International Journal of Application or Innovation in Engineering & Management (IJAIEM)
Web Site: www.ijaiem.org Email: editor@ijaiem.org
Volume 10, Issue 9, September 2021 ISSN 2319 - 4847

Analyzing Word Sense Disambiguation for Swedish language

Shreya Patankar[1], S.R. Devane[2]

[1]Assistant Professor, Datta Meghe College of Engineering, NaviMumbai

[2] Principal, Karmaveer bhaurao Patil college, Nashik

ABSTRACT

This paper presents word sense disambiguation for Swedish language where the research is not much focused on. Sense disambiguation is majorly focused on English language in the area of Natural language processing and various languages remained untouched owing to lack of resources. Now due to advent of multilingual dictionary various languages have started gaining attention. We explore Babel net dictionary to disambiguate senses for Swedish language text by making full use of ontological network. Synonym set and various relations of Babel net are explored and results are presented. It is observed that we achieved promising results and disambiguation for Swedish language opens up new perspective for multilingual word sense disambiguation.

Keywords: word sense disambiguation, Natural language processing

1. INTRODUCTION

Word sense disambiguation (WSD) is the most researched problem in the field of Natural language processing as it involves resolving the ambiguity of text to make it ready for further applications like text summarization, machine translation, etc. There are numerous polysemous words in practically any language and if ignored can hamper the presentation of all semantic NLP errands. Consequently, the undertaking of settling the polysemy and picking the most fitting importance of a word in setting has been an significant NLP task for quite a while. It is generally alluded to as Word Sense Disambiguation. It is complex issue since it includes drawing information from different sources. Huge measure of exertion has been placed into settling this issue in AI since its beginning yet the work is as yet continuous.

Lexical and semantic information from various languages can be an added advantage to performance improvement of WSD but earlier due to lack of resources, the development of the multilingual approach to WSD has been hampered. Invent of the multilingual dictionary has paved the idea of achieving sense disambiguation in a novel way. The combination of lexical and semantic language-independent information for disambiguation makes the understanding of text better and adds more knowledge to the disambiguation process.

Deficiency of resources obstructed the growth of multilingual WSD and it is observed that each language has a different set of ambiguities. Ambiguous words in one language may have the only single sense in other languages and retrieving this sense will boost the sense disambiguation process and making use of resources from the multilingual dictionary is an added advantage. The proposed work combines lexical-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 knowledge. The chapter concludes with observations and findings depicting the comparison between monolingual and multilingual WSD and presents the scores generated which improve the accuracy of multilingual WSD. The chapter concludes that information from various languages improves the accuracy of the system and opens up a new perspective for multilingual WSD.

English is the most researched language and many languages still need attention in research. We explore Swedish language for performing sense disambiguation using Babel Net multilingual dictionary to analyze the behavior of the system. The system relies on knowledge based and unsupervised approach. The results are presented and discussed and experiments have shown that word sense disambiguation performs well for Swedish language.

The paper is organized as follows. Section 2 presents the literature review which highlights the research work of various researchers, section 3 describes the proposed methodology used which includes working with Swedish language and making use of multilingual dictionary Babel net and the working of Word sense disambiguation engine. Section 4 focuses on results and discussions and section 5 sums up with conclusion.

2. LITERATURE REVIEW

The Knowledge-based approach is one of the approaches and fundamental concept of WSD. It makes use of external resources like a dictionary to perform sense disambiguation. Various algorithms which are part of the knowledge-based approach are discussed below. Lesk's algorithm also called as overlap approach [10] uses the dictionary definitions and the correct sense of the word in a given context is determined using the overlap between the definition of target words and the words of the current context. Tested on some short samples and news stories, the average accuracy observed was 50-60%. Banerjee *et al.* in [1] modified the original Lesk algorithm to include a richer knowledge base called Wordnet. They modified the original Lesk algorithm to take advantage of the strongly connected elements and relationship sets among synsets. The overall accuracy observed was 31.7% where 1374 of 4328 test instances were disambiguated correctly. Satisfactory performance of WSD is far from being achieved as this approach is dependent on a dictionary. Overlap based approached includes usage of dictionary for disambiguation and overlap between clue words from the context and dictionary definitions lack strong intersection as dictionary definitions are generally very small which lacks the distributional constraints of different word senses and leads to topic drift due to the wrong overlap.

The concept of selection preference or restrictions is best described by Philip Resnik in [2] and [3]. This approach used large corpora and models the selectional preferences of predicates by combining observed frequencies with knowledge about the semantic classes of their arguments and results achieved were 44% when tested on Brown corpus. The approach needs exhaustive knowledge base for disambiguation and if provided will boost the performance accuracy. Conceptual density approach using wordnet is proposed by Agirre *et al.* in [4] where the concept is to select a sense based on the relatedness of that word sense to the context. Measure of relatedness determines how close the concept represented by the word and concept represented by the context words are. It was tested on sense tagged version of brown corpus and the observed precision was 47.3%. The advantage is that it doesn't require tagged corpus but fails to capture strong clues provided by proper nouns in the context.

Random walks was worked by Agirre *et al.* and discussed in [5] which is a graph-based approach where a vertex is added for each possible sense of the word present in the context. Weighted edges are added to the graph-based on the semantic similarity and using the graph-based ranking algorithm, the score of each sense or vertex is retrieved and the vertex with the highest sense is selected as the winner sense. Weights capture the definition based semantic similarity and satisfactory accuracy was observed. The algorithm is portable, but the only requirement is the lexicon. The structured hierarchical semantic network was used to find the closeness between the words especially nouns to show how close the concept represented by word and concept represented by context words are. Algorithm fails to capture strong clues provided by proper nouns in the context. The approach to disambiguate senses was based on gloss expansion as discussed in [6] where additional information on ambiguous words are included in the context and provide a scoring method to resolve the ambiguity and 80.1% accuracy was observed when applied on TWA corpus.

The various knowledge-based approach discussed above is focussed more on English language and it was observed that very little work is reported on Indian languages to the best of our knowledge and a lot of research is still at the very basic level when it comes to Indian languages. Research has begun for Hindi language but still, more exposure is required. The research papers on Indian language WSD using knowledge-based approach were presented in [7-21]. Languages used were Punjabi and regional language Hindi and all the papers make use of wordnet and it is observed that knowledge-based approach are one of the good approaches to work for WSD as it doesn't require the burden of the corpus. It relies on external resources like wordnet or any lexicon which help in sense disambiguation. The drawbacks of this approach are that proper nouns are not part of the dictionary which results in topic drift and dictionary definitions are very small which results in the minimum overlap and poor accuracy.

The unsupervised approach works with plain raw text and is not dependent on labelled data. This approach divides the occurrence of the specific word into several classes to decide whether the occurrence of the word has the same sense or not. The most important task of this approach is to identify the sense clusters as the training is done on the raw corpus. The research papers on the unsupervised approach used for disambiguation are presented below.

Chaplot *et.al.* in [22] describes an unsupervised approach to WSD using sense dependency and selective dependency. This system was tested on SensEval-2, SensEval-3 mad SemEval 2007 for English all words dataset and this approach had outperformed many state-of-the-art systems. Navigli *et.al.* in [23] describes graph-based methods for unsupervised word sense disambiguation where they proposed a variety of measures that analyze the connectivity of graph structures,

thereby identifying the most relevant word senses. Tested on the semCor and Senseval-3, average accuracy is observed. The most famous algorithm depicting unsupervised approach was proposed by Yarowsky in [24]. The system worked with unlabelled data of English Language and the algorithm is based on constraints - which words tend to have one sense per discourse and one sense per collocation and accuracy crosses 96% for polysemy words. Automatic retrieval and clustering of similar words were discussed in [25] in which the similarity between two words is based on the information content of the single features. Hindi WSD was performed using the unsupervised approach based on the concept of network agglomeration in [26]. A sentence graph is created for a given sentence and from the sentence graph, interpretation graph is generated for each of the interpretation of the sentence. Network agglomeration is computed to identify the desired interpretation and tested on health and tourism corpus; this method yields an average accuracy of around 52%.

Chen *et al.* in [27] explored dependency knowledge for unsupervised approach and argued that existing systems are focussed on specific topics or words and broad coverage is required for WSD system. Dependency knowledge is acquired through parsing and the system was tested on SemEval 2007 dataset and achieved the performance nearly equal to the supervised setting. Nica *et al*. in [28] introduced a novel WSD algorithm to influence the performance of WSD systems in an unsupervised setting. The system was tested on Spanish language and restricted to nouns and tested on Senseval dataset. The method is easily adaptable across other languages provided a corpus of that language is available. Minkov *et al.* in [29] learned graph-based similarity measures for extracting the word synonym from a plain text that resolve the matter by calculating the similarity score which becomes difficult as every context ambiguous word is being connected to every clue word in some or other way. Multiple meanings of an ambiguous word being connected to clue words in the ontological network give a similar score. The machine fails to disambiguate due to similar score and hence does not achieve a satisfactory score for performance.

Agirre *et al.* in [30] discuss WordNet-based distributional similarity approaches on Spanish/English datasets. The authors make use of disambiguated glosses and achieved the best results for wordnet based systems on WordSim353 dataset. Distributional similarity concept was used to cover out of vocabulary words and provided the best results in an unsupervised setting. Various similarity measures are well explored by Meng *et al.* in [31]. The ambiguous word has multiple meanings and the algorithm needs to check all subtrees with different paths meanings to determine the measure and different paths may lead to similar score resulting in poor performance. From the literature review conducted for the unsupervised approach, it is observed that the accuracy of unsupervised systems is lowest as compared to the other two approaches as it works with raw untagged text which results in the formation of clusters. Two clusters may overlap in terms of similarity which may disturb the disambiguation process.

A survey of word sense disambiguation was observed for 158 languages using word embeddings and results were promising[32]. A work on Sense annotated Swedish corpus was carried out by Richard et al., where they developed new annotation tool[33]. SALDO lexicon was used to perform sense disambiguation for Swedish language and their approach showed promising results for nouns rather than verbs[34]. It is observed that not much work is reported on Word sense disambiguation in multilingual setting to the best of our knowledge and it needs to be explored using various state of the art Word sense disambiguation methods. Section 3 discusses the methodology and the working of the system.

3. METHODOLOGY

The philosophy utilizes knowledge based and unsupervised way to deal with sense disambiguation. At first we tried different things with knowledge based way to deal sense distinctions and investigated the ontological structure of Babel Net to decide the relations, for example, parent node, child node, sub class, meronyms , and so on. Synset and gloss in spite of the fact that aides in disambiguation however precision of the framework are at standard the best. The explanation is the definition of synset being too small and shine doesn't give adequate insights for disambiguation which decreases the general score. To fortify the exactness of the framework, different relations are investigated and it is seen that utilizing hypernyms, holonyms and hyponym alongside synset and gloss assists with boosting the framework precision. The dataset utilized is the unlabelled physically made corpus.

3.1 Babel Net

The Babelnet is a huge multilingual ontological network incorporating lexical semantic and syntactic knowledge from various languages [1]. Section 3.3 represents theworking of Word sense disambiguation with the algorithm for the same.

3.2 Working

The figure below represents the knowledge based approach to word sense disambiguation.

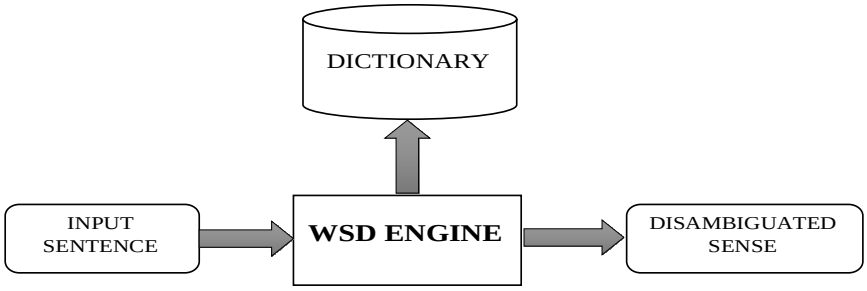

Figure 1 Proposed methodology

The methodology works by accepting input in Swedish language and exploring dictionary synset, gloss and various other relations such as hypernym, holonym and hyponym. The Babel Net displays the score of each value retrieved which helps in disambiguation and works towards improving the accuracy of the system. Various features of senses are explained below.

a) Hypernym:Each listed hypernym is superordinate to this word. This word's referent is (one of) the kind(s) of things each hypernym refers to. For example,placental is a hypernym of bat;
b) Hyponym: Each listed hyponym is subordinate to this word. For example,hen is a hyponym of bird; the bird is a hyponym of avertebrate;
c) Holonym: Each listed holonym has this entry's referent as a part of itself; this entry's referent is part of each listed holonym. For example,forest is a holonym of a tree (forests contain trees); The tree is a holonym of bark; The body is a holonym of an arm which is a holonym of elbow.

The working begins by representing a sentence in terms of a sequence of words μ and the different senses of the ambiguous word *t* are collected in *S* represented as synonym set from the BabelNet. Context words are collected in *Ctx* which includes all the words other than the target word *t*. Context words are the clue words which helps in sense disambiguation. A feature bag is generated which represents the ontological features such as Hypernym, Holonym and Hyponym of context words from the BabelNet. Ontological features of the target word are not considered as synset definition of target word is sufficient for disambiguation and features of target word generates too many scores which complicate the task. Feature bag of context words generates a score for each feature of the sense. The score represents the distance of the target word from the clue word in the ontology represented as a graph The generated score is added to the score of the synset of the target word. This generates a ΔScore and the minimum distance is calculated to represent the correct sense which is the global score returned as the answer.

3.3 Scoring distribution

Scoring distribution is calculated using the Inverse path length sum measure which scores each senseby summing over the inverse length of all paths which connect it to other senses in the graph.

$$\text{scorej} = p\left[\in paths(sj) \frac{1}{e^{length(p)-1}}\right] \sum \qquad (1)$$

where paths(sj) is the set of simple paths connecting sj to the senses of other context words. Length (p) is the number of edges in the path p and each path is scored with the exponentialinverse decay of the path length. Section 4 represents the results and observations and error analysis. The scores received for beach water sense disambiguation are presented below
strand vatten
0
strand
0.1875

0.375
bn:00009263n
An area of sand sloping down to the water of a sea or lake
1
vatten
0.15625
0.3125
bn:00011766n
The part of the earth's surface covered with water (such as a river or lake or ocean)
1
vatten
0.125
0.25
bn:00042379n
Binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent

4. RESULTS AND OBSERVATIONS

The snapshot of lexical sample task used for our experiments is represented below in Table 1and results followed by a discussion for each of the experiments. The goal of our experiments is to initially establish a competitive baseline using unsupervised learning algorithms and the best combination of features that yield the highest accuracy to boost the disambiguation system. WSD framework evaluation is performed on a manually created corpus for Swedish language consisting of polysemous nouns, for the Swedish lexical sample task.

Table 1: Ambiguous words

Sr.No	SWEDISH
1	Ärm (arm)
2	pommes frites (chips)
3	Fladdermus(Bat)
4	Bank
5	strand vatten (water)
6	Kran(Crane)
7	Aska(ash)
8	Stock
9	Pen(penna)
10	Träd(Tree)

Features of Babelnet senses are extracted from the synset(S), the gloss of synset member, hypernymy , hyponymy , synset gloss of Hypernymy-hyponymy relation , holonymy and gloss of holonymy. We tested these features on several instances and results are represented by taking the maximum of the global scores received. It is observed from the table above that combining all the features of BabelNet senses gives us an improved accuracy of 50%. The results after combining various relations are presented below in table 2.

Table 2: Performance analysis using BabelNet features

Features	Global Score	Accuracy in %
Synset	0.0869	20
Gloss	0.1254	22

Synset and Gloss	0.6532	23
Hypernym	0.1112	20
Hyponym	0.5642	30
Holonym	0.6323	30
Hypernym and Holonym	0.2525	35
Holonym and Hyponym	0.4323	38
Hypernym and Hyponym	0.4125	40
Hypernym, Holonym and Hyponym	0.6987	45
Synset, Gloss, Hypernym, Holonym and Hyponym	0.6989	50

To obtain a superior comprehension concerning why the framework neglected to disambiguate in certain cases and how the outcomes could be improved, we played out an itemized investigation of our test sentences. This showed that there are various explanations for the same as the lacking of context clues, relations such as ancestors and predecessors were drifted apart which failed the disambiguation and undetected multi-word articulations. In some cases we didi not receive the relations which gave a reduced score. The conclusions and future scope are represented in section 5.

5. CONCLUSIONS AND FUTURE SCOPE

The ontological structure of the dictionary using the knowledge-based approach also boosts the disambiguation accuracy and observed an overall accuracy of 50%. Limitations of the dictionary-based approach and lack of world knowledge leads to the creation of word and sense embeddings which are useful in NLP for disambiguation task as it gives promising results. The results discussed showed that sense embeddings after sense bag creation helped to improve the disambiguation accuracy and came close to baseline accuracy

References

[1] Banerjee S., Pedersen T.; “An adapted Lesk algorithm for word sense disambiguation using WordNet.”, Computational linguistics and intelligent text processing, Springer Berlin Heidelberg, 2002, pp. 136-145.

[2] Resnik P.; “Selection and information : A class based approach to lexicalrelationships.”, IRCS Technical Reports series, 1993.

[3] Resnik P.; "Using information content to evaluate semantic similarity in a taxonomy." arXiv preprint cmp-lg/9511007, 1995.

[4] Agirre E., Herriko E., Rigau G.; “Word sense disambiguation using conceptual distance.”, COLING 96 Proceedings of the 16th conference on Computational linguistics, Volume 1, pp. 16-22, 1996.

[5] Agirre., Lacella O., Soroa A.; “ Random walks for knowledge based word sense disambiguation”, Association for Computational Linguistics, 2014, pp 57-84.

[6] Fard M., Fakhrahmad S.,Sadreddini M.; “Word sense disambiguation based on gloss expansion”, Conference on Information and knowledge technology(IKT), 2014.

[7] Singh S., Siddiqui T.; "Role of Semantic Relations in Hindi Word Sense Disambiguation." Procedia Computer Science 46 , 2015 pp. 240-248.

[8] Agarwal M., Bajpai J.; "Correlation based Word Sense Disambiguation." Contemporary Computing (IC3), 2014 Seventh International Conference on. IEEE, 2014.

[9] Singh S., Singh V, Siddhiqui T., “Hindi WSD using semantic related measures”, Springer, 2013, pp. 247- 256.

[10] Jain A, Yadav S., Tayal D.; "Measuring context-meaning for open class words in Hindi language." Contemporary Computing (IC3), 2013 Sixth International Conference on. IEEE, 2013.

[11] Singh S., Siddiqui T.; "Evaluating effect of context window size, stemming and stop word removal on Hindi word

sense disambiguation.", Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on. IEEE, 2012.

[12] Sinha M., Kumar M., Pande P., Kashyap L., Bhattacharya P., "Hindi word sense disambiguation." International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India. 2004.

[13] Sawhney R., Kaur A.; "A modified technique for word sense disambiguation using Lesk algorithm in Hindi language", International conference on Advances in Computing, communications and informatics (ICACCI), IEEE, December 2014.

[14] Yadav P., Vishwakarma S.; " Mining association rules based approach for Hindi Language", International Journal of Emerging Technology and advanced Engineering, Volume 3, Issue 5, May 2013.

[15] Bala P.; "Knowledge based approach for word sense disambiguation using Hindi Wordnet", The International Journal of Engineering and Science(IJES), Voume 2, Issue 4, pp 36-41.

[16] Naskar S., Bandopadhyay S.; " Word sense disambiguation using extended Wordnet and Lesk approach", International conference on computing: Theory and applications (ICCTA 2007), IEEE, 2007.

[17] Sharma P., Joshi N.; " Knowledge based Method for word sense disambiguation by using Hindi wordnet", Engineering, Technology and Applied science research, Volume 9, No. 02,2019.

[18] Gautam C., Sharma D.; "Hindi word sense disambiguation using Lesk approach on Bigram and Trigram words", Proceedings of the International conference on Advances in information communication Technology and computing, Aug 2016, pp. 1-5.

[19] Pooja S., Nisheet J.; "Design and development of knowledge-based approach for WSD by using wordnet for Hindi",International Journal of Innovative Technology and Exploring Engineering (IJITEE) January 2019, Volume-8 Issue-3,

[20] Singh J., Singh I.; "Word sense disambiguation: enhanced Lesk approach in Punjabi language", International Journal of Computer, 2015.

[21] Jain S., Jain N., Tammewar A., Bhat R., Sharma D.; " Exploring semantic information in Hindi wordnet for Hindi dependency parsing", International joint conference on Natural language processing, October 2013, pp. 189-197.

[22] ChaplotD., Bhattacharyya P., Paranjape A.; "Unsupervised Word Sense Disambiguation Using Markov Random Field and Dependency Parser." *In AAAI*, pp. 2217-2223, 2015.

[23] [49] Navigli R.,Lapata M.; "Graph Connectivity Measures for Unsupervised Word Sense Disambiguation." *IJCAI*, 2007.

[24] Yarowsky D; "Unsupervised WSD rivalling supervised methods.", *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics, Massachusetts Institute of Technology*, Cambridge, MA., 1995.

[25] Lin D; "Automatic retrieval and clustering of similar words.", *Proceedings of the 17th International conference on Computational linguistics*-Volume 2. Association for Computational Linguistics, 1998.

[26] Jain A.,Lobiyal D.; "Unsupervised Hindi word sense disambiguation based on network agglomeration." *Computing for Sustainable Global Development (INDIACom), 2nd International Conference on. IEEE*, 2015.

[27] Chen P., Ding W., Bowes C., Brown D.; " A fully unsupervised word sense disambiguation method using dependency knowledge", *Human Language technologies: The 2009 Annual conference of the North American Chapter of the ACL*, June 2009, pp. 28-36.

[28] Nica L., Montoyo A., Vazquez S., Marti M.; "An unsupervised WSD algorithm for a NLP system", *International conference on Application of Natural Language to information systems*, 2004, pp. 288-298.

[29] Minkov E., Cohen W.; "Graph based similarity measures for Synonym extraction from parsed text", *Workshop proceedings of Text graph-7, Association for computional Linguistics*, 2012,pp 20-24.

[30] 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*, 2009, pp. 19-27.

[31] MengL., Huang R.,Gu J., 2013, "A review of semantic similarity measures in wordnet", *International Journal of Hybrid Information Technology*,Vol 6, No. 1.

[32] Varvara L., Denis Teslenko , Artem Shelmanov , Steffen Remus , Dmitry Ustalov , Andrey Kutuzov , Ekaterina Artemova , Chris Biemann , Simone Paolo Ponzetto , Alexander Panchenko; "Word Sense Disambiguation for 158 Languages using Word Embeddings Only", Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 5943–5952 Marseille, 11–16 May 2020.

[33] Richard Johansson, Yvonne Adesam , Gerlof Bouma , Karin Hedberg; "A Multi-domain Corpus of Swedish Word Sense Annotation", ACL anthology

[34] Ildikó Pilán; " Helping Swedish words come to their senses: word sense disambiguation based on sense associations from SALDO lexicon", Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)