Case Based Construal using Minimal Features to Decipher Ambiguityin Punjabi Language

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

[1,2]Amity University, Noida, Uttar Pradesh, India
[3]IET, Lucknow, Uttar Pradesh, India
[1]himdweep@yahoo.com, [2]ajay_rana@amity.edu, [3]vineetkansal@yahoo.com

Abstract: A minimal feature set is taken in order to find the closest context of the given ambiguous word by using case based construal model. The model proposed in this paper uses case-based reasoning to sort out the similar cases with vectors of size two (bigram), three (trigram) and four (n-gram) using Euclidean similarity function. These cases are then subjected to three different classifiers, namely – Bayes, k-Nearest Neighbor, and Decision Tree, to decipher the ambiguity.Vectorization eases the process of disambiguation as compared to when full sentences are processed. The similarity function helps to find similar cases of the given ambiguous word whereas the three classifiers helps in finding the right context of the given ambiguous word. Upon experimentation, Decision Tree classifier has achieved an accuracy of 84.88% using pre-bigram vectors.

Keywords: Word Sense Disambiguation, Case Based Reasoning, n-gram, similarity function, Bayes Classifiers, k-Nearest Neighbor, Decision Tree

I. INTRODUCTION

The natural language is laced with words having multiple meanings used in different contexts. Although this understanding of disambiguation comes naturally to humans, implementing the same with the machines is comparatively difficult. For instance, consider the word, "break", which has an astounding 75 different meanings ranging from "interruption" to "pause" to "time interval" to "social separation". This aspect makes Word Sense Disambiguation (WSD) an NP-hard problem in the Natural Language Processing (NLP) domain, and therefore it becomes imperative to draw algorithms / models that can help in correctly deciphering the ambiguity. This will be useful in improving the accuracy in language understanding, generation, machine translation, information retrieval, etc.

The work is this paper majorly focusses on an Indian Regional Language, Punjabi, spoken in the Indian subcontinent. Punjabi is ranked 10th in the world in the list of languages by number of native speakers. A number of books, newspapers, journals and documents are written in Punjabi and are widely in use. And like every language, this language too has its own set of ambiguous words. Consider the Punjabi word, "ਵੱਟ (vatt)", which has 8 different meanings ranging from "cramps in stomach" to "irritation" to "rolling of moustache" to "wrinkling of skin".

The dictionary approach to find the correct sense of the word is both complex as well as time consuming and the supervised machine learning approach is little hard to formalize as they require training the machine first. For this reason, Case Based Reasoning (CBR) is adopted which implements the concept of text comparison to disambiguate the words. CBR refers to previously documented cases with similarities to the new problem at hand to decipher the right context. One major concern though is the length of the text to be examined. For this reason, the input is given in the form of single sentence at a time which has shown better results. The proposed model takes the input in sizes two, three and four and then classifiers are used to find the closest context to disambiguate.

The further sections are chronologized as follows: section II describes the related research work, section III discusses the case based construal, section IV explain how the proposed model is executed and implemented, section V explains the results achieved and finally section VI concludes the paper.

II. LITERATURE REVIEW

In order to produce better results in applications like Machine Translation (MT), Information Retrieval (IR), Information Extraction (IE), Speech Recognition (SR), Speech Synthesis (SS) and likewise, it has become imperative that we improve the results of WSD algorithms. This has led to a number of research papers being published in this domain [1]. The approach that has been widely followed is knowledge-based and machine learning [1].

As the work in this paper is being carried out in an Indian Regional Language i.e. Punjabi, we went through the various research papers that have been written in Punjabi WSD [2, 3, 4]. The major work has been done in developing lexical resources in order to improve translations from English to Indian regional Languages and vice versa. Further various dictionary based and supervised and unsupervised approached have been implemented for deciphering the correct sense of the ambiguous word. The concept of CBR in Punjabi WSD has

978-1-5386-9346-9/19/$31.00 ©2019 IEEE

been introduced in the paper by H.Walia, et.al [17], which has been accepted.

CBR is a different approach that is being followed in artificial intelligence to help to arrive at the solutions [13]. The concept of CBR is that we refer to similar cases as the one in hand. The exciting case can be even adapted so that we find the solution. The similarity between the existing case and new case can be determined using similarity functions [14, 15].

Data sparseness and inconsistency in vocabulary are the major problems that we encounter in NLP. The reason is that in order to decipher ambiguity, we need a context window [11] i.e. the number of surrounding words with the ambiguous word in order to find the meaning of the given ambiguous word. In the paper by H. Walia, et.al. [16], out of the 3 window sizes – 3, 5, 7 – which were experimented with, the window size of 7 gave the best results.

In order to remove sparseness, it is important that we reduce the size of context window. In the paper by T. Pedersen [12], the author experimented with bigrams i.e. two words - ambiguous word and either pre or post word alongside the ambiguous word. Using the Naive Bayes and Decision tree classifiers, the author them deciphered the ambiguity.

In the papers by P. Tamilselvi, et.al. [14, 15], the authors have used cased based reasoning to find similar cases and then experimented with different feature sets. In the paper [15], they took two feature vectors – pre-bigram and post-bigram which was then processed using k-NN and ANN for deciphering the ambiguity. The experimentation showed that pre-bigram gave better results. In their paper [14], they took three feature vectors – bigram, trigram and n-gram. Using three different similarity functions and then subjecting the data to three different classifiers, they concluded that pre-bigram feature vector with Euclidean distance function with Bayes classifier produces the best result.

III. CASE BASED CONSTRUAL

The case based construal is based on a 4-fold process – pre disambiguation, vector representation, case extraction and sense disambiguation.

A. Pre-Disambiguation Process

The size of the processing document is usually huge, implying that most of the sentences are lengthy and having conjunctions joining two or more sentences together. This makes it a little difficult to be processed using CBR. Therefore, the document is pre-processed and broken into single sentences. These single sentences are then tokenized i.e. we remove all the stop words (i.e. prepositions, conjunctions, etc.) (The Punjabi language has 184 stop words [11]) from the sentence and are left with "bag of words". Also we separate the compound words like "ਮਰਜਾਵਾਂ (mrrjawan)" into "ਮਰ (mrr)" and "ਜਾਵਾਂ(jawan)".

Fig. 1. CASE BASED CONSTRUAL

Following this, we identify the ambiguous words in the given corpus [7] from the list of Punjabi ambiguous words identified in Punjabi Word Net[6].

B. Vector Representation

To create a vector we require semantic markers along with the ambiguous word. The semantic markers refer to the words positioned on the left and right side of the given ambiguous case. Here we are taking 3 words on the left side (L_{w1}, L_{w2}, L_{w3}) and 3 words on the right side (R_{w1}, R_{w2}, R_{w3}) of the given word. Feature Vector Representation is defined in the table given below:

TABLE I: FEATURE VECTOR REPRESENTATION

Column	Fields	Description
C1	Case	Ambiguous
C2	Sense_value	Sense valu
C3	Sense_tag	Sense tag
C4	L_{w1} L_{w2} L_{w3}	Weight of t

After identifying the ambiguous words in the given corpus along with their semantic markers, we need to identify the number of features that we want to study for our case analysis. In this paper, we are providing the model with three different inputs – 2 (bigram), 3 (trigram) and 4 (n-gram). Table II gives the interpretation of these three feature types.

T1 represents pre-bigram which means the ambiguous word along with immediate next word. T2 represents post-bigram which means the ambiguous word along with immediate previous word. T3 represents pre-trigram which means the ambiguous word along with immediate next two words. T4 represents in-gram which means the ambiguous word along with previous one word and next one word. T5 represents post-trigram which means the ambiguous word along with two previous words. T6 represents pre-bigram which means the ambiguous word along with previous two words and one next word.

TABLE II: FEATURE TYPES

	Feature type	F1	F2	F3
T1	Pre-bigram	W	R_{w1}	-
T2	Post-bigram	L_{w1}	W	-
T3	Pre-trigram	W	R_{w1}	R_{w2}
T4	In-trigram	L	W	R

C. Case Extraction

The case extraction process is based on the approach of CBR which has been carried out in three steps as shown in Fig 2.

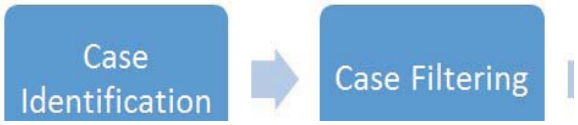

Fig. 2. STEPS IN CASE EXTRACTION

In case identification, we referred Word Net to pick ambiguous words (for our paper, we picked 4 such words – ਉੱਤਰ, ਕੱਚਾ, ਹਾਰ, ਉਲਟਾ) and then extracted the cases with these ambiguous words. In case filtering, we filtered the words with PoS similar to that of the given ambiguous word and rest were discarded. And finally in case selection, the alike cases were selected with the help of similarity function like Euclidean.

TABLE III: NUMBER OF SENSES FOR GIVEN AMBIGUOUS WORD

Ambiguous Word	Nu se dic
ਉੱਤਰ	
ਕੱਚਾ	

D. Sense Disambiguation

The corpus [7] is provided as the input to the model which is then subjected to pre-disambiguation process where the corpus is reduced into a set of single sentences. The stop words and compound words are processed to get what is known as the "bag of words". We then refer to the Punjabi Word Net to collect the ambiguous words to be deciphered.

The next step is to prepare vectors of bigram, trigram and n-gram for a given ambiguous word from the bag of words generated during the pre-disambiguation step. Now applying the concept of CBR, we picked the cases having similar PoS as our vector representation of the ambiguous word. Then using the Euclidean similarity function, we are able to find the similar cases.

The three classifiers – Bayes, k-NN and Decision Tree are then used to seek the best case. The proposed case based model for disambiguation is as shown in Fig 3 below:

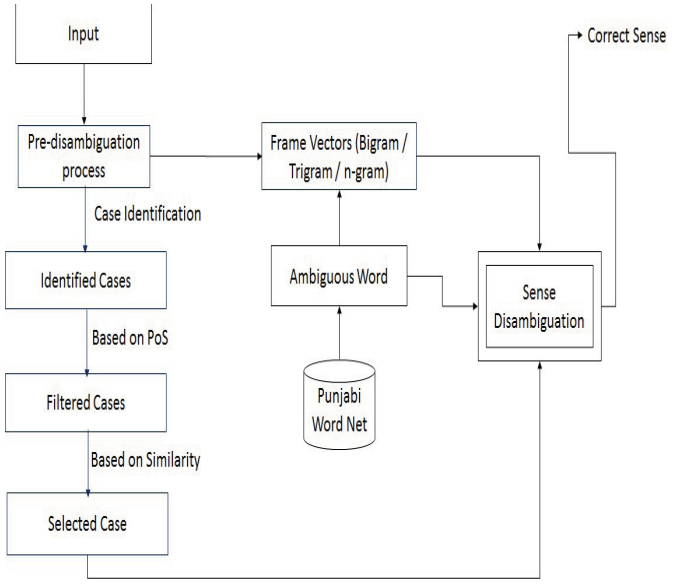

Fig. 3. CASE BASED MODEL FOR DISAMBIGUATION

IV. EXPERIMENTAL RESULTS

In this paper, we have incorporated the concept of CBR while deciphering the ambiguity. The problem with the supervised approaches have been that they require a context window of large size to decipher the meaning of an ambiguous word. This increases the sparseness due to inclusion of large number of features. By incorporating the concept of CBR and vector representation, we have tried to reduce the size of the context window.

By using the Euclidean similarity function, we were able to pull out similar cases and then these were compared with the input vector. The Euclidean similarity function alone was used as results in previous papers [14, 15] showed that this function had a better closeness than other similarity functions like Cityblock and Cosine.

We had prepared the vector representation with 6 different scenarios i.e. pre-bigram (T1), post-bigram (T2), pre-trigram (T3), in-trigram (T4), post-trigram (T5) and n-gram (T6). The cases similar with respect to their PoS for the given ambiguous words (refer Table III) were taken for our case study and the result is shown in Table IV.

TABLE IV: SENSE DISAMBIGUATION ACCURACY

Word	Learning Classifier	T1	T2	T3	T4
ਉੱਤਰ	Naïve Bayes	83.03	82.21	81.66	78.4
	K-NN	77.18	68.09	77.41	70.9
	Decision Tree	83.23	82.20	82.20	82.1
ਕੱਚਾ	Naïve Bayes	84.23	84.23	84.23	84.1
	K-NN	78.11	73.12	76.67	74.3
	Decision Tree	84.65	84.65	84.65	83.6
	Naïve Bayes	85.19	84.12	83.67	84.1
	K-NN	73.21	74.89	72.39	74.9

Of the three classifiers – Bayes, k-NN and Decision Tree, the Decision Tree showed the best result in case of pre-bigram vector. We calculated the average of all the four words used for the experimentation and it gave us the value of 84.88% accuracy.

V. CONCLUSION

In this paper, we have used the Euclidean similarity function to pull out the similar cases with respect to our input vector. The cases were extracted based on pre-bigram, post-bigram, pre-trigram, in-trigram, post-trigram and n.-gram. These vectors were then subjected to 3 classifiers, namely – Bayes, k-NN and Decision Tree. The best results were shown by the Decision Tree classifier with 84.88% using pre-bigram vector, followed by Bayes classifier Tree using pre-bigram vector and then k-NN using n-gram vector (where we used 4 features).

ACKNOWLEDGEMENTS

The author[1] acknowledges Amity University, Uttar Pradesh, where the author is registered as scholar for providing with the library facility and also an extensive online resources for research.

REFERENCES

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

[2] M. Bansal, "Word Sense Disambiguation: Literature Survey for Indian Languages", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 12, December 2015.

[3] H. Walia, A. Rana, and V. Kansal, "A study on different Word Sense Disambiguation Approaches and their application on Indian Regional Languages", Proceedings of International Conference on Technology and Trust (ICTT'17), 28-29 December, 2017.

[4] R. Kumar, and R. Khanna, "Natural Language Engineering: The Study of Word Sense Disambiguation in Punjabi", IJES, July2011.

[5] H. Walia, A. Rana, and V. Kansal, "Different Techniques Implemented in Gurumukhi Word Sense Disambiguation", International Journal of Advanced Technology in Engineering and Science, Volume 05, Issue 06, June 2017.

[6] Punjabi WordNet. [Online]. Available: http://tdil-dc.in/indowordnet/index.jsp

[7] Punjabi Corpora. Obtained from Evaluations and Language Resources Distribution Agency, Paris, France.

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

[9] J. Kaur, J. R. Saini, "Punjabi Stop Words: A Gurmukhi, Shahmukhi and Roman Scripted Chronical", Proceedings of the ACM Symposium on Women in research, March 2016, Page 32-37.

[10] J. Kaur, and V. Gupta, "Effective Approaches for extraction of Keywords", International Journal of Computer Science, Issue 6, Vol. 7, November 2010.

[11] S. Singh, and T. J. Siddiqui, "Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation", Proceedings of the International Conference on Information Retrieval & Knowledge Management, CAMP2012, 13-15 March, 2012, Malaysia, pp.1-5, IEEE Explorer.

[12] T. Pedersen, "A decision tree of bigrams is an accurate predictor of word senses", Presented at Second Annual Meeting of the North American Chapter of the Association for Computational Linguistics, 2001

[13] Janet L. Kolodner, "An Introduction to Case-Based reasoning", Artificial Intelligence Review 6, Pg 3-34, 1992.

[14] P. Tamilselvi, and S.K.Srivastsa, "Case Based Word Sense Disambiguation Using Optimal Features", Presented at International Conference on Information Communication and Management, 2011.

[15] P. Tamilselvi, and S.K.Srivastsa, "Word Sense Disambiguation using case based Approach with minimal Features Set", Indian Journal of Computer Science and Engineering, 2011.

[16] H. Walia, A. Rana, and V. Kansal, "Word Sense Disambiguation: Supervised Program Interpretation Methodology for Punjabi Language", Proceedings of IEEE 7th International Conference on reliability, Infocom Technologies and optimization (ICRITO'2018), 29-31 August, 2018, India.

[17] H. Walia, A. Rana, and V. Kansal, "Case Based Interpretation Model for Word Sense Disambiguation in Gurmukhi".