

Multilingual Approach to Word Sense Disambiguation using word2vec and Sense Embedding with Semantic Cloud

Shreya Patankar^a, Satish Devane^b

^a Datta Meghe College of Engineering, Airoli, NaviMumbai, 400708, India

^b Datta Meghe College of Engineering, Airoli, NaviMumbai, 400708, India

ARTICLE INFO

Article history:

Received 00 December 00

Received in revised form 00 January 00

Accepted 00 February 00

Keywords:

Natural language

Word Sense Disambiguation

Knowledge Based System

ABSTRACT

Word Sense Disambiguation(WSD) is an open problem in the area of Natural language processing. Many WSD systems uses knowledge based unsupervised approach as it bares us from using tagged corpus. But this approach depends extensively on knowledge. Enhancing the knowledge for ambiguity resolution is one of the important issues in WSD. Although Multilingual lexical resource is an asset for knowledge based systems(KBS), many KBS have not yield satisfactory results due to lack of sufficient information required for disambiguation. Multilingual information can help boost the performance of WSD in various languages. Combining lexical semantic information from different languages helps improve WSD task. We make use of Babel Net, a huge multilingual resource provider to attain the task and design a semantic word cloud to provide high performance disambiguation by imposing additional knowledge.

© 2019 NGCT and University of Petroleum and Energy Studies (UPES), Dehradun. Hosting by SSRN (ISN) All rights reserved.

Peer review under responsibility of UPES Dehradun and NGCT 2019.

1. Introduction

Here Large amount of data is available usually in the form of texts. Most of the textual information is available in English language which is also one of the natural languages in which human beings communicate. Increasing demands by the user to access text data in various languages opens up the doors of Multilingual natural language processing. It is an important concern in today's world to process this textual data and obtain useful and meaningful information from these texts and use it for various purposes like machine translation, information retrieval etc. WSD has proved to be a key step in performance improvement of many such systems. Lot of researchers have focused on English language WSD. However due to lack of sufficient resource availability, WSD in Multilingual setting has not achieved satisfactory results. With the availability of Multilingual dictionaries, Sense disambiguation can be enhanced using multilingual content which depicts the need for multilingual WSD which is ever growing.

Lexical and semantic information from various languages can be an added advantage to performance improvement of WSD. However earlier due to lack of resources, the development of Multilingual approach to WSD has been hampered. Due to invent of Multilingual dictionary, the idea of achieving sense disambiguation is reinvented in a novel way. Combining lexical and semantic language independent information for disambiguation makes understanding of text better and adds more knowledge to the disambiguation process. But the goal is far from being achieved as it is clearly evident that work on WSD is focussed on English language. Various KBS were designed in the past for Monolingual WSD which tried to enrich the knowledge of WSD systems but tend to show poor results in terms of accuracy. Magnini et.al., 2004 made use of domain knowledge to enrich the lexical resource synsets. Researchers have also attempted to showcase knowledge based approaches on Indian languages with accuracy upto 55% average as presented by Patankar et.al.,2017.

Deficiency of resources obstructed the growth of Multilingual WSD. It is evident that each language has different set of ambiguities. Ambiguous words in one language may have only single sense in other languages. Retrieving this sense will boost the sense disambiguation process. Making use of resources from the Multilingual dictionary is an added advantage. The proposed work combines lexico semantic information from various languages together to provide performance enhancement to the WSD system and also opens up a different way of approaching Multilingual WSD by making use of Babelnet, a wide ontological structure exploring semantic and syntactic knowledge. Relying only on Multilingual KBS may hamper the growth of WSD systems. Although multilingual dictionaries provide wide coverage exploring the interconnected ontology structure, various issues still remain to be seen. Proper nouns are not part of the dictionary. Correlation between most frequent words and rare contextual words lack dictionary coverage. This may hamper the disambiguation process and accuracy. Additional external knowledge by means of raw text is needed. Hence unsupervised approach is a natural choice. Popularity of word and sense embedding in the recent years have shown good performance accuracy in WSD systems. We aim to build an open framework for WSD by imposing external knowledge using word and sense representation by designing a semantic word sense cloud which will extract semantic

knowledge from thesaurus and external resource Wikipedia.

2. Review of Literature

Knowledge based approaches tend to use the external resources like Multilingual dictionary for WSD. Kulkarni et.al., 2010 explored use of monolingual information into another language. Navigli et.al., 2012 introduced a novel graph based algorithm to influence the performance of WSD systems and determine the most suitable approach for ambiguity resolution. Minkov et.al., 2012 learned graph based similarity measures for extracting word synonymy from plain text. Aggire et.al., 2009 WordNet based distributional similarity approaches. Various similarity measures are well explored by Meng et.al., 2013. Navigli et.al., 2004 explore the structural pattern recognition work. Recent works by Navigli et.al., 2012 also report on using Multilingual Dictionary to improve the Knowledge based systems. Our work is an inspiration from Navigli et.al, 2012 to work with Multilingual WSD. Instead of using parallel corpora for working with Multilingual WSD system, we make use of Babelnet, a huge multilingual lexical knowledge base covering as many as 284 languages. This approach provides a multilingual semantic information in association with Wikipedia's knowledge.

Word embedding are studied using Mikolov's model in Mikolov et.al., 2013. Word representation limits the polysemy words into a single vector. Polysemy disambiguation is crucial for various applications. Concept of sense representation became adherent with Multilingual thesaurus. Sudha et.al., 2015 beautifully presented sense representation in a novel way by considering the entire ontological relations like hypernym, hyponym, gloss etc. that exist between the different senses.

3. Proposed Method

We present our concept of providing novel prospective for Multilingual WSD with Babelnet as external Multilingual dictionary. We then represent our methodology for disambiguating senses using Multilingual WSD and the elements supporting it including graph based approach and similarity score calculation. Further we provide additional knowledge to the system by means of unsupervised disambiguation. It is a natural choice due to absence of resources in other languages. We then present word and sense representations for achieving satisfactory performance.

3.1 Multilingual WSD

We present our system of exploiting Multilingual WSD. One word ambiguous in one language may not be ambiguous in other languages. We present our system in Figure 1 below.

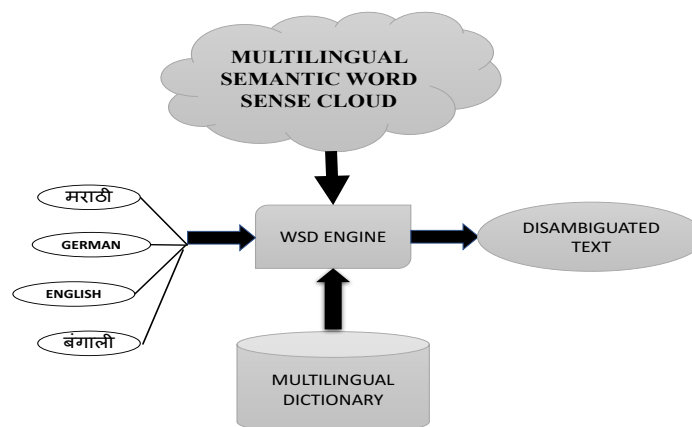


Fig. 1. Multilingual WSD

For example, पैरू in Marathi language having two senses is non-ambiguous in English and Hindi languages. Based on this concept, we take the translations of the ambiguous word in other languages. We design a Multilingual WSD system by taking the input in the form of context words, ambiguous word and its non-ambiguous translation making full use of Multilingual dictionary to provide the translated senses.

3.2 Babelnet

Babelnet is a huge multilingual ontological network incorporating lexical semantic and syntactic knowledge from various languages with enriched methodology. It represents labelled graph specifying semantic relations between various nodes and edges. It combines the knowledge of various language WordNet and largest multilingual encyclopaedia i.e. Wikipedia making it a huge multilingual encyclopaedic dictionary. The integration is performed via an automatic mapping and by filling in lexical gaps in resource poor language with the help of Machine Translation. Representation of Babelnet is shown in figure 2. We use Babelnet to represent sense definitions for our work.

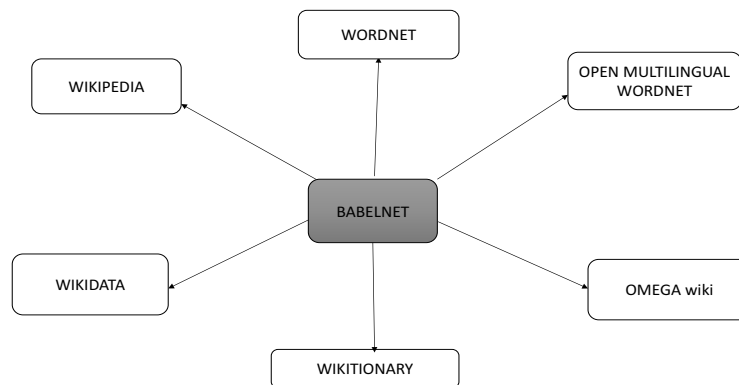


Fig. 2. Babelnet

3.3 Word Embedding

Word embedding are immensely popular in Natural language processing. They are based on Distributional Hypothesis which works under the assumption that similar words occur in similar contexts. Word representations mean converting plain text into numeric form as Machine learning and deep learning architectures are not capable of processing text in raw form. It represents embedding continuous vector space with lesser dimensions. Word representation is shown in figure 3.



Fig. 3. Word representation

Word representation is performed using two models namely Continuous Bag of Words (CBOW) and Skip Gram model. CBOW predicts the word representation based on neighbouring contextual information and Skip Gram model tries to maximize the probability of the word based on other words in the same sentence.

3.4 Context Sense Representation

Senses are also represented using numeric values just like word representations with the senses in the lexical semantic database. Formally we design a

semantic word sense representation using the following procedure.

- A Sentence comprising of words W and target word T has a context size CS which defines the number of words surrounding the target word T . The sentence is pre-processed, tokenized and each word is labelled with its Part of Speech Tag.
- Inverse path length sum is used to compute the similarity score given by Babelnet. $Score_w$
- We remove the stop words from the sentence and form the context vector as follows
- Context Vector $V = word2Vec \{W, T\}$
- We consider averaging of all the word embedding vectors of the context Vector $S = \sum_{i=1}^N \frac{V}{N}$

Where N is the number of words present in the Context Vector V

- We also retrieve different senses of the target word from the ontological structure such as synset members, gloss or example sentence, relations such as hypernym or hyponym to represent a sense bag for the target word as shown in figure 4 below. This sense bag $SB(T)$ represents Babelnet based features for T which are synset members, gloss, content words in the example sentence, hypernym and hyponym synsets and glosses, meronym synsets and glosses and example sentences of hypernym, hyponym and meronym.

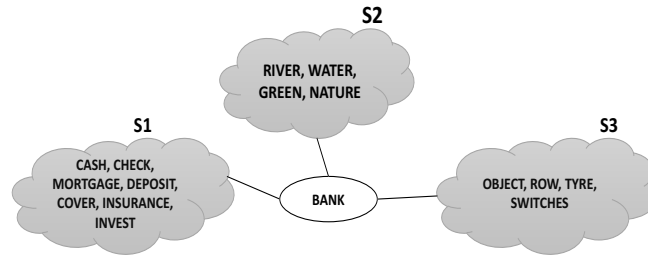


Fig. 4. Sense Representation

- We then consider the Wikipedia information of the target word represented in Figure 5 below.



Fig. 5 Wikipedia representation of Senses

- Let wk_i be the Wikipedia sense bag for the i th sense of the target word T represented in the sense representation. Features are represented by the example sentences of the word sense of the word.

$$WSB_i = \{Features(wk_i)\} \quad (4)$$

- Extended Lesk algorithm is applied to perform the overlap between the sense representation and Wikipedia representation of target word T to retrieve the most frequent words from the subset related to target word.
- Combining the data from (3) and (4) to represent the sense cloud SC for the sense representation is given below.

$$SC(T) = SB(T) \cap WSB_T \quad (5)$$

- Sense embedding are then represented using the most frequent words from the subset by taking the word embedding of each word from the subset of words in the sense representation. Let w be the word with t senses $w_{M1}, w_{M2}, \dots, w_{Mt}$ considered from the Babelnet senses.

$$\text{Vec}(\text{SC}(T)) = \frac{\sum \text{vec}(\text{SC}(T))}{N}$$

Where N is the number of words present in the sense cloud $\text{SC}(T)$

- We calculate the similarity measure by considering the cosine similarity between the word representation of context Vector and sense cloud representation.

$$\text{Cos}(\text{Context Vector } S, \text{SC}(T)) = \frac{\sum_{i=1}^n S * \text{SC}(T)}{\sqrt{\sum_{i=1}^n S^2} * \sqrt{\sum_{i=1}^n \text{SC}(T)^2}}$$

- Most frequent sense (MFS) for a given word is calculated by combining the scores of Eq. (1), Eq. (2) and Eq. (5) to boost the disambiguation accuracy.

$$\text{MFS}_w = \sum [\text{Score}(w), \text{Cos}(\text{vec}(w_i), \text{vec}(\text{SC}_{w_{si}}))]$$

4. Experiments and Results

We collected our instances from www.google.com and represent various domains like general, news, medicine etc. Our dataset consists of 530 test instances representing 20 polysemy nouns. We also considered translations of the target word represented using the Babelnet Multilingual Dictionary. We evaluated the performance of our system using Precision and Recall. Precision is defined as the ratio of the correctly answered instances and the total number of test instances for the target word. Recall is defined as the ratio of the correctly answered instances and total number of test instances to be answered for the target word.

4.1 Training the Word Embedding

Words are represented as Vectors for performing the score calculation.

4.2 Sense Representation

Senses are retrieved related to the ambiguous words and are converted into vectors. We also experimented with the context size to determine the most suitable context size for disambiguation process. Table 1 shows the tasks.

Table 1. Context Size

Representation	Precision	Recall
Context size 4	0.411	0.421
Context size 6	0.423	0.426
Context size 8	0.475	0.465
Context size 10	0.472	0.455

From Table 1, we observed a good response for the context size of 8. Further we represent senses by building a context vector by averaging the summation of the word vectors of the words present in the context and the target word. After computing the cosine similarity and calculating the Most Frequent sense for disambiguation, we compared the performance of our Multilingual method with other systems in terms of precision and recall represented in Table 2.

Table 2. Performance of our method and other Unsupervised systems in English lexical sample task.

System	Precision	Recall
Most frequent sense baseline	0.552	0.552
Lesk algorithm	0.097	0.053
Adapted Lesk	0.240	0.234
UMFS (Bhingardive)	0.433	0.432
Multilingual WSD with Sense embedding (Our method)	0.535	0.535

5. Conclusion

In our paper, we presented Multilingual approach to WSD and an unsupervised approach for determining the most frequent sense for polysemy nouns by means of word and sense representations. Multilingual WSD exploit graph based method to collect evidences from translations in various languages. This is beneficial to boost the disambiguation accuracy. The results show that our method of Multilingual WSD with sense embedding outperforms various baselines. The framework is open and can be easily adapted across other languages. The framework could also benefit by exploring various Multilingual dictionaries and various similarity measures. Also, we plan to explore sense representation for multiple languages.

REFERENCES

- Magnini, B., Giampiccolo, D., Vallian A., (2004). The Italian lexical sample task at Senseval-3, In Proceedings of the 3rd International workshop on the Evaluation of Systems for the Semantic Analysis of Text (SENSEVAL-3), pp. 17-20.
- Patankar, S., Devane, S., (2017), Issues in resolving word sense disambiguation for Multilingual translation framework, International conference on Intelligent Computing and Control Systems ICI CCS, pp.230–235.
- Kulkarni, A., Khapra, M., Sohoney, S., Bhattacharya, P., (2010), CFILT: Resource conscious approaches for all-word domain specific WSD, In proceedings of the 5th International workshop on Semantic Evaluations (SemEval-2010), pp.421–426.
- Navigli, R., Lapata, M., (2012), IEEE Transactions on pattern analysis and Machine intelligence, IEEE computer society, vol 32.
- Minkov, E., Cohen, W., (2012), Graph based similarity measures for Synonym extraction from parsed text, Workshop proceedings of Text graph-7, Association for computational Linguistics, pp 20-24.
- Agirre, E., Alfonseca, E., Hall, K., Kravalova, J., Pasca, M., Soroa, A., (2009), A study of similarity and relatedness using distributional and wordnet based approaches, Human language Technologies: The 2009 Annual conference of the North American Chapter of the ACL, pp. 19-27.
- Meng, L., Huang, R., Gu, J., (2013), A review of semantic similarity measures in wordnet, International Journal of Hybrid Information Technology, Vol 6, No. 1.
- Navigli, R., Velardi, P., (2004), Structural Semantic Interconnection: a knowledge based approach to word sense disambiguation, SENSEVAL-3: Third International workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain.
- Navigli, R., Ponzotto, P., (2012), Joining Forces pays off: Multilingual joint word sense disambiguation, Proceedings of the 2012 joint conference on empirical methods on natural language processing and computational natural language learning, Korea, pp. 1399-1410.
- Mikolov, T., Kai, C., Greg, C., Jeffery, D., (2013), Efficient Estimation of Word representations in vector space, In Proceedings of workshop at ICLR.
- Bhingardive, S., Singh, D., Murthy, R., Redkar H., Bhattacharya, P., (2015), Unsupervised Most frequent sense detection using word embeddings, Proceedings of the 2015 Conference of the North American Chapter of the Association of Computational Linguistics: Human Language technologies, Denver, Colorado.