Word Sense Disambiguation For Kashmiri Language Using Supervised Machine Learning

Tawseef Ahmad Mir[1] , Aadil Ahmad Lawaye[2]
Baba Ghulam Shah Badshah University Rajouri - Jammu & Kashmir – 185234
[1] tawseefmir1191@gmail.com
[2] aadil.lawaye@gmail.com

1 Introduction

Every language spoken by people in this world contain words that have more than one meaning. Meaning of such words at a particular time depends on the context in which the word has been used. The process of selecting the meaning of ambiguous word from the set of possible meanings is called word sense disambiguation (WSD). WSD is one of the hot research topics in the natural language processing (NLP) domain. For humans it seems to be very easy to understand the meaning of ambiguous words but it is a very complex problem for machines to do so. Consider the following sentences in English:

- This saw is blunt

  I saw a horror dream yesterday

In the first sentence the word saw is used as noun and its meaning is a tool used to cut hard material like wood, metal. In the second sentence the word saw is past form of verb see.
Similarly consider the following sentence in Kashmiri:

- Thave dare yel
  Open the window
- Rache daare zeeth.
  Keep long beard.

In the above two sentences the word *daare* is having two different meanings. In the first sentence it means window where as in the second sentence it means beard.

In NLP WSD is considered as an AI Complete problem, that is, a problem whose solution presumes a solution to understanding natural language or common-sense reasoning (Ide et al, 1998).The meaning of an ambiguous word depends heavily on the words surrounding it. To resolve the ambiguity of words number of approaches have been designed till date and work is still going on. The research on WSD actually started in 1940's making it one of the oldest problems in the computational linguistics. Some important research works for handling WSD in various languages are (Abid et al, 2017), (Borah et al, 2019), (Khaled et al, 2012), (Richard et al, 2014), (Basuki et al, 2019), (Rajat & Sudip , 2015), (Himdweep et al, 2017), (Tarjni & Amit, 2019). For resolving ambiguity in Kashmiri language no work is cited. The objective of this research is to propose the WSD for Kashmiri using Supervised Machine Learning approaches.

2 Motivation

The driving motivation for this research is that WSD is an intermediate step for the various NLP applications like Machine Translation, grammatical analysis, speech processing, Information Retrieval and hypertext navigation (Ide & Veronis, 1998) and developing efficient WSD system is very crucial for the better performance of these NLP applications.

Since there is no work cited in Kashmiri language for handling WSD problem this is the first attempt in this direction, so this also motivated me for this research.

The third point is that research in NLP applications for Kashmiri language is in infancy stage this research will boost the research in this field and attract researchers to work in this field of Artificial Intelligence.

3 Research Challenges

Word sense disambiguation is a computationally

Proceedings of the 17th International Conference on Natural Language Processing, pages 243–245
Patna, India, December 18 - 21, 2020.

complex task and poses a lot of difficulties to the researcher. As far as this study is concerned there are a number of challenges. Notable challenges include:

Resource Scarcity: Kashmiri language is a resource poor language as adequate resources are not available for research which makes our task difficult. Only work done in Kashmiri so far in this domain is the development of some corpus and few linguistic tools under the project "Development of Language Tools and Linguistic Resources for Kashmiri" at the Department of Linguistics, University of Kashmir (Aadil et al, 2009), (Aadil et al, 2009), (Aadil et al,2010),(Aadil et al, 2013) , (Aadil et al, 2012) , (Aadil et al, 2012), (Aadil et al,2011).

Sense selection: One important issue related to WSD is to select senses of ambiguous word as different sources provide different divisions of words into senses.

Inter-Judge Variance: This study is the first attempt towards resolving ambiguity in Kashmiri language so the only option to evaluate the WSD system for Kashmiri language is human-judgement. But different humans may give different meanings for the same word. This increases the complexity of WSD task.

Discreteness of senses: WordNet contain very fine-grained senses and often it is very difficult to differentiate between these senses. This causes the disagreements among the lexicographers to specify which senses should be considered different ones for a particular word.

3 Grammar Formalism and Issues specific to Kashmiri Language

Grammar formalism is of great importance for creating syntactic annotation corpus and frameworks available can be categorized into two types: Dependency based annotation scheme and Constituency based annotation. In constituent-based annotation scheme sentence is depicted as hierarchically organized phrases and relation between and within constituents is not represented explicitly. In dependency based annotation the sentence is organized as dependency graph consisting of a head and dependent with labelled arc specifying relationship between them.

Kashmiri language being inflectionally rich dependency annotation scheme existing for Hindi-Urdu is considered suitable for annotation. But some issues needed to be addressed. These issues include V2 phenomenon, discrepancy that exists between coordinating and subordinating conjuncts, rift in complex predicates, pronominal clitics etc.(Bhat , 2012).

5 Methodology

In this study Supervised Machine Learning approaches are to be used to handle WSD in Kashmiri language. The Supervised Machine learning approaches work in two phases i.e, training phase and test phase. In training phase classifier is trained how to resolve ambiguity of a polysemous word (words having multiple meanings). In the testing phase the classifier assigns most appropriate sense to ambiguous word. The flowchart below depicts the methodology to be used:

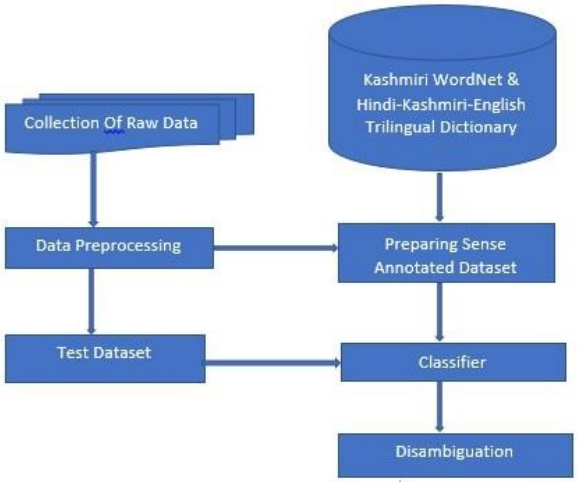

Fig 1. Proposed WSD System for Kashmiri

5.1 Collection of Raw Data

Data for this study will be collected from online (newspapers, blogs etc.) and offline resources. It's a very challenging task in this study.

5.2 Data Preprocessing

Usually the data is not readily available for research so it needs to be preprocessed to make it suitable for research. Data preprocessing usually involves stop word removal, data cleaning, stemming, removing inconsistencies in data.

5.3 Preparation of dataset

The data collected is divided into two sets. Training dataset and test dataset. The trained dataset sense tagged using Kashmiri WordNet and Hindi-Kashmiri-English Trilingual dictionary is used to

train the classifier so that it can disambiguate the ambiguous word for which it has been trained.

5.4 Classification

The supervised machine learning approach would be used to train the classifier and the classifier would be used to predict the meaning of the polysemous word in the test phase.

6 Experimental Outcomes

The main outcomes of the study are as follows:

1. Sense Annotated Corpus for Kashmiri Language.
2. WSD Data Set.
3. Word Sense Disambiguation System for Kashmir Language.

References

Aadil Amin Kak, Nazima Mehdi and Aadil Ahmad Lawaye 2009. What should be and What should not be? Developing a POS tagset for Kashmiri. Interdisciplinary Journal Of Linguistics (IJL), 2 185-196

Aadil Amin Kak, Nazima Mehdi and Aadil Ahmad Lawaye.. 2009. Towards Developing A Tagset For Kashmiri. Nepalese Linguistics, 49-60.

Aadil Amin Kak, Nazima Mehdi and Aadil Ahmad Lawaye 2010.Building a Cross Script Kashmiri Converter. Issues and solutions. *Proceedings of Oriental COCOSDA (The International Committee for the Co-ordination and Standardization of Speech Databases and Assessment Techniques)*.

Aadil Ahmad Lawaye and Prof. Bipul Syam Purkayastha2 2013. Towards Developing A Hierarchical Part Of Speech Tagger for Kashmiri: Hybrid Approach. *Proceedings of the 2nd National Conference on Advancements in the Era of Multidisciplinary Systems, Elsevier Publications, 187-192.*

Aadil Ahmad Lawaye, and Dixit N. 2012. Multilingual Dictionary Generation Using Indo-Wordnet:A Proposal', THE COMMUNICATIONS-Journal of Applied Research in Open and Distance Education, 188-191.

Aadil Ahmad Lawaye, Bipul Syam Purkayastha. 2014. *Kashmir Part of Speech Tagger Using CRF", Indian Journal of Research, 37-38*

Basuki, Setio, Ali Sofyan Kholimi, Agus Eko Minarno, Fauzi Dwi Setiawan Sumadi, and M. Rizal Arif Effendy. 2019 Word Sense Disambiguation (WSD) for Indonesian Homograph Word Meaning Determination by LESK Algorithm Application," *12th International Conference on Information & Communication Technology and System (ICTS), Surabaya, Indonesia, 2019,* pp. 8-15

Bhat, S.M., 2012, December. Introducing Kashmiri dependency treebank. *In Proceedings of the Workshop on Machine Translation and Parsing in Indian Languages* (pp. 53-60).

Borah, Pranjal Protim, Gitimoni Talukdar, and Arup Baruah. 2019 WSD for Assamese Language. *In Recent Developments in Machine Learning and Data Analytics*, pp. 119-128. Springer, Singapore

Himdweep Walia, Ajay Rana, Vineet Kansal. 2017. A Naïve Bayes Approach for working on Gurmukhi Word Sense Disambiguation", *6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions), Amity University Uttar Pradesh, Noida, India*

Ide, Nancy & Jean Véronis. 1998. Word sense disambiguation: The state of the art. Computational Linguistics

Khaled Abdalgader M. OmarAndrew Skabar. 2012 Sense disambiguation using context vectors and sentential word importance," ACM Transactions on Speech and Language Processing, vol. 9, no. 1, pp. 1-21

Muhammad Abid, Asad Habib, Jawad Ashraf, and Abdul Shahid. 2017. Urdu word sense disambiguation using machine learning approach". Cluster Computing, pages 1—8

Nancy Ide and Jean Veronis 1998, Introduction to the Special Issue on Word Sense Disambiguation: The State of the Art

Nazima Mehdi and Aadil Ahmad Lawaye 2011. Development of Unicode Complaint Kashmiri Font: Issues and Resolution. Interdisciplinary Journal of Linguistics (IJL), 4,195-200

Pandit, Rajat, and Sudip Kumar Naskar. 2015. A memory based approach to word sense disambiguation in Bengali using k-NN method. *2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS),* Kolkata, pp. 383-386.

Singh, Richard Laishram, Krishnendu Ghosh, Kishorjit Nongmeikapam, and Sivaji Bandyopadhyay 2014. A Decision Tree Based Word Sense Disambiguation System In Manipuri Language. Advanced Computing: An International Journal 5(4) p 17.

Tarjni Vyas, Amit Ganatra 2019. Gujarati Language Model:Word Sense Disambiguation using Supervised Technique. International Journal of Recent Technology and Engineering (IJRTE) ISSN: 2277-3878, Volume-8, Issue-2S11, September 2019