DIFFERENT TECHNIQUES IMPLEMENTED IN GURUMUKHI WORD SENSE DISAMBIGUATION

Himdweep Walia[1], Ajay Rana[2], Vineet Kansal[3]

[1]Department of Computer Science & Engineering, Greater Noida Institute of Technology (India)

[2]Amity Technical Placement Centre, Amity University, Uttar Pradesh (India)

[3]Department of Computer Science & Engineering, I.T.S. Engineering College (India)

ABSTRACT

One of the most important issues in the field of Natural Language Engineering is Word Sense Disambiguation (WSD).Gurumukhi or more commonly known as Punjabi, is world's 12th most widely spoken language and this language is morphologically rich. But surprisingly, there are relatively less efforts in the field of computerization and development of lexical resources of this language. It is therefore motivating to develop a corpus of Punjabi Language that will help in tagging the sense of the words.The availability of sense tagged corpora contribute a lot in advances in WSD. Most accurate WSD systems use supervised learning algorithm to learn contextual rules or classification models automatically from sense-annotated examples, like Naïve Bayes, k-NN and Support Vector Machine (SVM) classifiers have shown high accuracy in WSD. The majority of work on WSD is focused on English and other European languages and standard test corpora are available for these languages. The lack of such standards put a major hindrance on WSD research for Punjabi and other Regional Indian languages. Thus, this defines the objective of this survey.

Keywords : Word Sense Disambiguation(WSD), WSD in Punjabi, Knowledge- Methods, Supervised disambiguation

I INTRODUCTION

Ambiguity is the quality of being open to more than one interpretation. All the Natural Languages known to us are full of ambiguous words i.e. the words have more than one meaning and this can be understood in the context in which they are being used. But as humans know in which context the particular word is being used, they understand the zest of it. Consider the following example:

i. The yogi is praying near the bank.

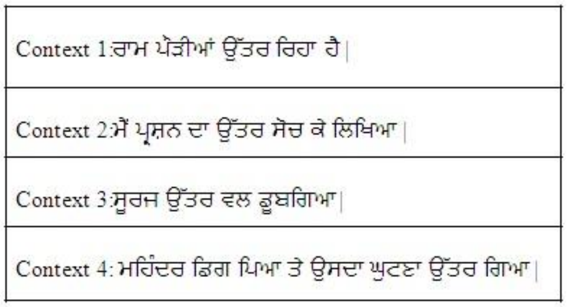
ii. The bank is closed on Sunday.

In the above sentences, the word “bank” is common in both the sentences.For humans, it is easy to comprehend that the word “bank” in the first sentence refers to the river bank, whereas in the second sentence it refers to institution that offers services like safekeeping and lending of money.

This is not the case when the interpretation is to be made by a machine. The machine will only be able to understand the right context if it has a set of predetermined rules or meaning or set of meanings associated with that word.

In a similar fashion, all the languages have this sort of ambiguity which is difficult for the machine to detect easily. In our language of study, Gurumukhi, popularly known as Punjabi, we too have many ambiguous words. Consider the word, **ਉੱਤਰ**.

Table-1 Examples of ambiguous word“ਉੱਤਰ”



Context 1:ਰਾਮ ਪੌੜੀਆਂ ਉੱਤਰ ਰਿਹਾ ਹੈ|
Context 2:ਮੈਂ ਪ੍ਰਸ਼ਨ ਦਾ ਉੱਤਰ ਸੋਚ ਕੇ ਲਿਖਿਆ|
Context 3:ਸੂਰਜ ਉੱਤਰ ਵਲ ਡੂਬਗਿਆ|
Context 4: ਮਹਿੰਦਰ ਡਿਗ ਪਿਆ ਤੇ ਉਸਦਾ ਘੁਟਣਾ ਉੱਤਰ ਗਿਆ|

The word “**ਉੱਤਰ**”, is common in all the four sentences and convey different meanings in all of the above. It is easier for a human to differentiate that in which context the word is being used. For example, in the first sentence the word means climbing down, in the second sentence it means answer, in the third sentence it means direction and in the last sentence it means dislocation of bone.

The study of Word Sense Disambiguation (WSD) has become a pivotal point in Natural Language Processing (NLP). In order for a machine to interact i.e. understand, interpret and speak, to humans, the machine needs precise, unambiguous and highly-structured set of rules. And for this WSD is imperative and so is its study to enable machines to use natural languages.

II SEARCH METHODS

This section explains in detail the procedure that was followed in order to study the various techniques that have been used in Gurumukhi Word Sense Disambiguation.

2.1 Search questions

In this review we have tried to analyze the work that has been done in Punjabi Language for Word Sense Disambiguation [1, 2]. The following are the research questions which were kept in mind while going through the various papers on the given subject:

i. Which all techniques exist in WSD?
ii. Which all techniques have been used in Punjabi WSD?
iii. Which all techniques can be implemented to give good results for Punjabi WSD?
iv. Can combining of techniques yield better results?

2.2 Sources of information

Keeping in view the recommendations enlisted in Kitchenham's guidelines [3], the search was carried out via electronic sources. The following are the list of databases used for the same:

i. ACM Digital Library (www.dl.acm.org)
ii. Springer (www.springer.com)
iii. IEEE Xplore (www.ieeexplore.ieee.org)
iv. Academia (www.academia.edu)
v. CiteSeerX (www.citeseerx.ist.psu.edu)
vi. ResearchGate (www.researchgate.net)
vii. Association for the Advancement of Artificial Intelligence (www.aaai.org)
viii. Association for Computational Linguistics (www.aclweb.org)

Word Sense Disambiguation has generated a huge amount of response from researchers and so the electronic databases are in abundance with the papers available on this topic. In order to streamline the search, the databases were limited to those mentioned above.

2.3 Search criteria

Initial search criteria were broad in order to include articles with different use of terminology. The keywords used were <WSD techniques>, <Knowledge-based PunjabiWSD>, <Supervised PunjabiWSD>, <Unsupervised Punjabi WSD>. The start year was set to 2001as before that not much work has been done related to Punjabi language in the domain of Word Sense Disambiguation.

2.4 Paper selection

The papers were screened through a three-stage process. In the first stage the papers were shortlisted based on titles. In our case, the number of irrelevant papers is very large as the topic "word sense disambiguation" is a new topic and a very active area of research. A number of techniques are implemented in various languages like English, European languages, Spanish, Dutch, Indian regional languages, like Hindi, Punjabi, Bengali [4, 5], to name a few. In the second stage, the information in abstracts were analyzed and the papers were classified based on Knowledge-based and Corpus-based techniques used in English and Indian regional languages.

In the third stage, the papers were selected based on full text. And by following this procedure, we shortlisted 23 papers to provide the ground work required for research in the area of WSD techniques used in Punjabi.

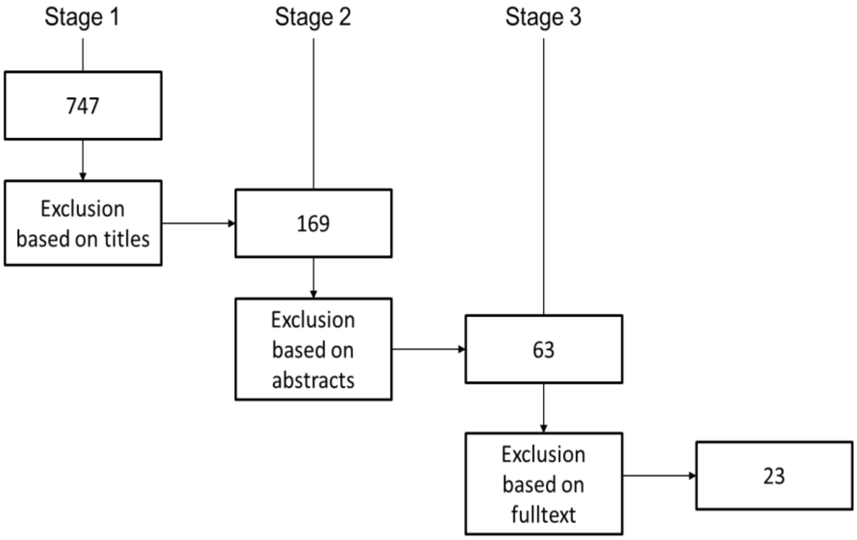

Stage 1
Stage 2
Stage 3
747
Exclusion based on titles
169
Exclusion based on abstracts
63
Exclusion based on fulltext
23

Figure 1: Paper selection procedure[6]

III RESULTS AND DISCUSSIONS

The data set or knowledge bank is crucial for WSD. We can broadly categorized it in two – structured resources [7] (which include thesauri, machine-readable dictionaries and ontologies)[8, 9] and unstructured resources (which includes corpora – raw and sense-annotated, collocation resources) [7].

The Word Sense Disambiguation (WSD) has three conventional approachesas discussed below:

3.1 Knowledge-based Disambiguation

These rely primarily on dictionaries, thesauri, and lexical knowledge bases to infer the senses of the words in context. A simple approach used here is that of overlap of sense definition, more popularly known as Lesk Algorithm. The Lesk algorithm[10] is based on the hypothesis that given a two-word context, the senses of the target word whose definition has the highest overlap is assumed to be the correct[11].

A significant amount of work has been done in Punjabi using Lesk Algorithm [12, 13, 14]. Enhancements in the base algorithm have been made and the results are very promising.

3.2 Supervised Disambiguation

These use machine-learning techniques for inducing a classifier from manually sense-annotated data sets. Some of the popular supervised techniques are decision list, decision trees, Naïve Bayes [15, 16, 17], neural networks, instance-based learning and Support Vector Machines (SVM)[18, 19]. Although Naïve Bayes and SVM are very popular techniques and have implemented for various languages giving excellent results but no work has been done in Punjabi Language.

3.3 Unsupervised Disambiguation

These approaches are "based on the idea that the same sense of a word will have similar neighboring words. They are able to induce word senses from input text by clustering word occurrences and then classifying new occurrences into the induced clusters" [7].The techniques which come under this are content clustering, word clustering and co-occurrence graphs. These techniques are being used for English language [20] but again no work is done in Punjabi language.

Table-2 Number of papers using a given approach for Punjabi WSD

Approach	**#**
Lesk Algorithm	5
Walker Algorithm	1
Naïve Bayes	8
SVM	2
Instance-based learning	2
Unsupervised	1
Combination of approaches	5

IV CONCLUSIONS

The survey has revealed that although a lot of work has been done in digitally advanced languages, such as English, but relatively less work has been done in Indian Languages, and one such language is Punjabi. The two important approaches used in WSD are knowledge-based disambiguation and supervised disambiguation. Algorithms such as Lesk's Algorithm, Walker's algorithm, etc have been popular among researchers under knowledge-based approach. A considerable amount of work has been done in Punjabi language as well using Lesk Algorithm. The knowledge set used in this case is also limited to structured resources. Not much work has been done in Punjabi using unstructured resources i.e. corpora. The primarily reason being lack of availability of sense-tagged corpora.

Also relatively less work has been done on supervised disambiguation for Punjabi Languages. Algorithms such as Naïve Bayes, n-grams, SVM, etc can be thus used which would yield better results for Punjabi Language and will help in building the corpora.

Combining of techniques has yielded brilliant results in other languages [21], so if similar conceptswill be implemented for Punjabi language as well then results will be better too.

V ACKNOWLEDGEMENTS

The author acknowledgesthe Management of Greater Noida Institute of Technology, Gr. Noida who encouraged to pursue doctorate and have provided all the necessary help for the same.

The author also acknowledges Amity University, Uttar Pradesh, who has given opportunity to pursue doctorate and has provided library facility and extensive online resources.

REFERENCES

[1] R. Kumar, R. Khanna, Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi, IJES (ISSN: 2229-6913), July, 2011.

[2] R. Kumar, R. Khanna, and V. Goyal, A Review of Literature on Word Sense Disambiguation, International Journal of Engineering Sciences, ISSN: 2229-6913, Vol. 6, July, 2012.

[3] B. Kitchenham, S. und Charters, Guidelines for performing Systematic Literature Reviews in Software Engineering, EBSE 2007-001, Keele University and Durham University Joint Report, 2007.

[4] M. Bansal, Word Sense Disambiguation: Literature Survey for Indian Languages, in International Journal of Advanced Research in Computer Science and Software Engineering (ISSN: 2277 128X), Volume 5, Issue 12, December 2015.

[5] M.M. Khapra, P. Bhattacharyya, S. Chauhan, S. Nair, A. Sharma, Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting, available at ResearchGate (https://www.researchgate.net/publication/228981276), last visited 2013.

[6] E. Engstrom, M. Skoglund, P. Runeson, Empirical Evaluations of regression Test Selection Techniques: A Systematic Review, ESEM'08, October, 2008, ACM.

[7] R. Navigli, Word Sense Disambiguation: A Survey, ACM Computing Surveys, Vol. 41, no.2, Article 10, February, 2009.

[8] M. M. Khapra, P. Bhattacharyya, S. Chauhan, S. Nair, A. Sharma, Domain Specific Iterative Word Sense Disambiguation in a Multilingual Setting, available at ResearchGate (https://www.researchgate.net/publication/228981276), last visited 2013.

[9] R. Kaur, R. K. Sharma, S. Preet, P. Kumar, Punjabi WordNet Relations and Categorization of Synsets, available at (http://www.cfilt.iitb.ac.in/wordnet/webhwn/IndoWordnetPapers/12_iwn_Punjabi%20WordNet%20Relations%20and%20Categorization%20of%20Synsets.pdf), last visited 2016.

[10] S. Banerjee, T. Pedersen, An adapted Lesk Algorithm for Word Sense Disambiguation using WordNet, in Proceedings of the 3rd International Conference on Computational Linguistics and Intelligent Text Processing, Mexico City, Mexico, pp.136-145, 2002.

[11] F. Vasilescu, P. Langlais, G. Lapalme, Evaluating Variants of the Lesk Approach for Disambiguating Words, available at www.iro.umontreal.ca/~felipe/Papers/paper-lrec-2004.pdf, last visited 2012.

[12] J. Kaur, A. Singh, A Customized Adaptation of Traditional Lesk Method for Sense Disambiguation of Punjabi Words from Medical Domain, IJRASET (ISSN: 2321-9653), Vol-3, Issue-V, May 2015.

[13] P. Rana, P. Kumar, Word Sense Disambiguation for Punjabi Language using Overlap Based Approach, Chapter in Advances in Intelligent Informatics, Volume 320 of the series Advances in Intelligent Systems and Computing, pp 607-619, Springer.

[14] J. Singh, I. Singh, Word Sense Disambiguation: Enhanced Lesk Approach in Punjabi Language, In International Journal of Computer Applications (0975-8887), Volume 129 – No.6, November 2015.

[15] D. Jurafsky, J. H. Martin, Naive Bayes Classifier Approach to Word Sense Disambiguation, Chapter 20 Computational Lexical Semantics, available at (http://www.let.rug.nl/nerbonne/teach/rema-stats-meth-seminar/presentations/Olango-Naive-Bayes-2009.pdf), last seen 2015.

[16] C. A. Le, A. Shimazu, High WSD accuracy using Naive Bayesian classifier with rich features, in Proceedings of PACLIC 18, Tokyo, Japan, pp.105-113, 2004.

[17] S. Singh, T. J. Siddiqui, S. K. Sharma, Naïve Bayes classifier for Hindi Word Sense Disambiguation, In Proceedings of 7th ACM India Compute Conference (Compute'14), Nagpur, India, 9 – 11 October, 2014, Article No. 1, ACM Digital Library.

[18] S. Garg, A. K. Mittal,A Performance of SVM with Modified Lesk Approach for Word Sense Disambiguation in Hindi Language, International Journal of Research in Engineering and Technology (IJRET) eISSN: 2319-1163, Volume: 04, Issue: 08, August, 2015.

[19] Y. K. Lee, H. T. Ng, T. K. Chia, Supervised Word Sense Disambiguation with Support Vector Machines and Multiple Knowledge Sources, in the 3rd International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, July 2004.

[20] T. Pedersen, Unsupervised Corpus-based Methods for WSD, chapter, Word Sense Disambiguation, Volume 33 of the series Text, Speech and Language, pp 133-166, Springer.

[21] G. S. Josan, G. S. Lehal, Size of N for Word Sense Disambiguation using N gram model for Punjabi Language, available at http://learnpunjabi.org/pdf/gslehal-pap19.pdf, last visited 2015.

[22] A. Montoyo, A. Suarez, G. Rigau, M. Palomar, Combining Knowledge and Corpus-based Word-Sense Disambiguation Methods, Journal of Artificial Intelligence Research, March, 2005.

[23] J. Kaur, A. Singh, A Customized Adaptation of Traditional Lesk Method for Sense Disambiguation of Punjabi Words from Medical Domain, IJRASET (ISSN: 2321-9653), Vol-3, Issue-V, May 2015.

[24] A. Narang, R.K.Sharma, P. Kumar, Development of Punjabi WordNet, Springer, CSIT, December 2013.

[25] S. Singh, T. J. Siddiqui,Role of Semantic Relations in Hindi Word Sense Disambiguation, In Proceedings of the International Conference on Information and Communication Technologies (ICICT 2014), Kochi, India, 3-5 December, 2014, Elsevier Procedia Computer Science, Volume 46, 2015, pages 240-248.

[26] S. Singh, V. K. Singh, T. J. Siddiqui, Hindi Word Sense Disambiguation using Semantic Relatedness measure, In Proceedings of 7th Multi-Disciplinary Workshop on Artificial Intelligence (MIWAI 2013), 9-11 Dec. 2013, Krabi, Thailand, pages 247-256, LNCS, Springer.

[27] S. Singh, T. J. Siddiqui, Utilizing Corpus Statistics for Hindi Word Sense Disambiguation, In International Arab Journal of Information Technology (IAJIT), Volume 12, No. 6A, December 2015, pages 755 - 763 (SCI Expanded, Impact Factor 0.582).