Generalized and Unsupervised Word-Level Language Detection

Shruti Rijhwani*
Language Technologies Institute
Carnegie Mellon University
srijhwan@cs.cmu.edu

Royal Sequiera*
University of Waterloo
Waterloo, Canada
rdsequie@uwaterloo.ca

Monojit Choudhury
Microsoft Research
Bangalore, India
monojitc@microsoft.com

Kalika Bali
Microsoft Research
Bangalore, India
kalikab@microsoft.com

Abstract

Code-switching is the use of multiple languages in a single sentence or document. Word-level language detection is necessary for processing code-switched text. State-of-the-art language identification systems are restricted to either detecting a single language for the entire document or identifying code-switching between two specific languages. These fail in real-world scenarios as multilingual text input rarely has a priori information on the languages used. We present a novel unsupervised word-level language detection technique for code-switched text in an arbitrarily large set of languages. With monolingual and code-switched tweets in seven languages, our model shows a 74% relative error reduction in word-level language labeling with respect to competitive baselines. Further, results are comparable with systems trained using large amounts of annotated data.

1 Introduction

In stable multilingual societies, communication often features fluid alteration between two or more languages – a phenomenon known as *code-switching*[1] (Gumperz, 1982; Myers-Scotton, 1993). It has been studied extensively in linguistics, primarily as a speech phenomenon (Poplack, 1980; Milroy and Muysken, 1995; Auer, 2013).

no/es me/es lebante/es ahorita/es cuz/en I/en felt/en como/es si/es me/es kemara/es por/es dentro/es ! :o Then/en I/en started/en getting/en all/en red/en , I/en think/en im/en allergic/en a/es algo/es

Figure 1: Word-Level Language Detection. The languages are *es* (Spanish) and *en* (English)

However, the growing popularity of computer mediated communication, particularly social media, has resulted in language data in the text form which exhibits code-switching, among other speech-like characteristics (Crystal, 2001; Herring, 2003; Danet and Herring, 2007; Cardenas-Claros and Isharyanti, 2009).

With the large amount of online content generated by multilingual users around the globe, it becomes necessary to design techniques to process and analyze mixed language. Language detection (LD) is a prerequisite to several language processing tasks. For code-switched tweets, token-level LD is required because several languages can be mixed within a single input. Figure 1 shows an example of word-level language detection for a Spanish-English code-switched tweet.

Most state-of-the-art LD systems detect a single language for an entire document or sentence (Cavnar and Trenkle, 1994; Dunning, 1994; Lui and Baldwin, 2012). Such methods often fail to detect code-switching, which can occur within a sentence. In recent times, there has been some effort to build word-level LD for code-switching between a specific pair of languages (Nguyen and Dogruöz, 2013; Elfardy et al., 2013; Solorio et al., 2014; Barman et al., 2014). However, user-

* This work was done when the authors were affiliated with Microsoft Research.

[1] This paper uses the terms 'code-switching' and 'code-mixing' interchangeably.

generated text (e.g., on social media) generally has no prior information of the languages being used causing these systems to fail. Further, these methods rely on large amounts of data annotated with word-level language labels – this is expensive and challenging for a large number of languages and their code-switched combinations.

This paper proposes a novel technique for word-level LD that generalizes to an arbitrarily large set of languages. Training is done without annotated data, while achieving accuracies comparable to language-restricted systems trained with large amounts of labeled data. Specifically, we improve on previous work in the following ways:

- The number of supported languages can be arbitrarily large
- Any number of the supported languages can be mixed within a single input
- The languages in the input do not need to be known a priori
- Any number of language switch points are allowed in the input.
- No manual annotation is required for training

2 Method

The intuition behind the model is simple – a person who is familiar with k languages can easily recognize (and also understand) the words when any of those languages are code-switched, even if s/he has never seen any mixed language text before. Analogously, *is it possible that monolingual language models, when combined, can identify code-switched text accurately?*

To formalize the task, let $\mathcal{L} = \{l_1, l_2, \ldots, l_k\}$ be a set of k natural languages. These are the languages the model can identify. We also define *universal tokens* like digits, emoticons, URLs, and punctuation, which do not belong to any specific natural language, labeled by $\mathcal{X}_\mathcal{L} = \{xl_1, xl_2, \ldots, xl_k\}$. Using xl_i instead of generically labeling all such tokens xl allows preserving linguistic context when memoryless models like Hidden Markov Models (HMM) are used.

Imagine we have k HMMs, where the ith HMM has two states l_i and xl_i. Each state can label a word. The HMMs are independent, but they are tied to a common start state s and end state e, forming a word-level LD model for monolingual text. Now, we add transitions from $l_i \rightarrow l_j$, where $i \neq j$. This HMM, shown in Fig. 2, is capable of labeling code-switched text between any of the k languages. Fig. 2 shows three languages, however, the number of languages can be arbitrarily large.

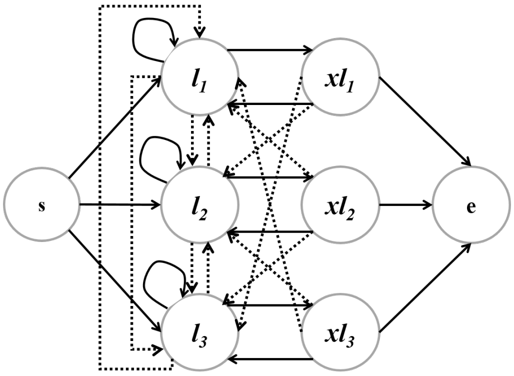

Figure 2: Hidden Markov Model LD System. $s \rightarrow xl_i$ and $l_i \rightarrow e$ transitions omitted for clarity.

The parameters to be learned are the transition and emission matrices of the HMM. However, obtaining word-level annotated monolingual and code-switched data for training is expensive and nearly infeasible for a large number of languages. Instead, using existing language identification systems, we automatically create weakly-labeled monolingual text (set $\mathcal{W}$) and use it to initialize the matrices. The emission values for states are initialized with the respective Kneser-Ney smoothed (Chen and Goodman, 1999) probabilities for all n-grams in $\mathcal{W}$.

Possible transitions for each monolingual HMM are $l_i \rightarrow l_i$, $l_i \rightarrow xl_i$ and $xl_i \rightarrow l_i$. We do not have the $xl_i \rightarrow xl_i$ transition, because preprocessing concatenates successive *universal tokens*. This does not change the output as the tokens can easily be separated, but is a useful simplification for the model. The transition values for are initialized by the probability of transitions in $\mathcal{W}$. The model supports code-switching by the addition of transitions $l_i \rightarrow l_j$, and $xl_i \rightarrow l_j$, for all $i \neq j$. We initialize these new edges with a small probability π, which we call the code-switch probability. This is a hyperparameter tuned on a validation set.

We reestimate the initialized parameters using the EM-like *Baum-Welch* algorithm (Welch, 2003) over a large set of unlabeled text.

After training, the Viterbi decoding algorithm is then used with the HMM parameters to perform word-level LD on an input sentence.

	Accuracy
Baseline	0.853
Our Model	**0.963**

Table 1: Accuracy on the Test Set

3 Results

We conduct experiments on monolingual and code-switched tweets in seven languages – Dutch (*nl*), English (*en*), French (*fr*), German (*de*), Portuguese (*pt*), Spanish (*es*) and Turkish (*tr*). The validation and test data contain 33981 and 58221 tokens respectively, with monolingual tweets and tweets that code-switch each of these languages with English (labeled by six annotators), *es-en* data from Solorio et al. (2014) and *nl-tr* data from Nguyen and Dogruöz (2013).

The weakly-labeled $\mathcal{W}$ is created using an existing LD system (Gella et al., 2013). This contains 700K tweets, 100K for each of the seven languages. The unlabeled set used for reestimating parameters contains 574K tweets in their natural distribution, collected using the Twitter API[2].

We compare with a strong dictionary-based baseline, since no general-purpose word-level LD system exists. The baseline finds the minimum number of languages required to explain all words in the input, based on monolingual dictionaries and word frequencies in those languages. The word-level LD accuracy comparison is shown in Table 1. Our model performs significantly better than the baseline.

We also compare with state-of-the-art results on existing annotated datasets (Solorio et al., 2014; Nguyen and Dogruöz, 2013). On *en-es*, Al-Badrashiny and Diab (2016) reports an $F1$-score of 0.964; our system obtains 0.978. Nguyen and Dogruöz (2013) report 0.976 Acc on the *nl-tr* test set. We obtain a less competitive 0.936. Notably, unlike our system, both these models use large amounts of annotated data for training and are restricted to detecting only two languages.

4 Error Analysis

We conducted a thorough analysis of the errors from all languages. *nl* words marked as *en* account for nearly 14% of all errors. We observe that most of these are actually *en* words with *nl* gold-standard labels, which is the convention used by dataset creators (Nguyen and Dogruöz, 2013). In fact, our model labels these *en* words accurately.

We also observe that 13% and 7% of the word-level errors come from confusion between *es-en* and *nl-tr* respectively. A large number of these are named entities (`Twitter`, `Orhan Pamuk`) and ambiguous words (`a`, `no`). 41% of the *es-en* errors are undetected single words language switches, likely because the model is inclined to remain in the same language for unseen words. It must be noted that over 70% of all single-word code-switching in es-en, including ambiguous and misspelled instances, are correctly labeled.

Further, confusion between *pt* and *es* contribute 10% of the total errors because these languages have several common words.

Over-detection, that is, detecting languages that are not present in the tweet, accounts for a sizable fraction (39.6%) of all word-level errors. Several of these overlap with the errors discussed previously and the causes are similar – named entities, misspelled words and ambiguous words.

Not detecting a language switch causes 7.7% of the word-level errors. 93.5% of these occur with fragments containing less than 3 words. As noted earlier, the model generally performs well for such short phrases and these errors typically contain out-of-vocabulary and ambiguous words.

5 Conclusion

We present a technique for word-level language detection for an arbitrarily large set of languages that is completely unsupervised, that significantly out-performs existing baselines and is comparable to supervised models. Our system can be deployed not only for better NLP tools for mixed-language data, but also for rigorous and large-scale sociolinguistic studies on code-switching from across the globe, which we plan to explore as future work. Further, we would like to extend our system to several more languages as well as consider a subword-level model to handle unknown words and misspelling more appropriately.

Acknowledgments

We are grateful to Prof. Shambavi Pradeep and her students from BMS College of Engineering for assisting with data annotation, and to Ashutosh Baheti from Microsoft Research (Bangalore, India) for help with data organization.

[2] `https://dev.twitter.com/rest/public`

References

Mohamed Al-Badrashiny and Mona Diab. 2016. Lili: A simple language independent approach for language identification. In *Proceedings of the 26th International Conference on Computational Linguistics (COLING). Osaka, Japan.*

Peter Auer. 2013. *Code-switching in conversation: Language, interaction and identity*. Routledge.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*.

Mónica Stella Cardenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Between yes, ya, and si a case study. In *The JALT CALL Journal, 5*.

William B Cavnar and John M Trenkle. 1994. N-gram-based text categorization .

Stanley F Chen and Joshua Goodman. 1999. An empirical study of smoothing techniques for language modeling. *Computer Speech & Language* 13(4):359–393.

David Crystal. 2001. *Language and the Internet*. Cambridge University Press.

Brenda Danet and Susan Herring. 2007. *The Multilingual Internet: Language, Culture, and Communication Online*. Oxford University Press., New York.

Ted Dunning. 1994. *Statistical identification of language*. Computing Research Laboratory, New Mexico State University.

Heba Elfardy, Mohamed Al-Badrashiny, and Mona Diab. 2013. Code switch point detection in arabic. In *Natural Language Processing and Information Systems*, Springer, pages 412–416.

Spandana Gella, Jatin Sharma, and Kalika Bali. 2013. Query word labeling and back transliteration for indian languages: Shared task system description .

John. J. Gumperz. 1982. *Discourse strategies*. Cambridge University Press, Cambridge.

Susan Herring, editor. 2003. *Media and Language Change*. Special issue of Journal of Historical Pragmatics 4:1.

Marco Lui and Timothy Baldwin. 2012. langid.py: An off-the-shelf language identification tool. In *In Proceedings of the ACL 2012 System Demonstrations*. pages 25–30.

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

Carol Myers-Scotton. 1993. *Dueling Languages: Grammatical Structure in Code-Switching*. Claredon, Oxford.

Dong Nguyen and A. Seza Dogruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.

Shana Poplack. 1980. Sometimes Ill start a sentence in Spanish y termino en espaol. *Linguistics* 18:581–618.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Gohneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, et al. 2014. Overview for the first shared task on language identification in code-switched data. *Proceedings of The First Workshop on Computational Approaches to Code Switching* .

Lloyd R Welch. 2003. Hidden markov models and the baum-welch algorithm. *IEEE Information Theory Society Newsletter* 53(4):10–13.