Polysemy Identification for Dogri Language

Shubhnandan S. Jamwal
jamwalsnj@gmail.com
Department of Computer Science and IT, University of Jammu, Jammu

Abstract

Many of the NLP systems fails to identify the Word Sense Disambiguation and because of the Polysemy. This problem persists in almost all the NLP systems for the Indian languages. The approach for the identification of this problem depend on language specific knowledge based on the context of the polysemy words and some pre-existing natural language processing (NLP) tools for the low resourced languages like Dogri. In this paper a model is proposed for the polysemy identifications and management based on the context. The proposed model will suggest the meaning based on the context.

Keyword

Polysemy, Dogri, Word Sense Disambiguation

Introduction

Dogri is an Indo-Aryan language and is a mother tongue of 422 million people and recently it has been introduced as an official language of the Jammu and Kashmir. It is the second prominent language of J&K State, presence of Dogri language can also be felt in northern Punjab, Himachal Pradesh and other places. Natural language processing is a field of Artificial Intelligence that deals with the methods of communicating with computers in natural languages like English, Hindi, Dogri etc. The researchers are developing number of NLP techniques for the different Indian languages for computational and analyzing processes which can enable a computer to understand the language.

The term polysemy refers to a word having more than one meaning. The Word sense disambiguation is a very complex problem which can be solved to some extent by properly managing the different context. The polysemy words and word sense disambiguation are the problems of NLP which are growing a lot in the written text rapidly in the digital world. Dogri language is also growing, although slowly, in the digital world but the problem of WSD will remain a challenge for the researchers of this language also. If any Indian language, will not grow digitally, it will raise a huge gap of information and disparity that exists in most countries between those who have quick access to information in regional language and those who do not.

The failure of the NLP systems to identify the Word Sense Disambiguation and Polysemy can cause problems in NLP systems. The approach for the identification of this problem depend on language specific knowledge based on the context of the polysemy words and some pre-existing natural language processing (NLP) tools for the low resourced languages like Dogri. This is not only the case of Dogri but many Indian languages have less resources and tools as compared to English for WSD.

Literature Review

R. Rao and J. S. Kallimani [1] has created Kannada polysemy word analyzer to solve the ambiguity. It allows the user to select a Kannada sentence from a list provided. It uses the

ISSN: 2229-6913(Print), ISSN: 2320-0332(Online)
Received: 15-12-2020; Revised: 25-03-2021; Accepted: 13-05-2021

Shallow parser, which gives the parts of speech information of each word in a sentence. Different meanings of a polysemy word are stored in a database. When the correct match is found between the shallow parser results and the database the exact meaning of the polysemy word used in the sentence is highlighted. Y. Lin, M. Yu and C. Lin [2] had shown that using language models to solve the polysemy problems can have very good results. We propose a combined approach to the polysemy problems in this paper. There are six words with polysemy problems to be solved in this paper. They are (you), (I), (he), (no), (up), and (down). The numbers of pronunciation of these six words are 2, 2, 2, 6, 3, and 4, respectively. Results show that the proposed combined approach can achieve higher accuracy than the existing methods, WU and DLC. H. Miao and Y. Zhang [3] paper adopted the inverted index to construct the polysemy index table which is based on the annotated corpus, and achieve a small search engine based on the Lucene technology. After studying the frequency and distribution of the Polysemy in the labeled corpus, they constructed the index table to improve the speed of searching polysemy in corpus. The small search engine realized the function of web resources crawling and extracting, indexing, search word processing, resources querying and sorting. It extended the application of inverted index technique in the network information retrieval. C. Prasad and J. S. Kallimani [4] proposed a method to extract various meanings of the polysemy words in Kannada where, identification of correct meaning of a polysemy word based on the context is very important in abstractive text summarization. The algorithm takes the word as an input and performs pattern matching with the words available in the dictionary and finally displays the result. This idea will be initially developed as a normal application and also proposed that later an attempt will be made to deploy it on the cloud and to make a cross-platform mobile application so that it can be used in any device that the user is comfortable with. U. R. Dhungana, S. Shakya, K. Baral and B. Sharma [5] presented a new model of WordNet that is used to disambiguate the correct sense of polysemy word based on the clue words and then the related words for each sense of a polysemy word as well as single sense word are referred to as the clue words. The conventional WordNet organises nouns, verbs, adjectives and adverbs together into sets of synonyms called synsets each expressing a different concept. In contrast to the structure of WordNet, we developed a new model of WordNet that organizes the different senses of polysemy words as well as the single sense words based on the clue words. These clue words for each sense of a polysemy word as well as for single sense word are used to disambiguate the correct meaning of the polysemy word in the given context using knowledge-based Word Sense Disambiguation (WSD) algorithms. The clue word can be a noun, verb, adjective or adverb. Z. Qin, H. Lian, T. He and B. Luo[6] focused on the polysemy and synonymy issue in clustering process. Polysemy represents the ambiguity of an individual word or phrase that can be used (in different contexts) to express two or more different meanings. However, synonymy is the semantic relation that holds between two or more words that can (in a given context) express the same meaning. These two conditions will affect our results of clustering. In order that, we use bag of words model to distinguish contexts of the same words and word2vec to re-cluster word with the similar meaning. Cosine similarity is also use to measure of similarity between two nonzero vectors in these two models. Y. Lin, M. Yu and

ISSN: 2229-6913(Print), ISSN: 2320-0332(Online)
Received: 15-12-2020; Revised: 25-03-2021; Accepted: 13-05-2021

C. Huang[7] brings up an important issue, the polysemy problems, in a Chinese to Taiwanese TTS system. Polysemy means there are words with more than one meaning or pronunciation, such as "(we)", "(no)", "(you)", "(I)", "(want)", and so on. They focused on the Chinese word "(we)" to show how imperative the polysemy problem in a Chinese to Taiwanese TTS system is. There are two pronunciations of the word "(we)" in Taiwanese, /ghun/ and /lan/. The corresponding Chinese words are and. We propose two approaches and their combination to solve this problem. The results show that we have a 93.1% precision in translating the correct meaning and pronunciation of the word "(we)" from Chinese to Taiwanese. Maria Lapata [8] investigated polysemous adjectives whose meaning varies depending on the nouns they modify and acquired the meanings of these adjectives from a large corpus and propose a probabilistic model which provides a ranking on the set of possible interpretations. They identified lexical semantic information automatically by exploiting the consistent correspondences between surface syntactic cues and lexical meaning. We evaluate our results against paraphrase judgments elicited experimentally from humans and show that the model's ranking of meanings correlates reliably with human intuitions: meanings that are found highly probable by the model are also rated as plausible by the subjects.

Proposed Model for Polysemy Dogri Words

The Dogri language remains a low resourced language because a low number of researchers are working on the problems of Dogri. Recently work on verbs of the Dogri [9] and the development of the stemmer [10] for the Dogri language has been published but the problem of the polysemy and word sense disambiguation still remains a big challenge in the Indian languages. One of the major causes of difficulties in NLP tasks and in turn results in semantic clustering of words in a large corpus. The model proposed in the paper will be used for the identifying the polysemy words in the Dogri corpus. The proposed model will accept the input in the form of the Dogri sentence, which will go to the polysemy identifier engine. The engine will take details from the polysemy word array of all the words of the input sentence. The other database will keep the polysemy words along with the meanings with respect to the context. The output of the engine will be in two parts, first the different meaning of the words will be displayed and in the second part the meaning based on the context can be suggested.

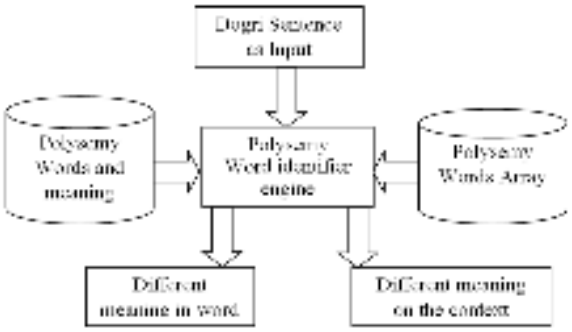

Fig 1: Dogri Polysemy engine

Conclusions

ISSN: 2229-6913(Print), ISSN: 2320-0332(Online)
Received: 15-12-2020; Revised: 25-03-2021; Accepted: 13-05-2021

The goal of this study is to develop a method for identifying the polysemy words from a Dogri sentence. The model can be implemented to increase the accuracy of a variety of natural language processing applications. As a result, generating a list of similar words having different meanings based on the context would be a crucial part of natural language processing. In this paper a model is proposed for the polysemy identifications and management based on the context. The proposed model will suggest the meaning based on the context.

References

[1] R. Rao and J. S. Kallimani, "Analysis of polysemy words in Kannada sentences based on parts of speech," 2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI), 2016, pp. 500-504, doi: 10.1109/ICACCI.2016.7732095.
[2] Y. Lin, M. Yu and C. Lin, "A combined approach to the polysemy problems in a Chinese to Taiwanese TTS system," 2010 7th International Symposium on Chinese Spoken Language Processing, 2010, pp. 455-459, doi: 10.1109/ISCSLP.2010.5684482.
[3] H. Miao and Y. Zhang, "Construction of Polysemy table and search engine based on inverted index," 2012 9th International Conference on Fuzzy Systems and Knowledge Discovery, 2012, pp. 2287-2290, doi: 10.1109/FSKD.2012.6234156.
[4] C. Prasad and J. S. Kallimani, "A novel approach to identify polysemy words in Indian regional language," 2016 International Conference on Electrical, Electronics, Communication, Computer and Optimization Techniques (ICEECCOT), 2016, pp. 179-182, doi: 10.1109/ICEECCOT.2016.7955210.
[5] U. R. Dhungana, S. Shakya, K. Baral and B. Sharma, "Word Sense Disambiguation using WSD specific WordNet of polysemy words," Proceedings of the 2015 IEEE 9th International Conference on Semantic Computing (IEEE ICSC 2015), 2015, pp. 148-152, doi: 10.1109/ICOSC.2015.7050794.
[6] Z. Qin, H. Lian, T. He and B. Luo, "Cluster Correction on Polysemy and Synonymy," 2017 14th Web Information Systems and Applications Conference (WISA), 2017, pp. 136-138, doi: 10.1109/WISA.2017.45.
[7] Y. Lin, M. Yu and C. Huang, "The Polysemy Problems, an Important Issue in a Chinese to Taiwanese TTS System," 2008 Congress on Image and Signal Processing, 2008, pp. 361-365, doi: 10.1109/CISP.2008.728.
[8] Maria Lapata. 2001. A corpus-based account of regular polysemy: the case of context-sensitive adjectives. In Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL '01). Association for Computational Linguistics, USA, 1–8. DOI:https://doi.org/10.3115/1073336.1073345
[9] Jamwal S.S., Gupta P., Sen V.S. (2021) Hybrid Model for Generation of Verbs of Dogri Language. In: Singh T.P., Tomar R., Choudhury T., Perumal T., Mahdi H.F. (eds) Data Driven Approach Towards Disruptive Technologies. Studies in Autonomic, Data-driven and Industrial Computing. Springer, Singapore. https://doi.org/10.1007/978-981-15-9873-9_39.
[10] Gupta P., Jamwal S.S. (2021) Designing and Development of Stemmer of Dogri Using Unsupervised Learning. In: Marriwala N., Tripathi C.C., Jain S., Mathapathi S. (eds) Soft Computing for Intelligent Systems. Algorithms for Intelligent Systems. Springer, Singapore. https://doi.org/10.1007/978-981-16-1048-6_11.