Project Proposal
Language Detection in Code-mixed Social Media Text

Alefiya Naseem, Paarth Kotak, Siddhesh Acharekar
CS6120 NLP Fall 2018
naseem.a@husky.neu.edu
kotak.p@husky.neu.edu
acharekar.s@husky.neu.edu

I. Introduction

The past few years have seen social media websites and networks penetrate greatly into multilingual societies. The linguistic and syntactic freedom that these environments on social media platforms afford their users lead to many deviations from what is the accepted norm of communication in a language. One of these deviations happens when users switch between languages i.e. code-mixing. Code-mixed text adopts the vocabulary and grammar of two or more languages and often forms new structures based on its users. Studies on the phenomenon can be found in linguistics literature [1] which majorly discuss conversational and sociological motivations behind code-mixing as well as it's linguistic nature. This proves as a road block for NLP tasks on social media text like sentiment analysis. Presence of phonetic typing of words leads to error in interpretation and hence mistakes in sentiment detection. Code-mixing is traditionally distinguished between *inter-sentence* (two or more languages in the same sentence), *intra-sentence* (two or more sentences with different languages), and, *intra-word* (code mixing within a single word) code mixing. This project attempts to evaluate previous word-level language identification experiments [2], suggest improvements to those, then build our own models and finally compare them.

We intend on incorporating more features: both extracted from our data and external sources to better language identification, an approach hitherto untried in code-mixed literature. We will see ahead that the simplest unsupervised dictionary-based approach is outperformed by supervised word-level classification without contextual clues and later that it is crucial to take context into consideration [2]. Further, we set forth on using word-level classification algorithms, building multiple n-gram generative language models, and deep learning-based classification algorithms for sequential text data. The next stage after language identification is Language Translation from one language to another and Sentiment Analysis, however, both those tasks by themselves are full-fledged domains which will require more research. Provided we attain convincing, and sizeable language detection models and if time permits, we would like to do a further investigation into the problem of Language Translation and Sentiment Analysis for code-mixed data in this research.

II. Related work

The problem of language identification has been investigated for half a century [3] and that of computational analysis of code switching for several decades [4], but there has been less work on automatic language identification for multilingual code-mixed texts. Prior research related to this task has mainly been focused on monolingual texts [5] due to their large-scale availability. However, in multilingual societies like India, usage of code-mixed languages (among which Hindi-English is most prominent) is quite common for conveying opinions online. Initial studies on Chinese-English code mixing [6] indicated that mainly linguistic motivations were triggering the code mixing in those highly bilingual societies. Reference [7] showed that Facebook users tend to mainly use inter-sentential switching over intra-sentential, and report that 45% of the switching was instigated by real lexical needs, 40% was used for talking about a topic, and 5% for content clarification. Turning to the work on automatic analysis of code mixing, there have been some studies on detecting code mixing in speech. Reference [8] tries to predict the points inside a set of spoken Spanish-English sentences where the speakers switch between the two languages. Other studies have looked at code mixing in different types of short texts, such as information retrieval queries and SMS messages. They explore several language identification approaches, including a Naive Bayes classifier for individual word-level classification and sequence labelling with Conditional Random Fields trained with Generalized Expectation criteria [9], which achieved the highest scores.

Another very recent work on this topic is [10]. They report on language identification experiments performed on Turkish and Dutch forum data. Experiments have been carried out using language models, dictionaries, logistic regression classification and Conditional Random Fields. They find that language models are more robust than dictionaries and that contextual information is helpful for the task.

III. Dataset

Considering the claim that code-mixing is most frequent among *young, multilingual* speakers [11], we have decided to use an Indian student community in the 20-30 age group as our data source. To save data collection and annotation time we have taken data with permission from previous researchers on this topic [2] who scraped a Facebook group and 11 of its users posts and comments. This data was then annotated by a team of

4 who were proficient in all 2 languages of our corpus (English and Hindi). Annotators were supplied with 4 basic tags (viz. *sentence, fragment, inclusion* and *wlcm*(word-level code mixing)) to annotate different levels of code mixing. Under each tag, five attributes were provided, viz. *English (en), Hindi (hi), Mixed (mixd), Universal (univ) and Undefined (undef).* The attribute *univ* is associated with *symbols, numbers, emoticons* and *universal expressions* (e.g. *hahaha, lol*). The attribute *undef* is specified for a sentence or a word for which no language tags can be attributed or cannot be categorized as *univ*. In addition, annotators were instructed to annotate named entities separately. Posts having words in any language other than Hindi or English were discarded, leaving us with 12936 code-mixed posts in total. Here is a description of the 4 basic tags:

Sentence (sent): This tag refers to a sentence and is used to mark *inter-sentential* code mixing.

Fragment (frag): This refers to a group of foreign words, grammatically related, in a sentence. This marks *intra-sentential code mixing.*

Inclusion (incl): This is a foreign word in a sentence or in a fragment used frequently in a native language.

Word-level code mixing (wlcm): Smallest unit of code-mixing which captures intra-word code mixing.

IV. Methodology

With 12936 posts, the first step will be to split the data into training, validation and testing sets. Next, we intend on trying the following approaches for word level language detection:

1) Unsupervised dictionary based language detection using the *British National Corpus (BNC), SEMEVAL 2013 Twitter Corpus* and *LexNormList,* word level classification without context and with context.

2) Supervised classification Algorithms with Naïve Bayes as baseline, and others like SVM and RandomForests.

3) Considering the entire training set as one corpus we want to know if we can capture sequential dependencies through n-gram language modelling.

4) Previous methods have not delved into a lot of feature extraction from the text. We intend on investigating more on that front with features like presence in dictionaries, major language used in sentence, etc. and experiment if that improves any of the aforementioned models.

5) We would like to explore if Deep Learning techniques could improve our classification task. Instead of manually passing the entire dataset to the neual network we could learn the most important features by Autoencoders and use those as our inputs to classifer algorithms to know if there are contextual features we cant manually capture.

Lastly, if the task of language identification is solved convincingly and with enough time to spare we would like to explore further into translation or sentiment analysis of codemixed data.

V. Evaluation

Since this is a classification task, we intend on using classification accuracy, sensitivity and specificity to quantify performance of our models.

References

[1] Lesley Milroy and Pieter Muysken, editors. 1995. One speaker, two languages: Cross-disciplinary perspectives on code-switching. Cambridge University Press.

[2] Code Mixing: A Challenge for Language Identification in the Language of Social Media Utsab Barman, Amitava Das, Joachim Wagner and Jennifer Foster CNGL Centre for Global Intelligent Content, National Centre for Language Technology Proceedings of The First Workshop on Computational Approaches to Code Switching, pages 13–23, October 25, 2014, Doha, Qatar, 2014.

[3] E Mark Gold. 1967. Language identification in the limit. Information and control , 10(5):447–474.

[4] Aravind K. Joshi. 1982. Processing of sentences with intra-sentential code-switching. In J. Horeck´y, editor, Proceedings of the 9th conference on Computational linguistics - Volume 1 (COLING'82) , pages 145–150. Academia Praha, North-Holland Publishing Company.

[5] Anna Schmidt and Michael Wiegand. 2017. A survey on hate speech detection using natural language processing. In Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media , pages 1–10.

[6] David C. S. Li. 2000. Cantonese-English codeswitching research in Hong Kong: a Y2K review. World Englishes , 19(3):305–322.

[7] Taofik Hidayat. 2012. An analysis of code switching used by facebookers: a case study in a social network site. Student essay for the study programme "Pendidikan Bahasa Inggris" (English Education) at STKIP Siliwangi, Bandung, Indonesia

[8] Thamar Solorio and Yang Liu. 2008b. Part-of-speech tagging for English-Spanish code-switched text. In Proceedings of the Conference on Empirical Methods in Natural Language Processing , pages 1051–1060. Association for Computational Linguistics.

[9] Gideon S. Mann and Andrew McCallum. 2008. Generalized expectation criteria for semi-supervised learning of conditional random fields. In Proceedings of ACL-08: HLT , pages 870–878, Columbus, Ohio. Association for Computational Linguistics.

[10] Dong Nguyen and A. Seza Do˘gru˘oz. 2013. Word level language identification in online multilingual communication. In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP 2013), pages 857–862, Seattle, Washington, USA. Association for Computational Linguistics.