International Journal of IT & Knowledge Management (ISSN: 0973-4414)
Volume 14 • Issue 2 pp. 22-25 Jan 2021 - June 2021 www.csjournals.com

Modeling Translation of Code Mixed English-Dogri Language

Shubhnandan S. Jamwal
PG Department of Computer Science and IT, University of Jammu, Jammu

Abstract: The code mixed languages are gaining importance day by day because almost every individual is comfortable in communication in the regional language. In this paper, a model is proposed for the identification and translation of the Dogri Roman and English code mixed language. The model is composed of the Input phase, Cleaning, Tokenization, Text Identification Engine, English word net, Dogri Romanized Word net, Translation of Dogri and English text and the final output phase.

Introduction

The common use of the internet and generation of the data on social media in the form of the text, audio and video is generating new challenges in every domain. The management and processing of the textual data is posing new challenges for the researchers of the NLP. The textual data processing is also posing big challenges for the automatic monolingual sentiment detection systems. Since, India is a multilingual country and a large part of its population speaks more than one language and speakers generally keep on switching between languages while communicating because of which the code is mixed and a text of mixed code are generated. Recently, on the social media a new form of communication is observed in the form of the multilingual and code-mixed texts.

Since the people are very comfortable while communicating in the regional language and when the languages are mixed the code mixed data is generated which exists in many different forms. The techniques of NLP application are analyzed to trace the words which can be used both in Indian languages and in English. In this paper, a model is proposed for the identification and translation of the Dogri Roman and English code mixed language.

Literature Review

P. Ranjan, B. Raja, R. Priyadharshini and R. C. Balabantaray[1] compared and experimented results of code mixed data with the normal text. We first identify the Languages present in social media text, in the case of code mixed data existing language detector fails to detect language at the word level because of the use of roman script to write their own language. So they bootstrapped language identification step and calculated the Code Mix Index to show the amount of code mix in the corpora. They used the RNNLM to create a language model of code mixed data as well as pen tree bank data and also evaluated the model. B. S. Sowmya Lakshmi and B. R. Shambhavi[2] focused on the problem of word-level LID for code-mixed data. Dataset collected contains English and Kannada code mixed sentences from social media posts. Experiments on various supervised classifiers are performed by embedding a dictionary module to handle word level code mixing. K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar and S. Saini[3] have worked on sentiment classification for one of the most common code-mixed language pairs in India i.e. Hindi-English. The conventional sentiment analysis techniques designed for a single language don't provide satisfactory results for such texts. They have proposed two approaches for better sentiment classification. They have proposed an Ensembling based approach which is based on hybridization of Naive Bayes, SVM, Linear Regression, and SGD classifiers. We have also developed a bidirectional LSTM based novel approach. The approaches provide quite satisfactory results for the code-mixed Hindi-English text. S. Shekhar, D. Kumar Sharma and M. M. Sufyan Beg [4] developed a method for identifying context words by classifying the intent for using the ambiguous word in code mixed sentence. A well-known hierarchical LSTM model is used in the reserch for context-based sub-word-level ambiguity detection to identify the language of the word. The work on Language Identification in the code-mixed text using character-based embedding for processing ambiguous word is a novel approach and shows promising results. A. Malte and S. Sonawane [5] analyzed code-mixed Hindi/English(Hinglish) text. Firstly, we generate a large scale code-mixed corpus that would aid in further research of code mixed text on social media. High-quality word embeddings are trained on this code-mixed text and demonstrated the efficacy of our proposed method by training machine learning models that improve upon the previous state-of-the-art using a much lighter and explainable architecture. S. Shekhar and D. K. Sharma [6] handcrafted rule based technique which is applied that accepts input in mixed script format and on the basis of illustrated rules, the system is suppose to extract temporal expressions from the sentence. The evaluation measures use rule-based approach validations along with the evaluations based on statistical measures. To further validate the results a voting technique is

applied that selects the most valid and suitable temporal tag for that word. The experimental results suggests that the voting method performs applied here gives the best result in context to accurate temporal tagging. S. Dutta,

T. Saha, S. Banerjee and S. K. Naskar [7] addressed the problem of text normalization, an often overlooked problem in natural language processing, in code-mixed social media text. The objective of the work presented was to correct English spelling errors in code-mixed social media text that contains English words as well as Romanized transliteration of words from another language, in this case Bangla. The targeted research problem also entails solving another problem, that of word-level language identification in code-mixed social media text. We employ a CRF based machine learning approach followed by post-processing heuristics for the word-level language identification task. For spelling correction, we used the noisy channel model of spelling correction. N. Sabri, A. Edalat and B. Bahrak[8], collected, label and thus created a dataset of Persian-English code-mixed tweets and then proceed to introduce a model which uses BERT pre-trained embeddings as well as translation models to automatically learn the polarity scores of these Tweets. A. Younas, R. Nasim, S. Ali, G. Wang and F. Qi[9] performed a sentiment analysis of code-mixed Roman Urdu and English social media text using deep learning models. The proposed work is independent of lexical normalization, language dictionary, and code transfer indication. They performed sentiment analysis using Multilingual BERT (mBERT) and XLM-RoBERTa (XLM-R) models. The results reveal that performance of XLM-R model with tuned hyperparameters for code-mixed Roman Urdu and English social media text is better than the mBERT model with F1 score of 71%.

As far as the development of the NLP applications for the Dogri Language is concerned the, models for generation of the verbs of the language [10] and the automatic generation of the Noun using Machine learning [11], English-Dogri MT using Moses [12], rule based transliteration system for detection of proper nouns for Dogri to English[13] and stemmer[14] for the Dogri language has been proposed, developed and implemented.

English-Dogri Mixed Code Model

The model will fetch the English words which are generally abbreviated in the communication on social medial like *hlo, omg* in addition to other English text. The model will also translate the Romanized Dogri text to the normal Devanagari text of the Dogri language which is mixed in the code. The proposed model which can be employed for the identification and conversion of the text is composed of the following phases:

1. Input Sentence
2. Cleaning of the sentence.
3. Tokenization
4. Text Identification Engine
   A. English word net
   B. Dogri Romanized Word net
   C. Translation of Dogri and English text
5. Output Regular Text

Input: The sentence which is the input contains the text of the English and Dogri in Romanized form. The proposed model can also be extended for the regular Dogri script in devanagari script. The input text in the form of the raw text format to the model is given for identification and proper translation of the code mixed language. The Dogri text is in Romanized format which we generally write while communicating on the social media.

International Journal of IT & Knowledge Management (ISSN: 0973-4414)
Volume 14 • Issue 2 pp. 22-25 Jan 2021 - June 2021 www.csjournals.com

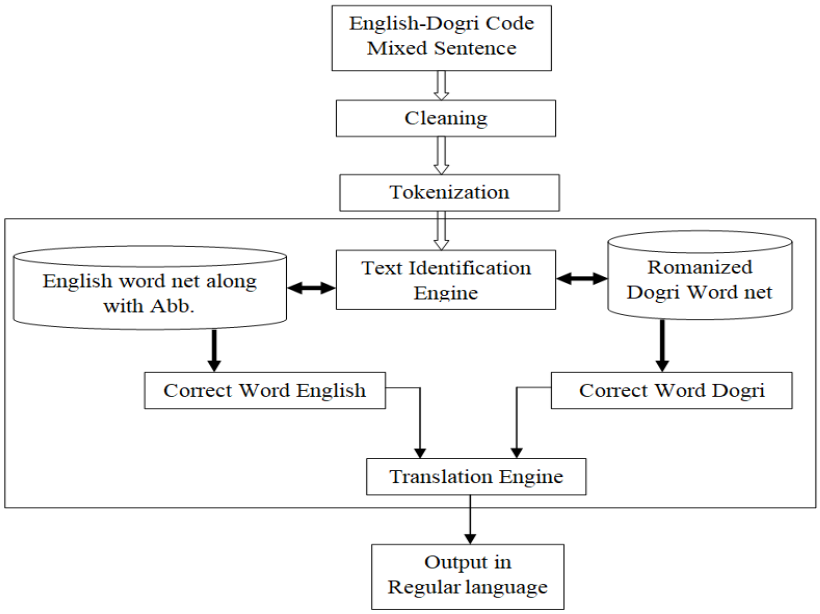

Fig 1: Identification of English Dogri Code Mixed Language model

Cleaning: The cleaning of the sentence involves the removal of the repeat words, by removing any extra characters, blank spaces or any other unwanted data such as emoji. This process is very important in order to have the precise output. The result of the output given is based on the correct input.

Tokenizer: The basic activity performed on the text before any further processing of the NLP activities is the tokenization. In this phase the input text in the code mixed language is converted into tokens. When the tokens are formed, they can be arranged in any of the data structures for any other NLP tasks.

Text Identification Engine: The text identification engine of the model is further composed of the English word net, Dogri Romanized Word net and their translations of Dogri and English text. This is the main phase which is composed of the sub phases and is responsible for the correct English words and corrects Dogri words translation from the Romanized text. The engine uses a data base of the commonly shot cut spelled words of the English like *omg, hlo* etc. On the other hand this engine also has the data base of the Dogri word net in Romanized-Dogri format. By taking the inputs from these databases, these sentences are converted into correct words of the English and Dogri. This model can also work in the supervised ML form for updating of the databases.

Output Regular Text: In the proposed model, the input is a sentence in a code mixed form of English and Romanized Dogri text. The output is displayed into normal text of the English and Devanagari script of the Dogri.

Conclusion

Since the generation of the social media text is gaining importance day by day and the text data generated by the individuals also creates a big challenge for the development of the NLP application. The identification of the language and the sentiments reflected in the code mixed data poses a big challenge. In this paper, a model is proposed for the identification of the Dogri and English language. The Dogri text is converted into Devanagari in the final output, which is generally written in the Romanized form. The words of the English language which are spelled wrongly are also translated in the standard form of the English.

References

[1] P. Ranjan, B. Raja, R. Priyadharshini and R. C. Balabantaray, "A comparative study on code-mixed data of Indian social media vs formal text," 2016 2nd International Conference on Contemporary Computing and Informatics (IC3I), 2016, pp. 608-611, doi: 10.1109/IC3I.2016.7918035.

[2] B. S. Sowmya Lakshmi and B. R. Shambhavi, "An Automatic Language Identification System for Code-Mixed English-Kannada Social Media Text," 2017 2nd International Conference on Computational Systems and Information Technology for Sustainable Solution (CSITSS), 2017, pp. 1-5, doi: 10.1109/CSITSS.2017.8447784.

[3] K. Yadav, A. Lamba, D. Gupta, A. Gupta, P. Karmakar and S. Saini, "Bi-LSTM and Ensemble based Bilingual Sentiment Analysis for a Code-mixed Hindi-English Social Media Text," 2020 IEEE 17th India Council International Conference (INDICON), 2020, pp. 1-6, doi: 10.1109/INDICON49873.2020.9342241.

[4] S. Shekhar, D. Kumar Sharma and M. M. Sufyan Beg, "Embedding Framework for Identifying Ambiguous Words in Code-Mixed Social Media Text," 2019 International Conference on contemporary Computing and Informatics (IC3I), 2019, pp. 59-63, doi: 10.1109/IC3I46837.2019.9055679.

[5] A. Malte and S. Sonawane, "Effective Distributed Representation of Code-Mixed Text," 2019 IEEE 16th India Council International Conference (INDICON), 2019, pp. 1-4, doi: 10.1109/INDICON47234.2019.9028960.

[6] S. Shekhar and D. K. Sharma, "H-LSTM Framework for Temporal Information Retrieval in Code-Mixed Social Media Text," 2020 8th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO), 2020, pp. 1315-1319, doi: 10.1109/ICRITO48877.2020.9197932.

[7] S. Dutta, T. Saha, S. Banerjee and S. K. Naskar, "Text normalization in code-mixed social media text," 2015 IEEE 2nd International Conference on Recent Trends in Information Systems (ReTIS), 2015, pp. 378-382, doi: 10.1109/ReTIS.2015.7232908.

[8] N. Sabri, A. Edalat and B. Bahrak, "Sentiment Analysis of Persian-English Code-mixed Texts," 2021 26th International Computer Conference, Computer Society of Iran (CSICC), 2021, pp. 1-4, doi: 10.1109/CSICC52343.2021.9420605.

[9] A. Younas, R. Nasim, S. Ali, G. Wang and F. Qi, "Sentiment Analysis of Code-Mixed Roman Urdu-English Social Media Text using Deep Learning Approaches," 2020 IEEE 23rd International Conference on Computational Science and Engineering (CSE), 2020, pp. 66-71, doi: 10.1109/CSE50738.2020.00017.

[10] 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.

[11] Shubhnandan S. Jamwal, Named Entity Recognition for Dogri using ML, International Journal of IT & Knowledge Management, Jan-June 2017 Volume-10, Number-2 pp. 141-144.

[12] Singh, Avinash & Kour, Asmeet & Jamwal, Shubhnandan. (2016). English-Dogri Translation System using MOSES. Circulation in Computer Science. 1. 45-49. 10.22632/ccs-2016-251-25.

[13] Tejpaul S. Sasan, Dr. Shubhnandan S. Jamwal, Transliteration of Name Entities Using Rule Based Approach, International Journal of Advanced Research in Computer Science and Software Engineering, ISSN: 2277 128X, Volume 6, Issue 6, June 2016.

[14] 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