-C-DistSimi+WnSimi	62.75	61.19	**56.22**	**60.99**	**61.30**
EM-C-DistSimi	63.09	61.82	55.60	43.69	58.92
EM-C-WnSimi	62.90	62.54	53.63	52.49	59.77
EM	**63.88**	58.88	55.71	35.60	58.03
WFS	59.35	**67.32**	38.12	34.91	52.57
RB	33.83	38.76	37.68	18.49	32.45

Table 2: Comparison(F-Score) of our approach (EM-C-DistSimi-WnSimi and EM-C-DistSimi) with other WSD algorithms on Marathi-Health domain

Our formulation is based on Bhingardive et al. (2013) formulation, where we use distributional similarity measures along with WordNet based similarity measures for finding the sense distribution. As shown previously, in E-step and M-step, $simi(v, b)$, $simi(x, b)$, $simi(u, a)$ and $simi(y, a)$ are computed as the distributional similarities (cosine distance) calculated from the large untagged text.

4 Experimental setup

In this section, we describe various datasets used in our experiments. We discuss how we obtained word embeddings and evaluated their quality.

4.1 Datasets used for WSD

In our experiments, we used the same corpus as used by Khapra et al. (2011) . This corpus is publicly available[1] for Health domain.

4.2 Training Word Embeddings

The word embeddings were obtained using *word2vec*[2] tool (Mikolov et al., 2013). This tool provides two broad techniques for creating word embeddings : Continuous Bag of Words (CBOW) and Skip-gram models. CBOW model predicts the current word based on the surrounding context whereas, the Skip-gram model tries to maximize the probability of seeing the context word given the word under consideration (Mikolov et al., 2013).

We have used the most widely used hyperparameter settings for training word embeddings. The Skip-gram model is used with 300 dimensions along with the window size equal to 5 (i.e. $w = 5$).

Word Embeddings for Hindi

For obtaining the word embeddings for Hindi, we used Bojar et al. (2014) corpus. This corpus contains around 812.6 million words along with POS and lemma information. We have trained the word embeddings using the lemmatized version of the corpus.

Word Embeddings for Marathi

Marathi corpus was collected from various resources like Web & Wikipedia dumps[3], Leipzig corpus[4] , Newspaper corpus from Maharashtra Times [5] & e-Sakal, [6] *etc*. This corpus contains

[1] http://www.cfilt.iitb.ac.in/wsd/annotated_corpus/
[2] https://code.google.com/p/word2vec/
[3] http://ufal.mff.cuni.cz/ majlis/w2c/download.html
[4] http://corpora.uni-leipzig.de/
[5] http://maharashtratimes.indiatimes.com/
[6] http://online4.esakal.com/

approximately 26.3 million words. Marathi word embeddings are trained on this corpus using the same parameters as used for Hindi.

4.3 Evaluating the quality of Word Embeddings

For evaluating the quality of both Hindi and Marathi word embeddings, we have created a *similarity word pair* dataset by translating the standard similarity word pair dataset (Finkelstein et al., 2001) available for English. Three annotators were instructed to give the score for each word-pair based on the semantic similarity and relatedness. The scale was chosen between 0 - 10. The least similar word-pair was given a score of 0, while the most similar word-pair was given a score of 10. We calculated the average inter-annotator agreement using Spearman's correlation coefficient. The embeddings giving the best Pearson's correlation coefficient was used in our experiments.

4.4 WSD Experiments

We performed WSD experiments on all content words. The entire sentence was considered as the context for the word to be disambiguated. Experiments are performed on Hindi-Marathi health domain corpus.

The F-Score of the following WSD algorithms are reported.

Random Baseline [RB]

This algorithm assigns the senses randomly to the words to be disambiguated.

Wordnet First Sense [WFS]

WFS baseline assigns the first listed sense in the WordNet to the word irrespective of its context.

Basic EM [EM]

This is basic EM approach by Khapra et al., (2011) which estimates the sense probability of a word in one language by using the raw counts of its cross-linked words in another language.

EM-C-WnSimi

This is an extended EM approach where Bhingardive et al. (2013) modified the basic EM formulation by adding the contextual information and using the WordNet based similarity for approximating the co-occurrence counts.

EM-C-DistSimi

This is our approach where we modify the formulation of Bhingardive et al. (2013) using the distributional similarity measure for estimating the sense distributions.

EM-C-DistSimi-WnSimi

This is also our approach where we use combination of distributional and WordNet similarity for estimating the sense distributions.

5 Results

In this section, we discuss our results and compare it with other WSD approaches. Table 1 and Table 2 show the results of our approach on Hindi-Health and Marathi-Health domain respectively. Results are given in terms of F-score. EM-C-DitSimi-WnSimi and EM-C-DistSimi achieves better result as compared to EM-C-WnSimi and EM. Using EM-C-DistSimi-WnSimi approach, verb accuracy increases at the level 8.5% for Marathi and for Hindi, it is very close to the existing approaches. The adjective accuracy also improved by 7% for Hindi and 2.5% for Marathi. Results for noun and adverb are observed very close to the existing approaches. The overall F-Score obtained is comparable. The results show that word embeddings can be used as an approximation along with WordNet-Based similarity measures for bilingual WSD.

6 Discussion

6.1 Poor performance for verbs using Word Embeddings

It has been observed that if we use only distributional measure (EM-C-DistSimi) as an approximation then we get significant performance except for verbs. We believe the reason that the word embeddings of verbs fail to capture the semantic information resulting in poor performance. Therefore, the word embeddings of verb fails in finding out relevant context words and choose its correct sense. But if we use the combination of distributional similarity and wordnet similarity then we get better results for the same.

6.2 Misleading contexts

In our approach, we consider the entire sentence as the context for performing WSD. As a result, we end up choosing many context words which causes topic drift. The approach needs to be care-

ful while selecting the context words for the disambiguation task.

7 Related work

Recently, several unsupervised WSD algorithms have been proposed. McCarthy et. al (2004) used distributional methods for finding the context clues for unsupervised most frequent sense detection. They have shown that MFS can be detected without the need of any sense tagged corpora. Only untagged text is used for finding the predominant senses of words. Parallel or comparable corpora have also been explored for unsuperwised WSD (Diab and Resnik, 2002; Kaji and Morimoto, 2002; Mihalcea et al., 2004; Jean, 2004; Khapra et al., 2011). Chen et. al (2014) have presented a unified model which focused on creating sense representations using word embeddings and used the same for the disambiguation purpose.

8 Conclusion and Future Work

We explored the usefulness of word embeddings from a bilingual WSD perspective. We used the distributional similarity measure as an approximation to the co-occurrence counts in bilingual EM Context based WSD. We found that the word embeddings along with wordnet similarity measure are a reasonable approximation to the simple co-occurrence counts. We also observed that the word embeddings for verbs fail to capture the relevant semantic information. Much focus is needed on getting the good quality word embeddings for verbs. We would also like to explore the strategies for choosing the most informative context words for disambiguation depending on the POS category of the word.

References

Sudha Bhingardive, Samiulla Shaikh, and Pushpak Bhattacharyya. 2013. Neighbors help: Bilingual unsupervised wsd using context. In *ACL (2)*, pages 538–542. The Association for Computer Linguistics.

Ondrej Bojar, Vojtěch Diatka, Pavel Rychlý, Pavel Stranak, Vit Suchomel, Aleš Tamchyna, and Daniel Zeman. 2014. Hindencorp - hindi-english and hindi-only corpus for machine translation. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Hrafn Loftsson, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, Reykjavik, Iceland, may. European Language Resources Association (ELRA).

Xinxiong Chen, Zhiyuan Liu, and Maosong Sun. 2014. A unified model for word sense representation and disambiguation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1025–1035. Association for Computational Linguistics.

Ronan Collobert, Jason Weston, Léon Bottou, Michael Karlen, Koray Kavukcuoglu, and Pavel Kuksa. 2011. Natural language processing (almost) from scratch. *J. Mach. Learn. Res.*, 12:2493–2537, November.

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.

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2001. Placing search in context: The concept revisited. *ACM Trans. Inf. Syst.*

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, Anup Kulkarni, Saurabh Sohoney, and Pushpak Bhattacharyya. 2010. All words domain adapted wsd: Finding a middle ground between supervision and unsupervision. In Jan Hajic, Sandra Carberry, and Stephen Clark, editors, *ACL*, pages 1532–1541. The Association for Computer Linguistics.

Mitesh M Khapra, Salil Joshi, and Pushpak Bhattacharyya. 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.

Els Lefever and Veronique Hoste. 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. ACL.

Diana Mccarthy, Rob Koeling, Julie Weeds, and John Carroll. 2004. Finding predominant word senses in untagged text. In *In Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics*, pages 280–287.

Rada Mihalcea, Paul Tarau, and Elizabeth Figa. 2004. Pagerank on semantic networks, with application to word sense disambiguation. In *COLING*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

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. ACL.

Lucia Specia, Maria Das Graças, Volpe Nunes, and Mark Stevenson. 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.

Duyu Tang, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin. 2014. Learning sentiment-specific word embedding for twitter sentiment classification. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1555–1565. Association for Computational Linguistics.

David Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the 33rd Annual Meeting on Association for Computational Linguistics*, ACL '95, pages 189–196, Stroudsburg, PA, USA. Association for Computational Linguistics.