TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 10320 - 10328

Equivocation based Pun Detection using H-LSTM Framework in Code-Mixed Text

Shashi Shekhar, Dilip Kumar Sharma
Department of Computer Engineering and Applications,
GLA University, Mathura, India

Article Info
Volume 83
Page Number: 10320 - 10328
Publication Issue:
May - June 2020

Abstract:

The computer systems need to be equipped to extract emotional expressions from code mixed text to better understand the human language phenomena. Text *on* social media contains code-mixed contents which can be used to extract equivocation information. This equivocation information is hard to be retrieved in transliterated domain. The extraction of equivocation expression used by the people to express their opinions on web is a challenging task in code mixed environment. The work presents the comparison of different approaches in code mixed social media text in transliterated domain. A rule based approach is proposed, that accepts text in code-mixed format as input and based on the defined rules, the system provides the equivocation expression in the sentence. The hypothesis is evaluated on the basis of experiments undertaken for the rule-based approach along with standard statistical approaches. On the obtained results of rule based approach and statistical approach, a voting technique is applied which selects the best equivocation tag for a word based on majority. This voting tag is also useful when all the three approaches gives different tag for a word the voting approach helps in considering the best equivocal tag. This voting approach performs best among all the different approaches used in the experiment with high accuracy.

Article History
Article Received*: 19 November 2019*
Revised*: 27 January 2020*
Accepted*: 24 February 2020*
Publication*: 18 May 2020*

Keywords*: NLP, transliteration,ambiguity,embedding, equivocal,mixed script.*

1 Introduction

Equivocal expressions now days are frequently used in social media. Equivocal information processing is a challenging task in transliterated information retrieval. The equivocation helps in answering the ambiguous category of expressions by identifying the words based on its context meaning. Also it helps in summarizing the things in terms of ambiguity order. Many researches has been undertaken to produceequivocal annotation of text in several languages. But very few workshas been done in Indian language transliteration. The paper explores this area inretrieving ambiguity based information in transliterated Roman Hindi domain. Here the classification related to humor expressions in Roman Hindi is analyzed. The paper presents a voting based approach based on crafted rules for classifying equivocal humor data in internet domain. The rules are modeled subject to regular expression matching. When the matching is done, it checks for ambiguous data and the words are tagged as equivocal expression. The labeling parameters consisting of equivocal dimension are checked against the rule and label that word as equivocal expression. Here in this work, we first analyze the structures of words in sentence to identify the ambiguity in code-mixed text and classify them as ambiguous or non-ambiguous. The language identification task is used here as we are concern only about the transliterated Hindi words that are checked in context to the entire sentence for identifying the equivocal words. For context identification technique of H-LSTM based framework has been designed using CBOW

May – June 2020
ISSN: 0193-4120 Page No. 574 - 585

technique that targets majority of the ambiguous words. Word and language identification in user-generated text is tedious task, where the language is unknown. Now a day, it is a challenging task where the text is available in code–mixed format. This type of data is very common in social media. The main challenge here is due to availability of many transliteration variants for a given word. Lastly we test our approach on a dataset of Amul advertisements in India [1] and the proposed framework is able to recover equivocal words.

The available identification systems are not equipped to deal with equivocal data. This paper describes the use of equivocal data to identify the language as well as the dimensions of the humor context in which it has been used in the expression. This identification is necessary for the languages which are linguistically much related with each other. A special technique is needed to differentiate the words which are syntactically similar in both the languages. Natural language is one of the medium for communication in India. The processing of this by the machine requires specialized skill to extract meaningful information based on humor dimension. It is an emerging area of research for extracting intent of the user for using ambiguous humor expression for expressing opinions. With the huge use of social media platforms for information exchange, it is likely to have natural language data that needs to be processed by the machine to get information. These platforms are widely used by Indians to discuss any issue especially using their own native languages. Previously we were using mainly English language for such communication but in present scenario peoples are using mixed script contents for information exchange. Now a day's in Indian scenario, people are mixing more than one language for expressions to be posted on social media. These scenarios are leading to the field of code-mixing. To better understand the scenario of code-mixing an example has been illustrated from the advertisement of Amul, which describes the exploration of equivocal expression in present time. Transliterated Hindi-English code-mixed is described in the following sentence:

Sentence 1: Namaske President Trump
H/EQ E NE

Here, words in Roman Hindi are labeled as H, English as E, Named entity as NE and ambiguous equivocal as EQ. The ambiguous equivocal words describe the ambiguity expression in roman Hindi in the sentence. In sentence 1 the word Namaske is marked as EQ, it illustrates that word denotes the equivocal expression. The proposal describes an architecture that represents context level information for presenting the equivocal tag associated with context dimension words used in the sentence, especially to those words which are marked as Hindi word.

The rest of the paper is organized as follows: section 2 illustrates the state of art in equivocal retrieval. A discussion on the methodology proposed is discussed in section 3. The dataset description and result evaluation is described in section 4, and finally, section 5 concludes the work with a pointer to the future direction.

2Related Work

This section provides the literature review in recent techniques regarding temporal information in transliterated domain.

Code-mixing is an emerging area of research in the field of language classification. Identifying the language is the major task for any linguistic processing applications. Presently several type of research is going on in the field of code-mixing. The proposal of King etal.[2] utilizes supervised mechanism for language identification. The paper [3] implements CRF model for identifying the language. The proposal of given in [4] uses logistic regression, in code-switching environment. Das et.al. [5] proposed the use of dictionary along with

May-June 2020
ISSN: 0193-4120 Page No. 3672 - 3680

the concept of edit distance to find word origin in regard to word context.

The task conducted on Mixed Script Information Retrieval (MSIR), where language identification for Indian languages combined with other languages have been scheduled [6] focusing on the use of transliteration. The task of MSIR was evaluated using SVM attaining an accuracy of greater than 75% [7]. The proposal of [8] uses supervised learning for English-Hindi word identification. The use of Naive Bayes classifier [9] was proposed for Hindi-English data. The paper [10] proposed embedding technique as a feature for entity extraction.

The paper [11] describes the advancements in neural learning model through fusion of CNN and LSTM for language identification. The model also contains Bi-LSTM –CRF for context classification in terms of word-level language identification.

Moving towards ambiguity identification, recent work [12] on ambiguity detection using the LSTM model investigates the ambiguity in English by discovering hidden representations in regard to contextual information regarding the words giving future scope in improving the word contextual information. The below section describes the proposed model for ambiguity detection for those English words, which are also used in for representing Hindi words in Roman script [16][17]. Word embedding's in query translation [13] has been investigated for improving precision in a multilingual environment [18]. This embedding is done considering the approach that similar kind of words will have similar kind of vectors. The paper [14] claims the novel work in multilingual NER using deep learning. Phrase extraction has been handled by [15] using speech-based transliteration.

3Proposed Framework

The proposed work is inspired by the latest work [20][21] undertaken in the field of language pairs that have different lexicons representing different context meanings when combined with other words. The research is underway based on the neural learning architecture to understand equivocation in finding humor with the help of pre-trained embedding technique for building RNN based transliteration model. The proposal presented in this paper is based on related research findings in the area of code-mixing. The intricacy to identify the language of the temporal expression words in code mixed data is presented in work. The code mixed data includes more than one script. Due to this mixing, complication in processing is bound to arise. Language classification with accuracy is the foremost problem identified in this case. The problem of identifying language in these domains is more complex as the text contents are written in different languages and it is difficult to identify humor equivocation information in such cases. The following section describes the equivocation architecture based on HLSTM model for extracting humor sense from the data.

The figure 1 describes the proposed HLSTM system for equivocation data retrieval from code-mixed domain. The proposed model is trained at the word level using the hierarchical LSTM on the basis of features selected for identifying ambiguous equivocation words in code mixed data. The system takes code mixed input. The data is tokenized as embedding process is based on word embedding and character embedding. There exist many probable equivocation classes for Hindi roman words. Character embedding is done for Hindi roman words as per the defined classes presented in table 1 to find equivocation expression. The token matching is done on the basis of words and parts of speech available in input text for predicting equivocation expression based on hierarchical LSTM model. The features for words which can be used in roman

May – June 2020
ISSN: 0193-4120 Page No. 574 - 585

Hindi using equivocation classes are given as training sample.

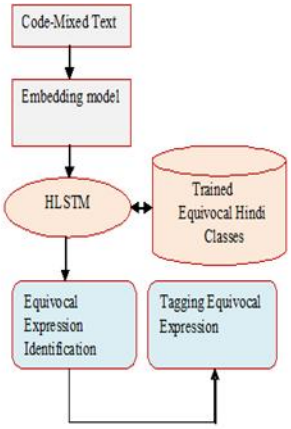

Figure 1: Hierarchical LSTM Framework

Table 1: Equivocation expression classes

Class	Examples
Puns in Headlines and Advertising	Happy Twenty Dineteen (Twenty)
Pun Time expression	3 gande (3 ghante/Hours)
Puns in Quotes	Namaske (Namaste) Trump
Pun- Past , present and future expression	pighlesaal (last year), is saal (this year), paglesaal (next year)
Pun- Quantifiers expression	Puch (few), bhaot (more), tanik (little)

To understand this, table1provides the details considering the equivocation classes to which the input word belongs to. Here forward and backward, LSTMs are used in the embedding layers. Finally in output layer softmax function is applied on the character vectors for labeling the token based on equivocation humor expression words. The hierarchical LSTM model treats the current token and the neighboring tokens for predicting the label for the current token by analyzing the context in which the word has been used. The proposal utilizes word-based embedding features and character-based context features for giving the final tag for equivocation expression words. The embedding model gives a comparative result for using character and word level analysis.

In this paper, equivocation expression in code mixed social media text recognition is proposed considering the following objectives.

- *Ambiguous expression identification in code mixed text.*
- *Tagging the identified ambiguous expression in proposed equivocal classes according to the sentence context.*

A. *Context lookup for identifying equivocation expression words.*

Here, we contextually look up the word using left context and the right context to get a list of all possible context variants of that word. We used the sentences from Hindi monolingual data from social media and Amul data [22] to build the language models for English and Hindi, respectively. It is desirable that the left and the right context of a pun word belong to different languages. We took the intersection of the left and right context of the word to predict the use of the word as English or Hindi. It helps us to calculate the similarity of the word at that position with all other words that can be available at that location depending on the context and classify the most appropriate word as the possible target word.

May-June 2020
ISSN: 0193-4120 Page No. 3672 - 3680

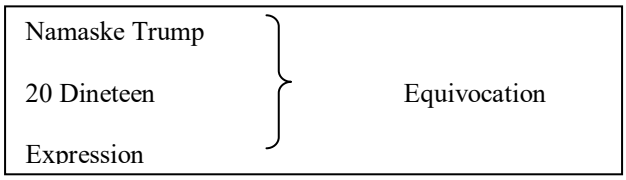

Figure 2: Equivocation context lookup

The word belonging to different is compared on a phonetic basis, we convert the word to WX notation , which denotes a standard way to represent Indian languages in the Roman script. However, OOV words cannot be converted to WX notation because of unavailability of phonetic transcription, therefore, we have used only roman Hindi words showing different use in Hindi context as illustrated in figure 2.

B. Embedding model

The next step is to process input data to the required word embedding model. Word embedding is word vectors of weights. The words can be represented in different dimensions and every word will have different weights in context to different dimensions. The meaning of word used as equivocation can easily be understood by the technique of embedding .This embedding technique of CBOW and Skip-gram technique will help to understand the context meaning of the word in connection with other word. Thus this technique helps in identifying the pun word used in the data set which has been used for evaluating the framework. CBOW technique is illustrated in figure 3.

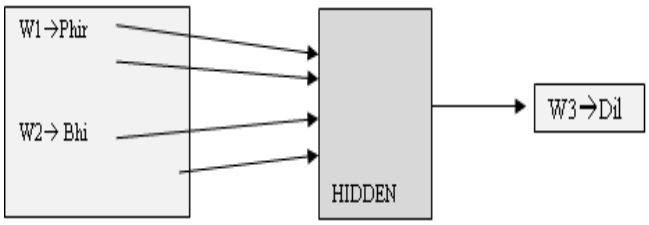

Figure 3: CBOW Technique

4Experimental Results

This section describes the evaluation scheme undertaken for the proposed model depicted in figure 1. The result description is presented by illustrating the use of dataset and its inference in this section. The dataset of code-mixed data used here is taken from the work of ICON-2016 and Amul advertisements [22]. It is Hindi-English text containing data of three social media texts along with advertisements hits of Amul containing headlines of advertisements.. The data description is illustrated in table 2. The data of these media texts has been labeled for HLSTM learning on four dimensions considering equivocation expression as base for classification. The four labeling parameters are H-EQ for Hindi words, E-EQ for English words, N-EQ for not recognized words and O for words belonging to other than Hindi and English. The labeling parameters and its corresponding description along with percentage are depicted in table 3.

Table 2: Dataset description

Code-mixed text	***No. of words***
ICON-2016	2631
Amul Advertisements	982
MSIR -2016	6139

Figure 4 provides the result analysis obtained for labeling accuracy on table 3 parameters. The four labels depicted in table 3 are evaluated on the data available in table 2. The labeling accuracy as per the f-score obtained is higher in case of Hindi words as compared to available English words. The figure 3 depicts the F-score obtained on the dataset.

May – June 2020
ISSN: 0193-4120 Page No. 574 - 585

Table 3 Labeling parameters

Label	Description	Hindi-English %
H-EQ	Equivocation words Hindi	57.76
EQ	Equivocation words English	20.41
N-EQ	Equivocation words not recognized	14.8
O	Other language words	7.03

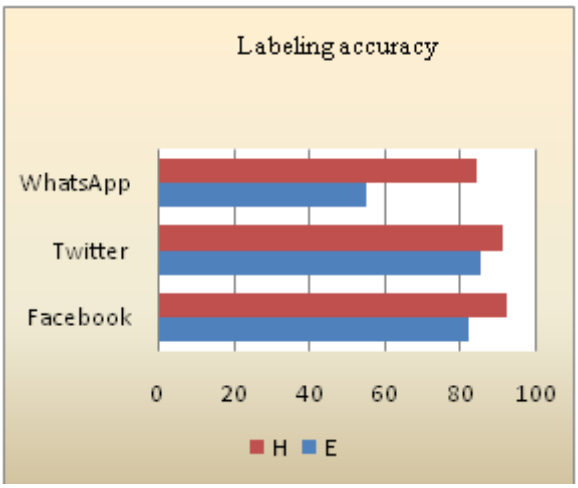

Figure 4: Labeling accuracy

The Figure 5 provides a result description of the model in regard to embedding technique used for context lookup. The figure describes the accuracy percentage of embedding model trained for character and words. The proposed HLSTM gives a clear separation for the different equivocation classes parameters, depicted in table 3. The result is compared with five different standard classification mechanisms based on equivocal pun expression words. The proposed HLSTM gives better accuracy as compared to other standard measures.

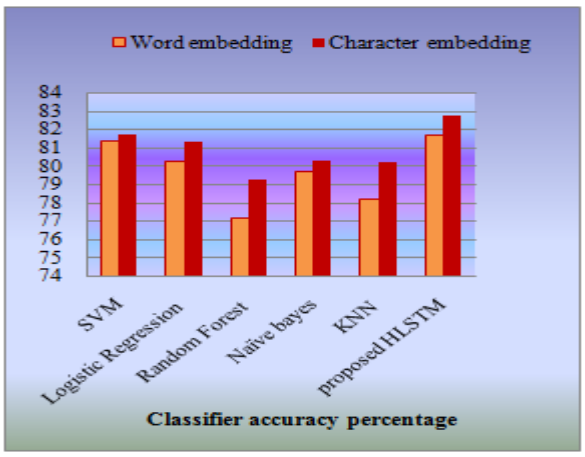

Figure 5: classifier accuracy comparison

Figure 6 illustrates the occurrences of words on which the experiment has been conducted, which comes under the category of pun expression words used in dataset. These words can be used in English as well as in Hindi for expressing the thoughts, but these are having different meanings when judged in context to other words in the sentence.

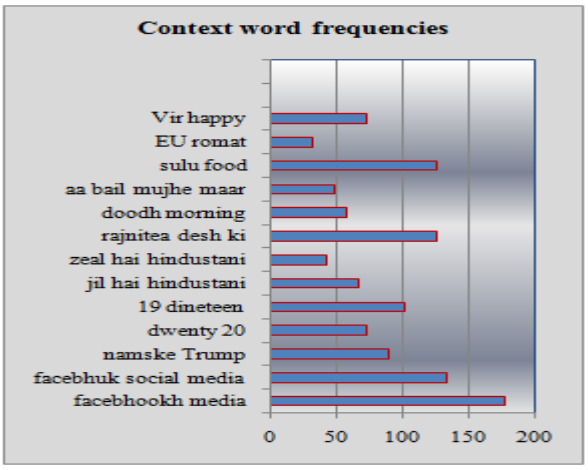

Figure 6: Equivocal expression words and their frequency

May-June 2020
ISSN: 0193-4120 Page No. 3672 - 3680

Figure 7 illustrates the accuracy of the framework as compared to other evaluation models like CRF, DT, SVM and rule based. The proposed voting mechanism gives more accuracy as compared to other benchmark standards.

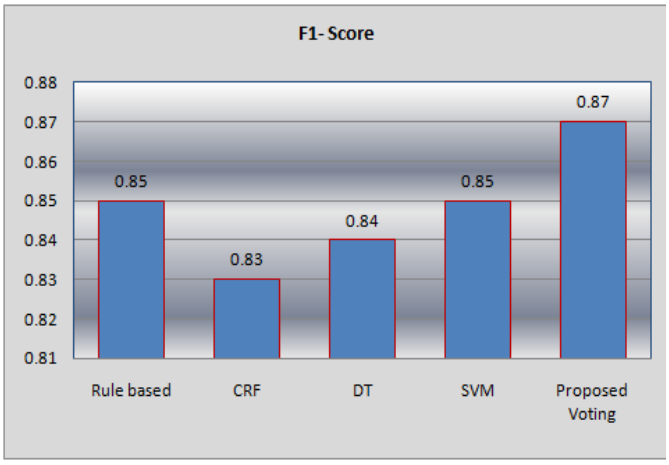

Figure 7: Accuracy representation

5Conclusions

The paper shows that equivocal expression retrieval is one of the prominent areas in information retrieval, where one can understand the context by identifying pun expressions. The learning strategy based on pre defined equivocal classes improves the labeling performance. This is one of the issues in language identification where equivocal expression or words need to be identified correctly in multi lingual environment. The multiple language use in code switching and code mixing environment is based on certain defined parameters like source of data, unstructured nature of data, switching and mixing percentages along with semantic relationship among the languages used for expression. We conclude that the equivocal expression words are often used on social media and advertisements according to the experiments conducted. It can be an interesting domain to investigate the patterns of words used to exhibit multiple contexts. The words which are used in Hindi as well as in English for expressions are needed to be examined for extracting equivocal dimension information. The paper provides state of art approaches for equivocal expression identification in Hindi. The paper illustrates different evaluation mechanism and comparison of proposed approach with standard approaches. It is being observed from the results that the proposed voting scheme gives better result in terms of F1 score. Our experiments were mainly on two language pairs based on bilingual learning approach. A HLSTM based learning approach has been proposed for classifying equivocal information in code-mixed text. The hierarchical LSTM system performs better as compared to other classifiers for pun words detection. The system can be enhanced to learn other patterns in data like hate or satire detection in social media text. In future, we would like to enhance this idea to extract information based on person's historical posts in regard to hate word detection.

References

1. Mamidi, Radhika. "Context and Humor: Understanding Amul advertisements of India." arXiv preprint arXiv:1804.05398 (2018).
2. King, Ben, and Steven Abney. "Labeling the languages of words in mixed-language documents using weakly supervised methods." Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2013..
3. Nguyen, Dong, and A. SezaDoğruöz. "Word level language identification in online multilingual communication." Proceedings of the 2013 conference on empirical methods in natural language processing. 2013.
4. Vyas, Yogarshi, et al. "Pos tagging of english-hindi code-mixed social media content." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014.

May – June 2020
ISSN: 0193-4120 Page No. 574 - 585

5. Das, Amitava, and BjörnGambäck. "Identifying languages at the word level in code-mixed indian social media text." Proceedings of the 11th International Conference on Natural Language Processing. 2014.
6. Sequiera, Royal, et al. "Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval." FIRE Workshops. Vol. 1587. 2015..
7. Rao, Pattabhi RK, and Sobha Lalitha Devi. "CMEE-IL: Code Mix Entity Extraction in Indian Languages from Social Media Text@ FIRE 2016-An Overview." FIRE (Working Notes). 2016.
8. Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg. "Hindi Roman Linguistic Framework for Retrieving Transliteration Variants using Bootstrapping." Procedia Computer Science 125,2018.
9. Ethiraj, Rampreeth, et al. "NELIS-Named Entity and Language Identification System: Shared Task System Description." FIRE Workshops. 2015.
10. Alekseev, Anton, and Sergey Nikolenko. "Word embeddings for user profiling in online social networks." Computación y Sistemas 21.2 ,203-226.2017.
11. Mandal, Soumil, and Anil Kumar Singh. "Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture." arXiv preprint arXiv:1808.07118 ,2018.
12. Aina, Laura, Kristina Gulordava, and Gemma Boleda. "Putting words in context: LSTM language models and lexical ambiguity." arXiv preprint arXiv:1906.05149,2019.
13. Bhattacharya, Paheli, PawanGoyal, and Sudeshna Sarkar. "Using Communities of Words Derived from Multilingual Word Vectors for Cross-Language Information Retrieval in Indian Languages." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 18.1,2019
14. Le, Ngoc Tan, et al. "Low-Resource Machine Transliteration Using Recurrent Neural Networks." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),2019.
15. Murthy, Rudra, Mitesh M. Khapra, and Pushpak Bhattacharyya. "Improving NER Tagging Performance in Low-Resource Languages via Multilingual Learning." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP),2018.
16. Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg. "Linguistic structural framework for encoding transliteration variants for word origin detection using bilingual lexicon." 2017 International Conference on Multimedia, Signal Processing and Communication Technologies (IMPACT). IEEE, 2017.
17. Gella S, Bali K, Choudhury M. "ye word kislangkahai bhai?" Testing the Limits of Word level Language Identification. InProceedings of the 11th International Conference on Natural Language Processing 2014 Dec (pp. 368-377).
18. Shekhar S, Sharma DK, Sufyan Beg MM. An effective cybernated word embedding system for analysis and language identification in code-mixed social media text. International Journal of Knowledge-based and Intelligent Engineering Systems. 2019 Jan 1;23(3):167-79.
19. Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg. "Computational linguistic retrieval framework using negative bootstrapping for retrieving transliteration variants." International Journal of Computational Vision and Robotics 10, no. 1 (2020): 79-101.
20. Ramrakhiyani, Nitin, and Prasenjit Majumder. "Approaches to temporal

May-June 2020
ISSN: 0193-4120 Page No. 3672 - 3680

expression recognition in Hindi." ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 14, no. 1 (2015): 1-22..

21. Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg. "Language identification framework in code-mixed social media text based on quantum LSTM—the word belongs to which language?." Modern Physics Letters B (2020): 2050086.
22. https://www.amul.com/m/amul-hits