Language Identification and Normalization of Code-Mixed English and Punjabi Text

Neetika Bansal[1], Vishal Goyal[2], Simpel Rani[3]
[1,2]Department of Computer Science, Punjabi University, Patiala, India
[3]Department of Computer Science and Engineering, YCOE, Talwandi Sabo, India
[1]sunshine_neetika@yahoo.com,[2]vishal.pup@gmail.com
[3]simpel_jindal@rediffmail.com

Abstract

Code mixing is prevalent when users use two or more languages while communicating. It becomes more complex when users prefer romanized text to Unicode typing. The automatic processing of social media data has become one of popular areas of interest. Especially since COVID period the involvement of youngsters has attained heights. Walking with the pace our intended software deals with Language Identification and Normalization of English and Punjabi code mixed text. The software designed follows a pipeline which includes data collection, pre-processing, language identification, handling Out of Vocabulary words, normalization and transliteration of English- Punjabi text. After applying five-fold cross validation on the corpus, the accuracy of 96.8% is achieved on a trained dataset of around 80025 tokens. After the prediction of the tags: the slangs, contractions in the user input are normalized to their standard form. In addition, the words with Punjabi as predicted tags are transliterated to Punjabi.

1 Introduction

India is the second largest online market in the world, ranked after China with over 560 million internet users.[1] Facebook is the largest social network with more than 2.7 billion monthly active users followed by WhatsApp, Twitter, and Instagram. Plenty of social media platforms are available nowadays but the most popular in context to Indic languages are Facebook, etc.

(Gold, 1967) was earliest to develop tools for automatic language identification by preparing a Language Learnability Model. (Gumperz, 1962; Scotton, 1997) stated that code-switching occurs when a user switches between different languages in written or spoken a single instance. Nowadays, code switching and code mixing are used alternatively. Word level language identification is one of the challenging tasks as code mixing takes place at word level, at sentence level and even at sub word level in an utterance. Challenges posed are numerous and keep changing with the intensity of languages in the utterance; still due to paucity of data the groundwork remains challenging.

2 Methodology

The main focus of current research is to identify the language of every word in the English Punjabi code mixed. The first and foremost task for developing the system is collection of Code Mixed Social Media Text (English- Punjabi) using API twitter threads for **Twitter**, selecting some prolific users comments for **Facebook** as data and some student community prolific users chat for **Whatsapp** followed by cleaning of extracted data.

(Gamback and Das, 2014) used Hindi, English, acronyms, universal tags along with Code Mixing Index. (Vyas et al., 2014) used English, Hindi and rest tags. In addition to the language tags (Chittaranjan et al., 2014) discussed named entity and ambiguous tags. (Gundapu and Mamidi, 2018)

[1]https://www.statista.com/statistics/262966/number-of-internet-users-in-selected-countries/

Proceedings of the 17th International Conference on Natural Language Processing: System Demonstrations, pages 30–31
Patna, India, December 18 - 21, 2020.

experimented with different possible combinations of available words, context and Part of Speech (POS) tags. (Jamatia et al., 2018) have used Hindi, English, universal, named entity, acronym, mixed and undefined tags.

The dataset used in the current research consists of 80025 tokens (after preprocessing) which have been tagged as en (English), pb (Punjabi), univ (Universal), mixed (mixing of two languages inside a word), ne (Named Entity), acro (Acronyms), rest (none of earlier mentioned tags). A supervised model is trained with Conditional Random Fields (CRF) which calculates the conditional probability of output tags given the values assigned to the input nodes. The features used are contextual features, capitalization features, special character features, character N-Gram features and lexicon features. After applying five-fold cross validation on the corpus, the accuracy of 96.8% is achieved on a trained dataset of around 80025 tokens.

In social media text people use creativity in spellings rather than traditional words. The deviation of text can be categorized as acronyms, slangs, misspellings, use of phonetic spellings etc. Contractions like hasn't- has not, ma'am-madam etc. which are handled by mapping. Plenty of common English words e.g. lyk – like, feb-February, gm- gud morning have changed their existence on social media. A dictionary of such out of vocabulary has been maintained in order to normalize them. A transliterated dictionary for the code mixed data contains transliterated pairs of Romanized text and its Punjabi equivalent. ***e.g*** kithey- ਕਿਥੇ, Janam – ਜਨਮ *etc.*

After the prediction of the tags: the slangs, contractions in the user input are normalized to their standard form words with Punjabi as predicted tags are transliterated to Punjabi language.

3 Results

On the bilingual English-Punjabi data set the CRF baseline approach reports an accuracy of 97.24 % with F1-score 96.8 % on the English-Punjabi language pair. Table 1 shows precision, recall and F1-score with different tag categories used in the system.

Tag Categories	Precision	Recall	F1-Score
acro	0.85	0.77	0.81
en	0.96	0.96	0.96
mixed	0.00	0.00	0.00
ne	0.88	0.92	0.90
pb	0.97	0.99	0.98
rest	0.87	0.54	0.67
univ	0.99	0.94	0.97
Accuracy			0.968

Table 1: CRF System Performance (Accuracy and F1-score) on the Test Data (%)

References

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali and Monojit Choudhury. 2014. Word-level language identification using CRF: Code-switching shared task report of MSR India system. In *Proceedings of The First Workshop on Computational Approaches to Code Switching,* pages 73-79.

Bjorn Gamback and Amitava Das. 2014. On Measuring the Complexity of Code-Mixing. In *Proceedings of the 11th International Conference on Natural Language Processing*, pages 1-7, Goa.

E. Mark Gold. 1967. Language Identification in the Limit. Information and Control: 10(5), pages 447-474.

John J. Gumperz. 1962. Discourse strategies. Cambridge University Press, Vol.1, Cambridge, UK.

Sunil Gundapu and Radhika Mamidi. 2018. Word Level Language Identification in English Telugu Code Mixed Data. In *PACLIC*.

Anupam Jamatia, Bjorn Gamback, and Amitava Das. 2018. Collecting and annotating Indian social media code-mixed corpora. In *International Conference on Intelligent Text Processing and Computational Linguistics.* pages 406-417, Springer.

Carol Myers-Scotton. 1992. Constructing the Frame in Intrasentential Codeswitching. Multilingua-Journal of Cross-Cultural and Interlanguage Communication, 11(1):101-128.

Yogarshi Vyas, Spandana Gella, Jatin Sharma, Kalika Bali and Monojit Choudhury. 2014. POS tagging of English-Hindi code-mixed social media content. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 974-979.