A Pun Identification Framework for Retrieving Equivocation Terms based on HLSTM Learning Model

Shashi Shekhar¹, Rishabh Sharma²
Department of Computer Engineering & Applications, GLA University, Mathura, India.
¹shashi.shekhar@gla.ac.in, ²rishabh.sharma_mtc19@gla.ac.in

Abstract. The artificial intelligence system is made linguistically intelligent through HLSTM model to identify pun expressions from code mixed text. The text available on social media domain is written in mixed script form 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 learning models and techniques applied in the area of transliteration for pun word retrieval. 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.

1. Introduction
Equivocal expressions now days are regularly used in the social media networks. Equivocal information processing is the challenging task in the transliterated domain. The equivocation helps in answering the ambiguous category of expressions by identifying the words based on its context meaning. Also it helps in illustrating the things in the 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. Language and word detection in the user-generated text is a very tedious task, where the language is un-known. Now a day, it is the difficult task where the text is available in the format of code-mixed sentence. This type of code-mixed data is common in the social media sites. 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 detection 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. To separate words that are syntactically similar in both languages, a special technique is required. The 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 mixed text is described in the below sentence:

Sentence 1: Namaske President Russia

| H/EQ | E   | NE |

Here, words in English as E, Roman Hindi are labeled as H, Named entity as NE and ambiguous equivocal as EQ. The ambiguous equivocal words describe the ambiguity expression in roman Hindi in sentence. In sentence 1 the word Namaske is labeled as EQ, it illustrates that the word denotes 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, which are marked as the 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

In this section provides the literature survey in recent techniques regarding temporal information in transliterated domain.

Code-mixing is an emerging area of research in the area 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 et al.[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 framework 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 the Recurrent Neural Network based transliteration model. The proposal presented in this paper is based on related research findings in the area of code-mixing. The intricacy to detect the language of the temporal expression words in code mixed data is presented in this work. The data which is present in the code-mixed format 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.

![Figure 1. HLSTM Framework.](image-url)

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

To understand this concept, Table 1 provides the in detail considering the equivocation classes to which the input word belongs to. Here backward and forward, 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 Hierarchical LSTM 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.

Table 1. Equivocation expression classes.

| Class                              | Examples                                |
|------------------------------------|-----------------------------------------|
| Puns in Headlines and Advertising  | Happy Twenty Nineteen                   |
| Pun Time expression                | 3 gande (3 ghante/ Hours)               |
| Puns in Quotes                     | Namaste Namaste Trump                   |
| Pun- Past, present and future expression | Pighle din(last day), is                |
|                                    | din(present day), pagle saal(next year) |

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.

3.1. 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 that belong to two 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.
3.2. 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.

![Figure 3. Embedding Technique.](image)

3.3. Experiment Evaluations

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.

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.](image)

| Code-mixed text       | No. Of words |
|-----------------------|--------------|
| ICON-2016             | 2823         |
| Amul Advertisements   | 1123         |
| MSIR-2016             | 5867         |
Table 3. Labeling parameters

| Tag | Description                                      | Hindi-English (Mixing) % |
|-----|--------------------------------------------------|--------------------------|
| H-EQ | Equivocation terms (Hindi)                       | 58.21                    |
| EQ   | Equivocation terms (English)                     | 21.05                    |
| N-EQ | Equivocation words not recognised                | 13.8                     |
| O    | Other language word                              | 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.

Figure 5. Classifier accuracy comparison.

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.

4. Conclusion and Future Work

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 predefined 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 2018 Context and Humor: Understanding Amul advertisements of India (arXiv preprint arXiv:1804.05398).
[2] King, Ben, and Steven Abney 2013 Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies pp.1110-1119.
[3] Nguyen, Dong, and A. Seza Doğruöz 2013 Word level language identification in online
multilingual communication. Proceedings of the 2013 conference on empirical methods in natural language processing p 857.

[4] Vyas, Yogarshi, et al. 2014 Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) p 974.

[5] Das, Amitava, and Björn Gambäck 2014 Identifying languages at the word level in code-mixed indian social media text p 169.

[6] Sequeria, Royal, et al. 2015 Overview of FIRE-2015 Shared Task on Mixed Script Information Retrieval FIRE Workshops. 1587 p 19.

[7] Rao, Pattabhi RK, and Sobha Lalitha Devi 2016 CMEE-IL: Code Mix Entity Extraction in Indian Languages from Social Media Text@ FIRE 2016-An Overview FIRE (Working Notes) p 289.

[8] Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg 2018 Hindi roman linguistic framework for retrieving transliteration variants using bootstrapping Procedia Computer Science 125 p 59.

[9] Ethiraj, Rampreeth, et al. 2015 NELIS-Named Entity and Language Identification System: Shared Task System Description FIRE Workshops p 43.

[10] Alekseev, Anton, and Sergey Nikolenko 2017 Word embeddings for user profiling in online social networks. Computación y Sistemas 21 p 203.

[11] Mandal, Soumil, and Anil Kumar Singh 2018 Language Identification in Code-Mixed Data using Multichannel Neural Networks and Context Capture arXiv preprint arXiv:1808.07118.

[12] Aina, Laura, Kristina Gulordava, and Gemma Boleda 2019 Putting words in context: LSTM language models and lexical ambiguity arXiv preprint arXiv:1906.05149.

[13] Bhattacharya, Paheli, Pawan Goyal, and Sudeshna Sarkar 2018 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 p 1.

[14] Shekhar, S., et al. 2020 Artificial Immune Systems-Based Classification Model for Code-Mixed Social Media Data IRBM.

[15] Murthy, Rudra, Mitesh M. Khapra, and Pushpak Bhattacharyya 2018 ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 18 p 1.

[16] Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg 2017 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) p 156.

[17] Mahrishi, M., Hiran, K. K., Meena, G., & Sharma, P. (2020). Machine Learning and Deep Learning in Real-Time Applications. IGI Global.

[18] Shekhar, Shashi, Dilip Kumar Sharma, and M. M. Sufyan Beg 2019 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 23 p 167.

[19] Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg 2020 "Computational linguistic retrieval framework using negative bootstrapping for retrieving transliteration variants." International Journal of Computational Vision and Robotics 10 p 79.

[20] Ramrakhiyani, Nitin, and Prasenjit Majumder 2015 Approaches to temporal expression recognition in Hindi ACM Transactions on Asian and Low-Resource Language Information Processing (TALLIP) 14 p 1.

[21] Shekhar, Shashi, Dilip Kumar Sharma, and MM Sufyan Beg 2020 Language identification framework in code-mixed social media text based on quantum LSTM—the word belongs to which language? Modern Physics Letters B Vol 34.

[22] Amul, https://www.amul.com/m/amul-hits. Accessed, Dec 2020.