Techniques for Disambiguation of Polysemy Words: A Review

Vandita Singh
Department of Computer Science
JIMS Engineering Management Technical Campus
Greater Noida, India

Krishan Kr. Saraswat
Department of Computer Science
JIMS Engineering Management Technical Campus
Greater Noida, India

ABSTRACT
The domain of Computational Linguistics involves the key task of Word Sense Disambiguation which aims to assign a meaning to particular word in terms of the context with which it is used in a sentence. The task of assigning the semantically correct meaning to a polysemy word in almost all the languages of the world stands out to be an open problem of research with considerably low accuracies achieved. The paper presents a meticulous review of the various techniques opted for disambiguation of polysemy words in various languages -English, Hindi, Nepalese, Tamil, Kannada, Telugu, Malayalam, Sinhala and German. Also, an insight into how the various approaches -supervised (involving corpora) and Unsupervised (clustering, meta thesaurus) to solving the above problems evolved over the years to get the accuracy improved. The applications include word processing, spell checking, content analysis, translation, improved search engines.

Keywords
Natural Language Processing, Word Sense Disambiguation, WordNet, Polysemy Words

1. INTRODUCTION
The field of Natural Language Processing deals with the task of disambiguating polysemous words which stands out to be an open problem in this particular field of research. Words have different meanings based on the context of the word usage in a sentence. Word sense is one of the meanings of a word. Most words have many possible meanings which is referred to as Polysemy, thus words can be interpreted in many possible ways depending on their context in a sentence. A computer program has no basis for knowing which one is appropriate, even if it is obvious to a human. *Word Sense Disambiguation* is the problem of selecting a sense for a word from a set of predefined possibilities.

Examples of ambiguity are – Sentence 1- The fisherman jumped off the bank and into the water; Sentence 2- The bank down the street was robbed.; Sentence 3 - Back in the day, we had an entire bank of computers devoted to this problem. The word bank has been used in different contexts in the above three sentences and thus is termed as a polysemy word which needs to semantically disambiguated while processing the natural language.

2. PROCESS OF WORD SENSE DISAMBIGUATION AND TERMINOLOGIES
2.1. Process of Word Sense Disambiguation
The resolution of various kinds of syntactic ambiguity can be resolved using parts of speech taggers with appreciably good levels of accuracy but the case of semantic ambiguity found in natural language processing can only be resolved using the task of word sense disambiguation. Usually the task of Word sense disambiguation involves two major steps, the first being the determination of all the different senses for every word in the text and secondly the assignment of each occurrence of a word to the appropriate sense either using the context of the ambiguous word or the external knowledge sources. The simplified process can be put up diagrammatically in Figure 1.

Princeton University developed WordNet which is a lexical database developed at for English Language and organizes various parts of speech-nouns, verbs, adjectives and adverbs into sets of synonyms called synsets. WordNet describes the relationships between these groups thus forming a semantic network among the words. Over the years lexical databases for other languages have also been developed such as Hindi WordNet [7], GermaNet[20], WordNet for Nepalese[10] ,few of which have been discussed in this paper as well.

Few terms have been used often while discussing the works of Word Sense Disambiguation in various languages such as Target Word, Context Word, Hypernymy [12]. Target Word is that word which has multiple meanings at different contexts and has to be disambiguated. Context words are those used in context with the target word. Hypernymy is semantic relation between two synsets to show super-set hood.

3. EVOLUTION OF TECHNIQUES FOR DISAMBIGUATION OF POLYSEMY WORDS IN VARIOUS LANGUAGES
The paper presents review of the techniques that evolved over the years for the task of Word Sense Disambiguation to assign semantically relevant meaning to polysemy words in various languages-English, Hindi, Nepalese, Dravidian Languages-Tamil, Telugu, Kannada, Malayalam, Sinhala and German and the respective lexical databases that have been developed for the processing of information in these particular languages.

3.1 English
Dealing with a class of problem involving disambiguation of words started with works based on Lesk Algorithm developed by Lesk Micheal[1] in the year 1986 for English Language to identify senses of polysemy words using Overlap of word definition from the Oxford Advanced Learners Dictionary of Current English. The approach given for the Lesk algorithm can be simply put up as follows. The gloss of each of its senses is compared to the glosses of every other word in the window of words. The first step is to obtain the glosses of the senses of the target word, followed by comparison of the gloss of each sense of the target word with the glosses of every other word in the given window of words, and keeping a count of the overlapping words in each sense pair. Finally, the one with highest count of overlaps will be the most

appropriate sense. The accuracy achieved was around 50 % for very short word samples.

In the year 2002,Banerjee and Pederson[2] developed a technique by making an adaptation to the Lesk Algorithm using Lexical database WordNet SenseVal2 for Word Sense Disambiguation .The purpose was to identify senses of polysemy words by finding the longest sequence of words occurring in two glosses where the final count was taken as sum of all overlaps ,here taken as the square of the number of words in the overlap.The accuracy achieved was 32%.Eneko Agirre and German Rigau[3] adapted the Lesk Algorithm to make a comparison of glosses alongside looking for occurrence of nearby word's senses in hierarchy of target word's senses.

W .J. Lee and E.Mit[4] suggested a bag of words approach which was a combination of supervised and unsupervised technique; where the domain-term distribution in the given text was identified and sense of the word was defined. The accuracy achieved was up to 70%.[4]Previously all the approaches adopted were mostly including a measure of the information content of the words and the amount of overlapping. Sayali Charhate, Anurag Dani and Rekha Sugandhi[5]discuss an unsupervised approach where more emphasis has been put over adding intelligence to the mere gloss overlap predicted by intersection score such as Path distance. This was done by considering various WordNet semantic relations and auto-filtration of content words before semantic graph generation.[5]

P . Sachdeva , S.V erma and S.K.Singh present an integrated approach in order to rule out the polysemy words based on the combined impact of three parameters - Intersection, Hierarchical level and distance[6].In addition, the number of common words along the entire hierarchy of the target and nearby words' senses are found out and the algorithm also takes into account the factor of distance, which is the distance between the target word and the nearby word in the input text.[6] The algorithm achieves a precision of 53.12% , 59.91% and 62.13% respectively for Top1 , Top2 and Top3 results which as stated by the author is comparatively better than other knowledge based approaches.[6]

3.2 Hindi

The first attempt on automatic word sense disambiguation system for Hindi was made by Sinha, Kashyap, Bhattacharyya, Pandey, and K. Reddy[7].They developed a statistical method for Hindi word sense disambiguation with a rule based algorithm [7]. The main idea behind their research was to compare the context of the word in a sentence with the contexts constructed from the Hindi WordNet and choose the winner. For assigning senses to words in Hindi, with the use of the context in which it has been mentioned, the information in the Hindi WordNet and the overlap between these two pieces of information was calculated and the sense with the maximum overlap was declared as the winner. The Hindi corpora from the Central Institute of Indian Languages(CIIL), Mysore was used as the test data source for sense disambiguation and their work is limited to nouns only. The accuracy varies from 40-70%.

S.Vishwakarma and C.Vishwakarma[8] have reviewed techniques for techniques for Word Sense Disambiguation in Hindi and discussed about a graph based approach for word sense disambiguation for Hindi Language.Sharma[9] applied knowledge based ,machine learning based and other hybrid approaches to develop word sense disambiguation for Hindi language using the Hindi WordNet developed by Sinha, Kashyap, Bhattacharyya, Pandey, and K. Reddy[9].

3.3 Nepalese Language

Shreshta N. et al. [10] has implemented the Lesk Algorithm to disambiguate the polysemy words in Nepali language. The Lesk algorithm was modified in such a way that context words did not include synset, gloss, example and hypernym and also number of example for each sense of the target word was taken as only one.

Later, Dhungana and Shakya [11] further adapted the Lesk Algorithm for disambiguation of Nepali words comprising a total of 348 words inclusive of 59 polysemy words along with the context words. The test data comprised of 201 Nepali words and the final accuracy of the system was 88.05% which was an increase of 16.41% in comparison to the work of Shreshta N. et al.

Further in 2015, U. R. Dhungana, S. Shakya, K. Bara and B. Sharma [12] performed experiment on same experimental setting as of Dhungana and Shakya [11] to disambiguate polysemy words and included synset, gloss, example and hypernym and number of examples for each sense of the context words. Accuracy 88.059%. They also adapted the Word Net to include Clue Words where grouping was done for each sense of polysemous word based on the verb, noun, adverb and adjective with which the sense of the polysemy word can be used in a sentence. Accuracy 91.543%.

3.4 Dravidian Languages

3.4.1 Tamil

Bhaskaran S and Vaidehi[13] present an unsupervised approach of clustering to group the occurrence of an ambiguous word in a trained corpus. The three types of ambiguities being -Polysemy, Homonymy and categorical ambiguity [13]. High accuracy was achieved using an efficient and large corpus for the collocation based Tamil Word Sense Disambiguation System using clustering

3.4.2 Telugu

Ch. Mandakini and K.V.N Sunitha[14] aim at disambiguating sense of a word using argument structure in Telugu sentences containing a single verb. The work focuses on describing the argument structure of the verb in context where argument is the main element required by the predicate in a Telugu sentence.

3.4.3 Kannada

This Kannada Word Sense Disambiguation System is based on the usage of a big corpora, nearly of the size of 5 million words which takes randomly selected sentences from the corpora [15], hence resulting into higher accuracy of the system.

3.4.4 Malayalam

The very first work done in Malayalam Language for Word Sense Disambiguation by R.P. Haroon[16] was using Lesk Algorithm and Conceptual Density. The work produces satisfactory results with limitations being the lack of a good corpus. Later a supervised Malayalam Word Sense Disambiguation system using Naïve Bayes Classifier was implemented by Sreelakshmi Gopal and R.P. Haroon[17] which provides 95% reliability using a corpora of over 1 lakh words.

3.5 Sinhala

Sinhala ,derived from old Indo Aryan Sanskrit through middle Indo Aryan Prakrit ;is the main language of Sri Lanka spoken by over 19 million people. Despite of the fact that it is spoken by over 19 million people and is one of the official languages of Sri Lanka, there has been very limited research on computational linguistic of Sinhala. The very first attempt to find a technique for Sinhala Word Sense Disambiguation was using a WordNet and rule based algorithm implemented by Arukgoda, Bandara, Bashani, Gamage and Wimalasuriya[18] to gain an F1 Score of 0.63 for Sinhala.

Prior to this work on Sinhala Language, previous study was carried out by A. Marasinghe, S. Herath, and A. Herath[19].They proposed a method to disambiguate Sinhala words based on an unsupervised leaning technique with the use of Susantha- Corpus [19].The two unsupervised learning methods - EM algorithm and Gibbs sampling were used to show the result of disambiguation of five words. The system could only disambiguate up to 5 nouns as the number of ambiguous target words used in this study was very small.

Table 1. Evolution of Word Sense Disambiguation techniques for English and Hindi Languages

S.No.	Languages : English(1-6) and Hindi(7-9) In-between Bottom	
	Author	**Approach**
1.	Michael Lesk[1]	Lesk Algorithm
2.	Eneko Agirre & German Rigau[3]	Adapted Lesk Algorithm
3.	Banerjee & Pederson[2]	Adpated Lesk Algorithm
4.	Wie Jan Lee and Edwin Mit	Bag of Words Approach
5.	Sayali Charhate, Anurag Dani and Rekha Sugandhi[5]	Adding intelligence to path distance(semantic relations and auto-filtration)
6.	Sachdeva P,Verma S,SinghS.K[6]	Hybrid Approach(intersection between word families+ hierarchical relationship + distance)
7.	M. Sinha, M. K. Reddy, P. Bhattacharyya, P. Pandey, and L. Kashyap[7]	Statistical Method -Rule Based Method
8.	S. Vishwakarma, and C. Vishwakarma[8]	Graph Based Approach
9.	R. Sharma[9]	HybridApproach(Knowledge Based + Machine Learning Approach)

3.6 German

Review work conducted by Verena Henrich and Erhard Hinrichs presents a wide range of algorithms implemented for word sense disambiguation in German Language. The major approaches discussed are Semantic relatedness measures such as path based , information content based and gloss based methods. The best results for German language Word Sense Disambiguation were obtained using a word overlap method derived from the Lesk Algorithm which uses Wiktionary glosses and GermaNet (German WordNet).GermaNet and WebCAGe corpus have been utilized as Sense inventory for German.[20]

Broscheit, Frank, Jehle, Ponzetto, Rehl, Summa, Suttner and Vola[21] developed Word Sense Disambiguation resources for German using GermaNet as a basis. Following previous unsupervised methods, predominant sense information was acquired and used as type based first sense heuristics for token level Word Sense Disambiguation.[21] The state of the art knowledge based Word Sense Disambiguation system was adapted to the GermaNet lexical resource. The study was conducted to investigate the hypothesis whether the two systems are complimentary by combining their output.

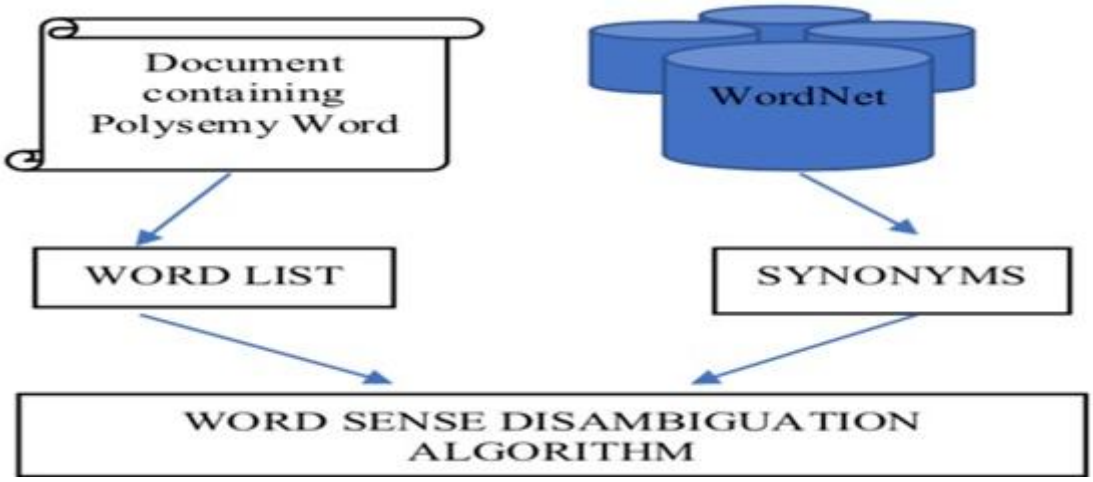

Fig 1: Process of Word Sense Disambiguation

4. CONCLUSION

The intent of covering these techniques was to have an overview of the existing word sense disambiguation techniques and perform a comparative analysis of various such techniques in relation with the natural language being taken into consideration. We can easily comprehend from the survey shown in Table 1 for English and Hindi Languages, that task of Word Sense Disambiguation dates back with advent of Lesk Algorithm primarily applied to the most popular language around the globe i.e. English and further adapted to other languages such as Malayalam, German and Nepalese.

For English the techniques evolved over time starting from information content based and knowledge based techniques(rule based algorithms) to graph based and machine learning based techniques. The application of this language is most widespread for tasks such as feature level sentiment analysis, Biomedical document disambiguation and various other natural language processing tasks. The task of Word Sense Disambiguation for Hindi Language started with the statistical approach using a rule based algorithm and state of the art techniques incorporate graph based algorithms for Hindi as well.

The techniques in Tamil language relied on a much newer technique of unsupervised approach called as clustering, yielding efficient results. The Telugu Language Word Sense Disambiguation task based itself over building a large sized word corpora and thus improving the accuracy rates. The Malayalam word sense disambiguation task was initiated by the implementation of Lesk Algorithm and later was put to improved accuracies using a machine learning based approach using Naïve Bayes Classifier. For Sinhala language an unsupervised approach usingGibbs Sampling and EM Algorithm[18] was opted as a firstattempt to solve the problem of disambiguation of polysemy words but later the accuracies largely improved for thelanguage by adopting a rule based algorithm. Also fordifferent languages different WordNet was developed to suit IEEE the requirements of the structure of the given natural language. Having an insight into these techniques wouldprovide a way to devise systems in future which are efficientin terms of time and storage i.e. gives higher accuracy with

considerable size of the corpora or employs some unsupervised approach of learning. Also the coverage ofother languages around the globe may be considered.

5. REFERENCES

[1] M. Lesk, "Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone," in Proc.5th annual international conference on Systems documentation, New York, USA, 1986, pp. 24 – 26. J. Clerk Maxwell, A Treatise on Electricity and Magnetism, 3rd ed., vol. 2. Oxford: Clarendon, 1892, pp.68–73.

[2] S. Banerjee and T. Pedersen, "An adapted lesk algorithm for word sense disambiguation using wordnet," in Third Interna tional Conference on Intelligent Text Processing and Computational Linguistics, Gelbukh, 2002.

[3] Eneko Agirre and German Rigau, "Word Sense Disambiguation using conceptual density," Proceedings of the 16th conference on Computational linguistics - Volume 1, 1996

[4] Wei Jan Lee and Edwin Mit, "Word Sense Disambiguation By Using Domain Knowledge", International Conference on Semantic Technology and Information Retrieval 28-29 June 2011, Putrajaya, Malaysia, 978-1-61284-353-7/11/$26.00 ©2011 IEEE

[5] Sayali Charhate, Anurag Dani and Rekha Sugandhi, "Adding Intelligence to Non-corpus based Word Sense Disambiguation", *2012 12th International Conference on Hybrid Intelligent Systems (HIS)* 978-1-4673-5116-4/12/$31.00 c 2012 IEEE

[6] P. Sachdeva , S.Verma and S.K.Singh, " An Improved Approach to Word Sense Disambiguation", 978-1-4799-1812-6/14/$31.00 ©2014

[7] M. Sinha, M. K. Reddy, P. Bhattacharyya, P. Pandey, and L. Kashyap, "Hindi word sense disambiguation," Master's thesis, Indian Institute of Technology Bombay,Mumbai, India, 2004

[8] S. Vishwakarma, and C. Vishwakarma, "A graph based approach to word sense disambiguation for Hindi language," India, International Journal of Scientific Research Engineering & Technology, vol. 1, pp. 313-318, August 2012.

[9] R. Sharma, "Word Sense Disambiguation for Hindi Language," Diss. Tharpar University, India, 2008.

[10] N. Shrestha, A. V. H. Patrick, and S. K. Bista, "Resources for nepali word sense disambiguation," in IEEE International conference on Natural Language Processing and Knowledge Engineering (IEEE NLP-KE'08), Beijing, China, 2008.

[11] U. R. Dhungana and S. Shakya, "Word sense disambiguation in nepali language," in The Fourth International Conference on Digital Information and Communication Technology and Its Application (DICTAP2014), Bangkok, Thailand, 20l4, pp. 46-50.

[12] Udaya Raj Dhungana1, Subarna Shakya, Kabita Bara and Bharat Sharma , in IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015) 2015, Anaheim, California, USA 978- 1-4799-7935-6/15

[13] Bhaskaran S and Vaidehi , "Collocation Based Word Sense Disambiguation using clustering for Tamil", V, K U Research Center,2003.

[14] Ch Mandakini and Dr K V N Sunitha, "Disambiguating the sense of verb in Telugu sentence using the argument structure", International Journal of Computational Linguistics and Natural Language Processing Vol 1 Issue 5 December 2012

[15] S Parameswarappa and V N Narayana,, "Kannada Word Sense Disambiguation for Machine Translation," International Journal of Computer Applications V olume 34– No.10, November 2011.

[16] R P Haroon "Malayalm Word Sense Disambiguation" Computational intelligence and computing research (ICCIC), IEEE,2010, E- ISBN:978- 1-4244-5967-4.

[17] Sreelakshmi Gopal and Rosna P Haroon, " Malayalam Word Sense Disambiguation using Naïve Bayes Classifier", International Conference on Advances in Human Machine Interaction (HMI - 2016), March 03-05, 2016, R. L. Jalappa Institute of Technology,

Doddaballapur, Bangalore, India , 978-1-4673-8810-8/16/$31.00 ©2016 IEEE

[18] C. Marasinghe, S. Herath, and A. Herath, "Word sense disambiguation of Sinhala language with unsupervised learning,"in Proc. International Conference on Information Technology and Applications, Bathurst, Australia, November 2002, pp. 25-29

[19] J. Arukgoda ,V. Bandara,S.Bashani,V.Gamage and D.Wimalasuriya, "A Word Sense Disambiguation Technique for Sinhala" , in 4th International Conference on Artificial Intelligence with Applications in Engineering and Technology , 978-1-4799-7910-3/14 $31.00 © 2014 IEEE DOI 10.1109/ICAIET.2014.42

[20] V. Henrich, and E. Hinrichs "A comparative evaluation of word sense disambiguation algorithms for German," in Proc. LREC'12,Istanbul, Turkey, 2012, pp. 576-583.

[21] S. Broscheit, A. Frank, D. Jehle, S. Ponzetto, D. Rehl, A. Summa, K.Suttner, and S. Vola, "Rapid bootstrapping of word sense disambiguation re-sources for German,"inProc.10th Konferenz zur Verarbeitung Natürlicher Sprache, Germany, 2010 pp. 19–27.