Evaluation of Word Embeddings for Semantic and Syntactic Tasks in Code-Mixed Social Media Text

ABHISHEK SRIVASTAVA[1,*], ROHAN DATTA[1], AND DOLLY SHARMA[3]

[1] *Student, Department of Computer Science & Engineering, School of Engineering*
[2] *Assistant Professor, Department of Computer Science & Engineering, School of Engineering*
[*] *Corresponding author: as878@snu.edu.in*

Compiled March 14, 2019

Multilingual speakers often switch between multiple languages while speaking and writing on social media. This phenomenon of using linguistic units from different languages in a single utterance or sentence is called code-mixing. The switching linguistic unit can be words or short extracts. While most NLP tasks have seen major improvements in monolingual setting, which can be attributed to the introduction of pre-trained word embeddings such as word2vec, it is not true for code-mixed text. In this project, we aim to develop and evaluate different word-level embeddings on natural language processing tasks on Hindi-English code-mixed social media text. We aim to compare the performance of bilingual embedding approaches on semantic and syntactic tasks.
Haan Haan

1. INTRODUCTION

Word embeddings are useful for a variety of natural language processing tasks, as they allow to generalize the system on much larger corpora than the annotated dataset for the task. They can significantly improve semantic and syntactic tasks such as text classification, named entity recognition and sentiment analysis, etc. However, such monolingual embeddings fail to perform satisfactorily on code-mixed datasets.

In this study, we focus on bilingual word embeddings where words from two languages are embedded into the same space. While offering same advantages as their monolingual counterpart, they also have a potential, yet unexplored application. They can be used in the processing of code-mixed language.

We first preprocess the web-crawled code-mixed data using different normalization techniques, prepare bilingual embeddings, and finally, evaluate their performance on different tasks. Following are some instances from a Twitter corpus of Hindi-English code-mixed texts also transliterated in English.

T1: "Finally India away series jeetne mein successful ho hi gayi :D"

Translation: "Finally India got success in winning the away series :D"

T2: "This is a big surprise that Rahul Gandhi congress ke naye president hain."

Translation: "This is a big surprise that Rahul Gandhi is the new president of Congress."

2. DATASET CREATION

We retrieve a set of about 300 Hindi words that are frequently used both in writing and speaking. The list of words can be easily found online on websites such as Wikitionary.com and have the roman transliterated version also present beside the Devanagari form. We reduce the transliterated words to the basic roman form without any accent. We then manually go through the list and intuitively remove the words we think are not used very often on social media such as Ahinsa, etc. After this step, we are left with a word count of about 200.

We then scrape 1,000 tweets containing code-mixed text using an API based on Twitter Advanced Search[1] for every word in the list. High frequency words are used as keywords for scraping the tweets. So, basically, every tweet has at least one high frequency word in it. The final dataset contains about 200,000 unique tweets.

3. WORD NORMALIZATION

The web-scraped tweets are cleaned, and all URLS and special characters are removed from them. All unique words are extracted from the dataset to create a dictionary. However, code-mixed data is full of variations. For example, *kaise* (How?) can be written as both *kaise* and *kese*. Also, due to informal nature of

[1] https://github.com/gutfeeling/twitass

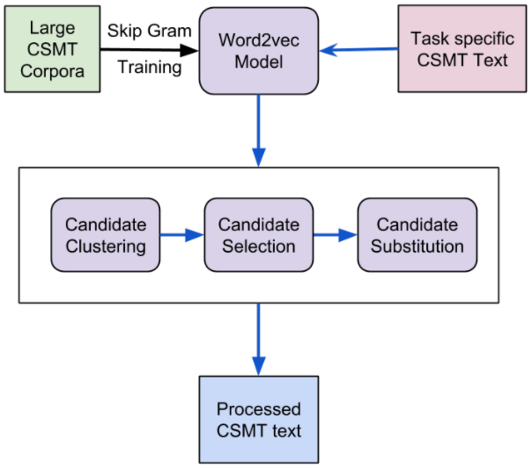

Fig. S1. Flow Diagram for Normalization of Words

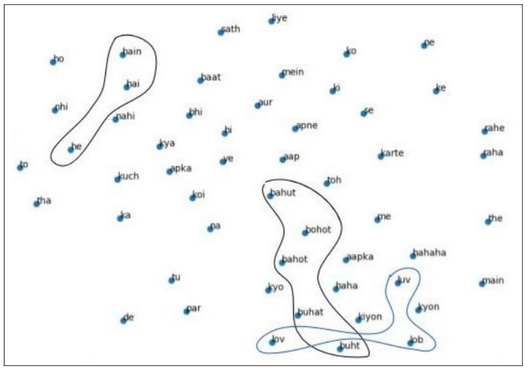

Fig. S2. Example of the clustering technique

social media, people often misspell, creating more tokens for the same base form of the word. Therefore, the number of tokens is very high when in fact semantically the actual unique words would be fewer in number.

In order to normalize different variations of words belonging to the same context, we use a clustering-based method [Singh et al., 2018] on the entire corpus. The entire process flow is described in Figure 1.

Using the approach proposed by [Singh et al. 2018], we train a word2vec model on the scraped tweets. We also train a FastText model and prepare representations for the words. Since both the embeddings model are monolingual, they treat the code-mixed corpus as one language and prepare the embeddings. We test the model and visualize the words in 2D Euclidean space. As it is evident in Fig. 2, the words which are used in similar context appear close together. Using this technique, we can identify the clusters and map the words to the most commonly used term.

We further add more constraints such as: the Levenshtein distance should be less than a threshold, and the Soundex equivalent of the word should be close to the final form. We use both or one of these constraints to get the final set of normalized words.

4. PREPARING EMBEDDINGS

In recent times, there has been some interest in bilingual word embeddings, where words from two languages are embedded into the same space. Here in this section, we outline how we go about preparing bilingual embeddings for our scraped, normalized corpus.

We use two embeddings preparation technique: (i) BiCVM - Hermann and Blunsom (2014) and (ii) Random translation replacement - Gouws and Søgaard (2015).

We prepare the embeddings using these techniques, and compare their performance with the monolingual embeddings prepared using word2vec and FastText models.

5. TASKS

We plan to evaluate our prepared embeddings on semantic and syntactic natural language processing tasks.

A. Semantic Tasks

Semantic tasks deal with the inherent meaning of the word or sentence as opposed to its grammar or syntax (Syntactic). For example, sentiment analysis is a semantic task because the meaning of the document as a whole is considered to decide the polarity of the text. We plan to evaluate the embeddings on sentiment analysis as well as sarcasm detection.

B. Syntactic Tasks

Syntactic tasks deal with the way the sentence is written. For example, named entity recognition is a syntactic task because the ordering of words and grammar matter. We plan to evaluate the embeddings on named entity recognition as well as POS tagging.

6. CONCLUSION

We demonstrate for the first time that code-mixed text with languages originally in different scripts (such as Devanagari-Roman) require a more nuanced pre-processing approach before the preparation of bilingual word embeddings. In this report, we have outlined the progress we have made so far in the project and have also given a broad view of our path ahead. We plan to proceed in the described manner and record our progress in the paper. The project could initiate a novel approach for working with social media corpora which is full of code-mixed texts and effectively improve many natural language processing tasks.

REFERENCES

1. Singh, Rajat & Choudhary, Nurendra & Shrivastava, Manish. "Automatic Normalization of Word Variations in Code-Mixed Social Media Text." (2018)
2. Hermann, Karl Moritz and Phil Blunsom. "Multilingual Models for Compositional Distributed Semantics." *ACL* (2014).
3. Gouws, Stephan & Søgaard, Anders. "Simple task-specific bilingual word embeddings."*Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (2015).