Neural Network Pun Material Identification Framework based on Artificial Intelligence Learning

Shashi Shekhar
Department of Computer Engineering and Applications,
GLA University, Mathura, India
shashi.shekhar@gla.ac.in

Abstract. The Neural based artificial intelligence system is made linguistically intelligent through simulation model to identify pun material expressions from code mixed text. The text available on social media consist contents which are written in mixed script format and from these content puns word identification is a challenging task in this scenario. The retrieval of pun and its corresponding equivocation terms is very hard to retrieve from the transliterated text. The pun retrieval and its equivocation representation are widely used to present the opinions over the network applications. The work described in the paper gives the comparative view of different neural network and its corresponding learning and simulation techniques applied in the area of transliteration. The rule framed approach is presented which accepts the roman form text as input and as per the defined rules the system is developed to give the equivocation words available in the sentence. The evaluation measures used here to validate the hypothesis is based on statistical measures along with HLSTM learning model. Further the result is validated using the voting technique that can choose appropriate equivocation label which are not identifies by the learning model. The use of voting technique here is to provide an extra edge when the proposed approaches suggest incorrect tag against the pun word. The voting approach enhances the overall result accuracy with high precision value.

Keywords: Neural network, Material, AI, Mixing, Embedding

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 produce equivocal annotation of text in several languages. But very few works has been done in Indian language transliteration. The paper explores this area in retrieving 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 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 America

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 part of this paper is structured as follows: section 2 illustrates the state of art in equivocal retrieval. The methodology description is available in section 3. The description of dataset and its corresponding evaluation is being contained in section. The section 5 provides the summarization and conclusion with future path of work.

2 Related 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 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.
A mixed script based language identification task was conducted for Indian Languages [11]. Here the use of machine learning techniques using SVM classifier [12] was proposed. The technique of classification and its related machine learning techniques for English-Hindi [10][11] languages were taken care. This task gives the opportunity for the emerging researched to enhance their learning and understanding the domain area covered under transliteration field [13]. Various emotion identification models have been described based on learning-approach [13] for language mining [14].

The work consisting of ambiguity removal in code mixed text needs to be handled [15] with the help of learning models mostly in native language domain for finding effective context meaning. The following section of proposed methodology tries to model the ambiguity problem available in code mixed data using embedding technique [16]. The embedding model is more concerned for those words which are commonly used in both the languages. As it the most common research issue [19] in multilingual dataset [17] used in case of NER [18] extraction in transliterated domain. The paper [14] suggests the work related to NER in multilingual environment based on learning model and correspondingly provides a research issue in the field of retrieving phrases having equivocal opportunity in transliteration.

3 Proposed 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.

![Hierarchical LSTM Framework](image)

Figure 1: Hierarchical LSTM Framework

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 Hindi using equivocation classes are given as training sample.

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               | Namases (Namaste) Trump                      |
| Pun- Past, present and future expression | pighle din (last day), is din (present day), pagle saal (next year) |
| Pun-Quantifiers expression   | Puch (few), bhaot (more), tanik (little)     |

To understand this, table 1 provides 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 model HLSTM considers the neighboring words to the pivot word for suggesting the tag through context analysis features. The embedding technique used by the learning model which takes word and character embedding features for equivocation analysis.

The objective behind finding equivocation words available in mixed script text is summarized below.

- Ambiguous words retrieval in mixed text.
- Labeling the ambiguous words against the defined equivocal classes.
- Intent retrieval for using equivocal words.

A. Context retrieval equivocation expression.

The consideration of the terms that exist prior to pivot term and next to pivot word forming the things as (i+1) term and (j+1) term, are used as word features for context finding. The Amul data [22] containing monolingual format sentences are used to frame the linguistic learning model for context tracing. The constraint applied for context tracing is that the equivocal words must be accompanied by the terms available in the left and right of that word should belong to other languages. The concept of intersection needs to be computed to find whether the word has been used in Hindi or English context. This modeling helps in finding the similarity which further helps in finding appropriate pun words.

The evaluation measure for context identification for finding the ambiguity is measured with the left and right context in regard to the used pivot word. The evaluation explores the base of the data discussed in for starting the self-learning approach. The condition used here in this case is that the left and right words to the pivot word must belong to two different languages. The set theory intersection
concept is applied for tagging. The context word is retrieved on the basis of WX notation. Thus considering this scenario the roman Hindi words describing different context has been pointed out in figure 2.

![Equivocation context lookup](image)

**Figure 2: Equivocation context lookup**

**B. Embedding model**

The next step is to process input data against the embedding model. Word embedding is the weighted vectors of terms. The words can be represented in different dimensions and every term contains 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.

![CBOW Technique](image)

**Figure 3: CBOW Technique**

**4 Experimental 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.

![Data description](image)

**Table 2: Data description**

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.

Table 2: Dataset description

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

Table 3 Labeling parameters

| Tag   | Depiction                          | Hindi-English (Mixing) % |
|-------|------------------------------------|--------------------------|
| H-EQ  | Equivocation terms (Hindi)         | 58.21                    |
| EQ    | Equivocation terms (English)       | 21.05                    |
| N-EQ  | Equivocation words not recognized | 13.8                     |
| O     | Other language words               | 6.94                     |

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 accuracy of developed HLSTM provides enhanced accuracy when compared with standard measures.

The Figure 6 pointed below provides the detailing of occurrences of context words which has been selected for evaluation. These words are categorized as pun expression terms belonging to the defined data sample. These words exhibit equivocation features and are highly used for expressing opinions. Their contextual meaning can be different when correlated against the other terms available in the sentence.

The figure 7 depicts the framework accuracy when compared against the standard models like CRF, DT, SVM and rule based. The developed voting technique provides greater accuracy when compared with the CRF, DT, SVM standard models.
5 Conclusions

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 multilingual 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 which gives better results. The system can be enhanced to learn other patterns in data like hate or satire detection in social media text.

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. Seza Doğruöz. "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.
5. Das, Amitava, and Björn Gambäck. "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, Pawan Goyal, 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. S. Shekhar, D.K. Sharma, D.K. Agarwal, Y. Pathak, "Artificial Immune Systems-Based Classification Model for Code-Mixed Social Media Data, IRBM, 2020
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 kis lang ka hai 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 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