Codeswitching language identification using Subword Information Enriched Word Vectors

Meng Xuan Xia
McGill University
3480 University, Rm. 318
Montreal, Quebec H3A 0E9, Canada
meng.xia@mail.mcgill.ca

Abstract

Codeswitching is a widely observed phenomenon among bilingual speakers. By combining subword information enriched word vectors with linear-chain Conditional Random Field, we develop a supervised machine learning model that identifies languages in a English-Spanish codeswitched tweets. Our computational method achieves a tweet-level weighted F1 of 0.83 and a token-level accuracy of 0.949 without using any external resource. The result demonstrates that named entity recognition remains a challenge in codeswitched texts and warrants further work.

1 Introduction

Codeswitching (CS) is a widely observed phenomenon in social media. Solorio et al. (2014) define CS broadly as a communication act, whether spoken or written, where two or more languages are being used interchangeably. Codeswitching is common among bilingual speakers, both in speech and in writing. Identifying the languages in a codeswitched input is a crucial first step before applying other natural language processing algorithms.

The second shared task, like the previous one (Solorio et al., 2014), challenges the participants to develop computational method for identifying the language of each word in a dataset of codeswitched tweets. For each word in the source, the goal is to identify whether the word is *lang1*, *lang2*, *mixed*, *other* (punctuation, emotion and everything that is not a word in neither *lang1* nor *lang2*), *ambiguous*, *ne* (named entity), *unknown* or *fw* (foreign word). *Lang1* and *lang2* are the two languages presented in a codeswitched language pair. There are two language pairs available in this shared task: Modern Standard Arabic-Arabic Dialects (MSA-DA) and English-Spanish (EN-ES). An example of token language identification is shown in Table 1.

Token	Gold standard label
Hay	lang2
Dios	ne
,	other
I	lang1
'm	lang1
tired	lang1

Table 1: Example of label assignments for a English-Spanish codeswitched tweet

Our work covers only the EN-ES language pair. We use FastText (Bojanowski et al., 2016) to train a subword information enhanced word vectors model from the datasets of the shared task. We then use these vectors and, in addition, custom features extracted from the words to train a linear-chain Conditional Random Field model that predicts the language label of each word. Our system requires only the dataset provided by the shared task, without any external resource. The final model scores 0.83 in weighted tweet-level F1 and 0.949 in overall token-level accuracy.

2 Related Work

Seven systems were submitted for the previous shared task. The system with the highest prediction result for the EN-ES language pair scores 0.822 in F-measure (Bar and Dershowitz, 2014). Solorio

Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 132–136,
Austin, TX, November 1, 2016.

et al. (2014) shows that Conditional Random Field (CRF) and Support Vector Machines (SVM) are the most popular supervised machine learning algorithms used for this task. Similar approach have been found outside of the share tasks, including Nguyen and Dogruoz (2013). All of these systems rely on external resources, while our system relies only on data prepared by the shared task.

Aside from the last shared task, previous work on identifying languages emphasizing on word-level identification includes Yamaguchi and Tanaka-Ishii (2012; VRL (2014; Zubiaga et al. (2015). There are also studies on multilingual documents, focusing on inter-sentential codeswitching (King and Abney, 2013; Singh and Gorla, 2007).

Previous work on language models that encode syntactic constraints from codeswitching theory includes Li and Fung (2013; Li and Fung (2014). These models require a parser for the codeswitched input, while our work only requires word-level tokenization.

3 System Description

Our system contains two steps to identify tokens in a codeswitched input.

In the first step, we use FastText (Bojanowski et al., 2016) to train a subword information enhanced skipgrams word vectors model, using the tweets presents in the train and the dev dataset. Word vectors are vector representations of the words learned from their raw form, using models such as Word2Vec (Mikolov and Dean, 2013). When used as the underlying input representation, word vectors have been shown to boost the performance in NLP tasks (Socher et al., 2013). FastText word vectors are used instead of standard word2vec because FastText can obtain representations of out-of-vocabulary words by summing the representations of character n-grams. This feature is particularly useful because the size of the training data is relatively small. We expect the test dataset to contain words not found in the training dataset. Another motivation for using FastText word vectors is for its ability to take into account morphological information. We trained a skipgram FastText word vector representation model from the train datasets, using the default parameters provided by FastText (size of word vectors: 100, size of the context window: 5, number of epochs 5, minimal number of word occurences: 5, max length of word ngram: 1 and loss function: ns)[1].

In the second step, We use supervised machine learning to train a Linear-Chain Conditional Random Field (CRF) (Lafferty et al., 2001) classifier that predicts the label of every token in the order given by the EN-ES token assigner. CRF is naturally suited for sequence labeling tasks and it has been shown to perform well in previous work on language identification tasks (King and Abney, 2013; Chittaranjan et al., 2014; Lin et al., 2014). We use CRFsuite(Okazaki, 2007) in our experiment.

3.1 Feature Extraction

For each token, we extract three types of features: word features, spelling features and intra-word features.

3.1.1 Word features

Word features contain the word vector representation of the current token and that of the token directly before and after the current one. We use the word vector model trained in the first step to obtain the word vector of each token. Word vectors of out-of-vocabulary tokens are automatically predicted in FastText by summing up the vector representations of character n-grams.

3.1.2 Spelling features

The following boolean features are extracted from the current token and that of the token directly before and after the current one:

- whether the token capitalized
- whether the token is all uppercase
- whether the token is all lowercase
- whether the token contains an uppercase character anywhere but in the beginning
- whether the token is alphanumeric
- whether the token contains any punctuation
- whether the token ends by an apostrophe

[1] https://github.com/facebookresearch/fastText#full-documentation

- whether the token contains no roman alphabets
- whether the token is in the beginning of the sentence
- whether the token is in the end of the sentence

Capitalization is a strong indicator of a proper noun, hence a named entity. However, this is not always the case with social media texts, where grammatical rules are not always followed. The boolean feature of the lack of roman alphabets is added because of our observation on the training data – most tokens classified as *other* do not have roman alphabets.

We also considered adding a boolean feature of whether the token contains Spanish-only accented characters (i.e. í, ú, é). However, it did not positively impact the prediction performance when tested against the dev dataset. This is possibly due to that social media users are more casual with spelling and replace accented characters with their non-accented counterpart. For example, both the correct spelling *así como*[2] and the incorrect spelling *asi como* are found in the training data.

3.1.3 Intra-word features

In contrast to English, Spanish is a morphologically rich language, demonstrating a complicated suffixed-based derivational morphology (Bar and Dershowitz, 2014). To capture repeating prefixes and suffixes that characterize each language, we extract the first 1-3 and the last 1-3 characters of each token as intra-word features. These features have also been shown to help predicting named entities and tokens labelled as *other*. For example, hashtags (tokens that begin with a # sign) are often named entities; twitter handles always begin with an @ sign.

4 Experiment

The shared task maintains three sets of dataset: a training dataset, a development dataset and a testing dataset. Each dataset contains rows of token extracted from EN-ES codeswitched tweets and the respective gold standard label for each token. We train an unsupervised FastText word vectors model using the training dataset. Then we train a supervised CRF model using the same dataset. The supervised model is validated on the development dataset by evaluating standard metrics: precision, recall and F-measure of the predictions (Powers, 2011). We make hyper-parameter tuning to the CRF classifier using grid search.

To verify that all our features were contributing to the model's performance, we also did an ablation study, removing one group of features at a time.

Using the final model which consist of all the features, we compute predictions for the tokens in the testing dataset and submit the result to the workshop as final result.

5 Result and Analysis

5.1 Feature ablation

Table 2 shows the F1 scores on the dev dataset resulting from training with each group of feature removed. Note that although the removal of word features has no impact on the overall average F1, we decide to keep it because of the extra boost in performance it provides for named entities.

5.2 Final model performance

Our final model, when evaluated on the test dataset, has a tweet-level performance of 0.83 in weighted F1 as shown in Table 3. In terms of token-level performance, our model has an overall token accuracy of 0.949. The detailed metrics for each label are shown in Table 4.

We observe that the model is not able to predict mixed, foreign word, ambiguous and unknown labels. This is due to the lack of sufficient training data for these labels.

Our model has relatively low precision and recall with the NE labels. This suggests that our system is weak in recognizing named entities. While Bar and Dershowitz (2014) describe improvements of name entity recognition by using a gazetteer of proper nouns, our system did not benefit from having such a gazetteer. In fact, when validating on the development set, having such a gazetteer feature introduces over-fitting and decreases the overall accuracy of the model. The result suggests that named entity recognition remains a challenge in the context of codeswitched text.

[2]Meaning *as well as* in English

Features	lang1	lang2	ne	other	unk	Average
All	**0.965**	**0.945**	**0.355**	0.995	0.095	**0.946**
– Word features	**0.965**	0.944	0.331	**0.997**	0.058	**0.946**
– Intra-word	0.957	0.936	0.228	0.981	**0.126**	0.936
– Spelling	0.963	0.944	0.353	0.989	0.093	0.944

Table 2: Feature ablation study. F1 on dev dataset after training with individual feature groups removed. The F1 for *mixed*, *fw* and *ambiguous* are all 0, hence omitted in this table. The average F1 is micro-averaged, taking into account all eight labels. The number of tokens for each label are the following: lang1: 16813, lang2: 8653, ne: 740, mixed: 14, ambiguous: 70, unk: 133, other: 6853 and fw: 0

Monolingual F1	Codeswitched F1	Weighted F1
0.86	0.79	0.83

Table 3: Tweet-level performance – there are in total 4626 codeswitched tweets and 6090 monolingual tweets.

Label	Recall	Precision	F1
lang1	0.879	0.866	0.873
lang2	0.968	0.962	0.965
other	0.993	0.994	0.993
ne	0.313	0.421	0.359
mixed	0	0	0
fw	0	0	0
ambiguous	0	0	0
unknown	0	0	0

Table 4: Token-level performance – the number of tokens for each label are the following: ambiguous: 4, lang1: 16944, lang2: 77047, mixed: 4, ne: 2092, fw: 19, other: 25311, unknown: 25, Total : 121446.

6 Conclusion

We participated in the shared task of the second codeswitching workshop by creating a supervised machine learning model that identifies the languages given a English-Spanish codeswitched input. Our model uses FastText to train a subword information enhanced word vectors model from the shared task datasets. In addition to these vectors, we add custom features extracted from the words to train a linear-chain Conditional Random Field model that predicts the language label of each word. Our system uses only the training data provided by the shared task and requires no external resource. The final model scores 0.83 in weighted tweet-level F1 and 0.949 in overall token-level accuracy. Our result suggests that named entity recognition remains difficult for codeswitched text and warrants future work.

Acknowledgment

The author thank Jackie Chi Kit Cheung for mentoring and the two anonymous reviewers for providing detailed constructive feedback.

References

Kfir Bar and Nachum Dershowitz. 2014. The Tel Aviv university system for the Code-Switching Workshop Shared Task. *EMNLP 2014*, page 139.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching Word Vectors with Subword Information. *arXiv preprint arXiv:1607.04606*.

Gokul Chittaranjan, Yogarshi Vyas, and Kalika Bali 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.

Ben King and Steven P Abney. 2013. Labeling the Languages of Words in Mixed-Language Documents using Weakly Supervised Methods. In *HLT-NAACL*, pages 1110–1119.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of the eighteenth international conference on machine learning, ICML*, volume 1, pages 282–289.

Ying Li and Pascale Fung. 2013. Improved mixed language speech recognition using asymmetric acoustic model and language model with code-switch inversion constraints. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 7368–7372. IEEE.

Ying Li and Pascale Fung. 2014. Code switch language modeling with functional head constraint. In *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 4913–4917. IEEE.

Chu-Cheng Lin, Waleed Ammar, Lori Levin, and Chris Dyer. 2014. The CMU submission for the shared task on language identification in code-switched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 80–86.

T Mikolov and J Dean. 2013. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*.

Dong-Phuong Nguyen and A Seza Dogruoz. 2013. Word level language identification in online multilingual communication. Association for Computational Linguistics.

Naoaki Okazaki. 2007. Crfsuite: a fast implementation of conditional random fields (crfs).

David Martin Powers. 2011. Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation.

Anil Kumar Singh and Jagadeesh Gorla. 2007. Identification of languages and encodings in a multilingual document. In *Building and Exploring Web Corpora (WAC3-2007): Proceedings of the 3rd Web as Corpus Workshop, Incorporating Cleaneval*, volume 4, page 95. Presses univ. de Louvain.

Richard Socher, John Bauer, Christopher D Manning, and Andrew Y Ng. 2013. Parsing with Compositional Vector Grammars. In *ACL (1)*, pages 455–465.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steven Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirschberg, Alison Chang, 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*, pages 62–72. Citeseer.

NICTA VRL. 2014. Accurate language identification of twitter messages. In *Proceedings of the 5th Workshop on Language Analysis for Social Media (LASM)@ EACL*, pages 17–25.

Hiroshi Yamaguchi and Kumiko Tanaka-Ishii. 2012. Text segmentation by language using minimum description length. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 969–978. Association for Computational Linguistics.

Arkaitz Zubiaga, Iñaki San Vicente, Pablo Gamallo, José Ramom Pichel, Iñaki Alegria, Nora Aranberri, Aitzol Ezeiza, and Víctor Fresno. 2015. TweetLID: a benchmark for tweet language identification. *Language Resources and Evaluation*, pages 1–38.