Translating Code-Mixed Tweets: A Language Detection Based System

Shruti Rijhwani, Royal Sequiera, Monojit Choudhury, Kalika Bali
Microsoft Research
Bangalore, India
{t-shruri,a-rosequ,monojitc,kalikab}@microsoft.com

Abstract

We demonstrate a system for the machine translation of code-mixed text in several Indian and European languages. We first perform word-level language detection and matrix language identification. We then use this information and an existing translator in order to translate code-mixed tweets into a language of the user's choice.

Keywords: code-mixing, language detection, machine translation

1. Introduction

Code-mixing is the alteration between two or more languages at the sentence, phrase, word, or morpheme level. It is prevalent in multilingual communities around the world (Gumperz, 1982). Although code-mixing has traditionally been observed in spoken language, informal text-based interaction on social media has seen the advent of code-mixed language in text as well (Bali et al., 2014; Das and Gambäck, 2014; Solorio et al., 2014).

In India, a significant percentage of the population fluently communicates in more than one language and often mixes these languages in both speech and text. Bali et al. (2014) found that 17.2% of the posts on public Facebook pages from India are code-mixed.

Machine Translation of Social Media text is a difficult problem (Carrera et al., 2009; Hassan and Menezes, 2013; Galinskaya et al., 2014), and the fact that many multilingual users use code-mixed language on social media, compounds this problem manifold. Most existing Natural Language Processing (NLP) techniques and systems, including Machine Translation, are designed for monolingual language data and break down in the presence of code-mixed text. With the pervasiveness of code-mixing in India, it becomes necessary to create language systems that can process mixed language data.

With this view, we propose a machine translation system for code-mixed text in several Indian and European languages.

2. System Architecture

Figure 1 describes the architecture of our system. Given a Twitter handle, several tweets belonging to the corresponding handle are collected. The user chooses one of these tweets to be translated. The modules involved in translation in the order of their working are as follows:

1. **Language detector:** The language detector identifies the language of each word in a given tweet. We use a Hidden Markov Model trained on Twitter data from our set of languages. The word-level language identification accuracy is around 95%.

2. **Matrix language identifier:** A matrix language of an utterance is defined as the language that governs the grammar of the utterance (Joshi, 1982). This module selects the language that the majority of words belong to as the matrix language of the tweet. Our initial observations showed that this simple heuristic works well in practice.

3. **Translate to matrix language:** We conducted an analysis of code-mixed data translations by a state-of-the-art machine translation system, which showed that translation quality is improved if the input is first translated to the matrix language. This module translates the tweet to the matrix language using the Bing Translator API.

4. **Translate to destination language:** Once the tweet is in its matrix language, it is translated to the destination language specified by the user by the Bing Translator API.

3. Conclusion

We present a demo of a system that given a stream of tweets, can identify code-mixed tweets, identify the language of mixing, and translate them into a single language using Machine Translation. While the current system is used to translate tweets, the underlying models, techniques and architecture can be used across any code-mixed text. Our future work will focus on expanding the set of languages as well as other scenarios.

4. Bibliographical References

Bali, K., Sharma, J., Choudhury, M., and Vyas, Y. (2014). I am borrowing ya mixing? an analysis of english-hindi code mixing in facebook. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics.

Carrera, J., Beregovaya, O., and Yanishevsky, A. (2009). Machine translation for cross-language social media.

Das, A. and Gambäck, B. (2014). Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on Natural Language Processing*.

Galinskaya, I., Gusev, V., Mescheryakova, E., and Shmatova, M. (2014). Measuring the impact of spelling errors on the quality of machine translation. In *Proceed-*

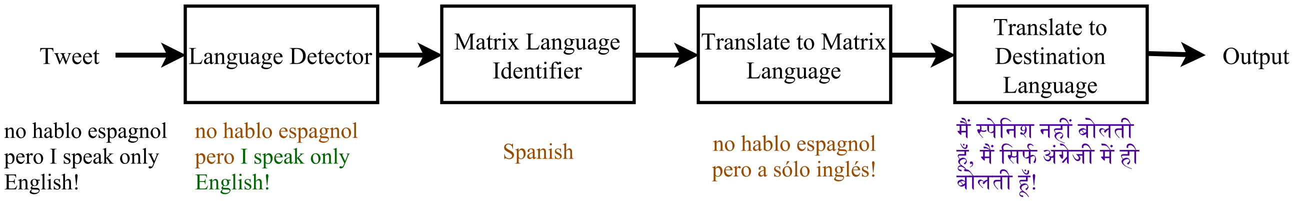
Tweet
Language Detector
Matrix Language Identifier
Translate to Matrix Language
Translate to Destination Language
Output
no hablo espagnol pero I speak only English!
no hablo espagnol pero I speak only English!
Spanish
no hablo espagnol pero a sólo inglés!
मैं स्पेनिश नहीं बोलती हूँ, मैं सिर्फ अंग्रेजी में ही बोलती हूँ!

Figure 1: System Architecture

ings of the Ninth International Conference on Language Resources and Evaluation (LREC).

Gumperz, J. J. (1982). *Discourse strategies*, volume 1. Cambridge University Press.

Hassan, H. and Menezes, A. (2013). Social text normalization using contextual graph random walks. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL).*

Joshi, A. K. (1982). Processing of sentences with intrasentential code-switching. In *Proceedings of the 9th International Conference on Computational Linguistics (COLING).*

Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Gohneim, M., Hawwari, A., AlGhamdi, F., Hirschberg, J., Chang, A., et al. (2014). Overview for the first shared task on language identification in code-switched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics.