Codeswitching Detection via Lexical Features using Conditional Random Fields

Prajwol Shrestha
Verscend Technologies Pvt. Ltd.
prajwol.shrestha@verscend.com

Abstract

Half of the world's population is estimated to be at least bilingual. Due to this fact many people use multiple languages interchangeably for effective communication. At the Second Workshop on Computational Approaches to Code Switching, we are presented with a task to label codeswitched, Spanish-English (ES-EN) and Modern Standard Arabic-Dialect Arabic (MSA-DA), tweets. We built a Conditional Random Field (CRF) using well-rounded features to capture not only the two languages but also the other classes. On the Spanish-English(ES-EN) classification task, we obtained weighted F1-score of 0.88 on the tweet level and an accuracy of 96.5% on the token level. On the MSA-DA classification task, our system managed to obtain F1-score of 0.66 on tweet level and overall token level accuracy of 74.7%.

1 Introduction

It is very common to find people adept in two or more languages (Ansaldo et al., 2008). During interactions and conversations in real life as well as in social media, multilingual people tend to switch between languages in written as well as verbal communication. Hale (2014) found that over 10% of Twitter users tweet in multiple languages and they are in general more active users than their monolingual counterparts. Modupeola (2013), in his research of an effective way of teaching English to Nigerian students, finds codeswitching from English to Nigerian by teachers as a necessity for generating interest, elaborating word meanings and the transfer of knowledge of the secondary language in an efficient manner. Codeswitching phenomenon is more prevalent in multi-cultural societies, where means of communication is not primarily in a single language (Cheng and Butler, 1989; Auer, 2013). It can also be attributed to the huge amount of exposure to other languages due to social media, TV shows and movies. Codeswitched text identification is often the first step for text to speech translation, automatic speech recognition and sentiment analysis.

Our paper deals with the phenomenon of codeswitching between Spanish and English (ES-EN) words and Modern Standard Arabic to Dialect Arabic (MSA-DA). The main aim of this paper is to describe our system submitted to the Second Workshop on Computational Approaches to Code Switching shared task (Molina et al., 2016). We use Conditional Random Fields (CRF) based system to categorize the codeswitched text into eight different categories: *lang1*, *lang2*, *other*, *ne*, *mixed*, *unknown*, *ambiguous*, and *fw*(foreign words). The categories *lang1* and *lang2* refer to the two main language pairs, while *ne* refers to named entities. ES-EN classification uses all these labels whereas MSA-DA classification does not use the labels *unk* and *fw* labels. First we elaborate on the various features we chose to identify and distinguish one category from the other. Then we elaborate on our results and findings. Finally we will analyze the strengths and weaknesses of our system.

2 Related Work

Research into codeswitching and codeswitching detection is not new. There have been many studies on

Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 121–126,
Austin, TX, November 1, 2016. ©2016 Association for Computational Linguistics

codeswitching at socio-cultural level. According to Heredia and Altarriba (2001), the main reason multilingual people codeswitch is the lack of a single language proficiency and frequency/ease of use of word of one language than the other in given context. Eldin (2014) highlight other factors such as clarification, persuasion, effective interaction and current mood and expression of the person for codeswitching.

Recently many researchers have tried various computational methods for language identification in codeswitched text. Yu et al. (2012), Garrette et al. (2015) and Bacatan et al. (2014) use language models and word n-grams with position and frequency data as part of their system to identify codeswitched text. Both achieved an accuracy of over 80%. Maharjan et al. (2015) collected codeswitched tweets for Spanish-English and Nepali-English language pairs. They first figured out some seed users who codeswitched frequently and then followed him/her to collect more codeswitched tweets.They obtained an accuracy of 86% and 87% for Spanish-English and Nepali-English dataset using CRF GE algorithm.

CRF has been used for many different tasks, especially dealing with sequence labeling such as POS tagging (Lafferty et al., 2001a; Silfverberg et al., 2014) and named entity recognition (McCallum and Li, 2003; Settles, 2004). Similar to us, three out of seven participating teams also used CRF for codeswitching detection for the EMNLP 2014 language identification shared task (Solorio et al., 2014). Other participants used Support Vector Machines (SVM),k Nearest Neighbour (KNN), Extended Markov Models (eMM) and spell checker methods with external resources to tackle the problem. In the same year of the shared task, we used dictionaries to store and search prefix ngrams and further used a spell checker to identify wrongly spelled words from social media lingo so they can be classified correctly (Shrestha and Dhulikhel, 2014).

3 Methodology

Our system uses CRF for modeling the sequence of codeswitched tweets. CRF is a very popular sequence classifier for when constituents of a sequence are dependent upon their adjacent context. Rather than modeling for each part separately, it makes more sense to model for the whole sequence. CRF calculates the probability of a whole sequence of labels given a sequence of tokens (words) by using Equation 1. Here $\vec{s}$ represents a sequence of labels and $\vec{x}$ represents a sequence of tokens. $\vec{\phi}$ and $\vec{w}$ are the feature vector and the weights for the feature vector respectively.

$$p(\vec{s}|\vec{x};\vec{w}) = \frac{exp(\vec{w}.\vec{\Phi}(\vec{x},\vec{s}))}{\sum_{\vec{s'}\epsilon S^m} exp(\vec{w}.\vec{\Phi}(\vec{x},\vec{s'}))} \quad (1)$$

Using a CRF with task-specific features, given a context can result in very robust systems for sequence labeling (Lafferty et al., 2001b; Ye et al., 2009). CRF fits the problem of codeswitching detection since frequent switching between languages is not very common. For example, if a token is in Spanish, it is very likely that the token before it and after it are also in Spanish. We describe our features in detail below.

3.1 Features

1. **Token**: The token itself is the biggest indicator of which language it might belong to. Fry and Kress (2012) estimate that the most frequent 1,000 words in the English language accounts for 89% of the text written in English. If the training data is large enough, even a dictionary based approach can provide a very good baseline to build upon. This feature is especially helpful for the tokens that are not shared between the two languages.

2. **Suffix**: Most languages have distinctive suffixes that tokens end in. For example: In English we add *ly* to adjectives to turn them into adverbs similarly in Spanish *-mente* is added to adjectives to turn them into adverbs. So *simply* in English is *simplemente* in Spanish. For the *ing* form of the verb in English there is a close representation in Spanish using the characters *ando,iendo*. English words ending in *-tion* have a Spanish word equivalent in *-cin*. Similarly English *ency, ancy* and Spanish *encia*, English *fic* and Spanish *fico,-fica* etc are some of the many equivalents. As there are many distinct suffixes in both language it is a very helpful feature. We use 1-4 letter suffixes as features.

3. **Prefix**: Although many common prefixes in English have identical prefixes in Spanish *(eg:anti,auto,inter)*, similar to suffixes, these languages also have distinctive prefixes. For example *pseudo* in English is *seudo* in Spanish. So *pseudoscience* in English is spelled as *seudociencia* in Spanish. Other examples include, words beginning in English with *s,ph,poly* generally begin in Spanish with *es,f,poli* respectively. To incorporate these distinct prefixes during classification, we use 1-4 letter prefixes as features.

4. **Prefix and Suffix bigrams**: Rather than just using prefixes and suffixes, we also used bigrams of prefixes and suffixes as features.

5. **Titlecase**: This feature can help catch named entities as the most characteristic attribute of a named entity in a properly written text is that it is in titlecase.

6. **Token is/has Punctuation(s)**: Punctuation tokens and tokens containing many punctuations are more likely to fall in the *other* category.

7. **has/all Unicode**: Spanish tokens contain accented Unicode characters, whereas English ones do not. Also, if the number of Unicode characters in a word is very high, the word is likely to be in the *other* category, for example, in emoticons.

8. **Number**: Tokens containing numbers do not belong to either language.

9. **Uppercase**: Token containing many uppercase characters are usually abbreviations, NE or slangs.

We use the same system with exactly the same features for both ES-EN and MSA-DA language pairs.

4 Implementation Details

We performed a very simple preprocessing step of removing the hashtags since hashtags do not have any significance for this task. The presence of hashtags might confuse the model instead. We also removed non-ASCII quotation marks since these would wrongly be accounted as unicode features. For CRF, we used the sklearn-crfsuite package (Korobov, 2015), which itself is a wrapper over CRFsuite (Okazaki, 2007). In order to obtain the best model, we experimented over all of the available algorithms in CRFsuite namely, *lbfgs*, *l2sgd*, *ap*, *pa*, and *arow*. Using the provided development set, we found the *pa* algorithm to perform the best among them. We also performed hyperparamter optimization by using randomized search over parameters specific to each algorithm. We obtained the highest accuracy on the development set with *pa_type=1* and *c=10.0*. We used these parameters to train a final model by combining both the training and the development datasets. We used this model to obtain our results on the test dataset.

Evaluation Metric	ES-EN	MSA-DA
Monolingual F1	0.90	0.72
Codeswitched F1	0.86	0.34
Weighted F1	0.88	0.66

Table 1: Tweet level results for both Spanish-English and Modern Standard Arabic-Dialect Arabic.

Label	Recall	Precision	F1 score
ambiguous	0.000	0.000	0.000
lang1	0.919	0.919	0.919
lang2	0.981	0.975	0.978
mixed	0.000	0.000	0.000
ne	0.422	0.560	0.481
fw	0.000	0.000	0.000
other	0.993	0.994	0.994
unknown	0.000	0.000	0.000

Table 2: Recall, Precision and F1 score of token level results for Spanish-English.

5 Results

Tweet level results for both ES-EN and MSA-DA are shown in Table 1. For ES-EN, our F1 scores for both monolingual and codeswitched tweets are similar, although the results for monolingual tweets are slightly better. Our weighted F1 score is only 0.03 points below the system with the best weighted F1. The token level results broken down by labels for ES-EN are in Table 2. We could catch most of the *lang1*, *lang2* and *other* tokens. Both our precision and recall are very high for these classes. We also

Label	Recall	Precision	F1 score
ambiguous	0.000	0.000	0.000
lang1	0.959	0.551	0.699
lang2	0.586	0.943	0.722
mixed	0.000	0.000	0.000
ne	0.662	0.851	0.745
other	0.977	0.973	0.975

Table 3: Recall, Precision and F1 score of token level results for Modern Standard Arabic-Dialect Arabic.

catch nearly half of the name entities. Named entities are harder to capture with our system since we do not have a specialized system for named entities, apart from the features that check for titlecase and uppercase. Our system did not catch any of the *ambiguous*, *mixed*, *fw* or *unknown* tokens. The main reason behind this might be how sparse these types of tokens are in training. These comprised of only 0.408% and 0.652% of the total tokens in the training set the test set respectively. Most of the other systems in the shared task also failed to catch these tokens. Despite this, overall, our system was able to obtain good results in the ES-EN language pair codeswitching.

However, we did not fare as well for the MSA-DA language pair. The monolingual F1 is a lot higher at 0.72 than the codeswitched F1 at 0.34 and as such the weighted F1 also suffers. Our system has the propensity of predicting a high number of tokens as lang1, as evidenced by the token level results in Table 3. The recall for *lang1* is very high, while the recall for *lang2* is low. This might owe to the fact that the ratio of tokens in lang1 and *lang2* are reversed in training and test sets. In the training set, 68.74% of the tokens are lang1 and 11.70% of the tokens are *lang2*. Whereas in the test set, only 28.10% of the tokens are lang1 and 46.62% of the tokens are *lang2*. This might have led our model to predict high number of *lang2* tokens as lang1. Similar as in ES-EN, we again perform very well for *other* tokens. A surprising finding with the MSA-DA results is that we have a lot better results for named entities than for ES-EN in both precision and recall, although the system used for both language pairs are exactly the same. One reason might be that the model has a higher number of named entities to learn from in MSA-DA dataset than in the ES-EN dataset. Unfortunately, apart from that, we cannot provide any further intuition into this due to our lack of knowledge of Arabic. Overall, our results for the MSA-DA language pair could definitely use some improvement.

State	Coefficient	Transition	Coefficient
prefix:@,other	2.0138	lang1 → lang1	0.2050
suffix:os,lang2	0.6566	lang2 → lang2	0.2034
is_number,other	0.5938	lang1 → ne	0.1629
has_punct,other	0.5831	unk → unk	0.1491
prefix:’,lang1	0.5704	ne → ne	0.1087
suffix:as,lang2	0.5629	fw → fw	0.1048
all_punct,other	0.5563	ne → lang1	0.0896
suffix:o,lang2	0.4872	lang1 → amb	0.0843
suffix:t,lang1	0.4663	lang2 → mixed	0.0776
all_unicode,other	0.4488	lang2 → amb	0.0712
suffix:a,lang2	0.4436	lang2 → ne	0.0651
suffix:tion,lang1	0.4264	mixed → lang1	0.0455
prefix:th,lang1	0.4230	other → lang1	0.0413
suffix:ed,lang1	0.3837	amb → lang1	0.0412
suffix:ien,lang2	0.3830	unk → other	0.0403
suffix:ing,lang1	0.3726	mixed → lang2	0.0382
prefix:w,lang1	0.3517	unk → lang1	0.0373
suffix:oy,lang2	0.3511	lang1 → unk	0.0371
prefix:Th,lang1	0.3302	lang1 → mixed	0.0314
suffix:ly,lang1	0.3196	other → unk	0.0276

Table 4: Most likely states (state:feature,label) and transitions for the ES-EN model.

6 Analysis

To gain an insight into what our model might be learning, we look at the most likely states and transitions of our CRF model trained on the ES-EN language pair dataset. These are shown in Table 4. Most of the top likely states deal with prefixes and suffixes. The state with the highest coefficient is the one for tokens that have a prefix as @ being *other*. This is very intuitive as most of the tokens that start with @ are Twitter usernames and do indeed have *other* as their label. The suffixes *-os*, *as* and *a* have been associated with Spanish while *t*, *tion*, *ed*, *ing*, and *ly* have been taken as indicators of the token being English by our model. Similarly, the tokens that start with ’ as the prefix having label *other* has high coefficient. This works for tokens such as *’s*, *’ll*, *’d*. Similarly most other top likely states are intuitive and are aligned our knowledge of Spanish and English. The associations of *is_number*, *has_punct* and *all_unicode* with other also have high weight. This reflects the data as numbers do not belong to either language and tokens with all unicode characters are mostly emoticons.

Four out of five most likely transitions are between the same two labels. This shows that language users do not switch between labels too often. The most likely transitions are from English to English and from Spanish to Spanish. The transition from English to Spanish or vice versa did not fall under the top 20 likely transitions. This might be because a person can only use so many *ne* or *other* tokens before switching back to a language but when a person switches from one language to another, there is no such constraint and people are likely to keep on using the other language rather than switching back.

We also looked at some of the instances in the development set where our system predicted the wrong label. We made the most number of mistakes for *ne*. Almost all of these mistakes were for tokens that were not in titlecase. Since this was our only feature specific to named entities, our model could not catch these. There were also some cases where our model did correctly label the tokens as named entities, such as *twitter*, *RT*, *iphones*, etc but the labels given to them in the dataset were different. The most confusion between English and Spanish were for words that are present in both languages such as *yo*, *a*, *no*, *senior*, etc. But this was rare and most of the time our model correctly labeled even these words.

7 Conclusion and Future Works

The task of codeswitching detection is highly suited to be solved using CRF as evidenced by our results. The performance of the system depends on the features chosen and in our case, these features worked for ES-EN and did not work too well for MSA-DA. Our lack of knowledge of Arabic definitely proved to be a hindrance in crafting sensible features for MSA-DA. Nonetheless, we were able to obtain competitive scores for both language pairs. We were also able to see the characteristics of both English and Spanish being captured by our model, especially the most common prefixes and suffixes of both languages. In our further research, we will work towards finding effective features for both ES-EN as well as MSA-DA language pairs. One of the shortcomings of our system is also not being able to capture named entities. In the future, we will also look into named entity recognition systems that work for codeswitched texts.

Acknowledgments

We would like to thank the organizers of the Computational Approaches to Code Switching shared task for providing us with this opportunity.

References

Ana Inés Ansaldo, Karine Marcotte, Lilian Scherer, and Gaelle Raboyeau. 2008. Language therapy and bilingual aphasia: Clinical implications of psycholinguistic and neuroimaging research. *Journal of Neurolinguistics*, 21(6):539–557.

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

Arianna Clarisse R Bacatan, Bryan Loren D Castillo, Marjorie Janelle T Majan, Verlia F Palermo, and Ria A Sagum. 2014. Detection of intra-sentential code-switching points using word bigram and unigram frequency count. *International Journal of Computer and Communication Engineering*, 3(3):184.

Li-Rong Cheng and Katharine Butler. 1989. Code-switching: a natural phenomenon vs language deficiency. *World Englishes*, 8(3):293–309.

Ahmad Abdel Tawwab Sharaf Eldin. 2014. Socio linguistic study of code switching of the arabic language speakers on social networking. *International journal of English linguistics*, 4(6):78.

Edward B Fry and Jacqueline E Kress. 2012. *The reading teacher's book of lists*, volume 55. John Wiley & Sons.

Dan Garrette, Hannah Alpert-Abrams, Taylor Berg-Kirkpatrick, and Dan Klein. 2015. Unsupervised code-switching for multilingual historical document transcription. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1036–1041.

Scott A Hale. 2014. Global connectivity and multilinguals in the twitter network. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 833–842. ACM.

Roberto R Heredia and Jeanette Altarriba. 2001. Bilingual language mixing: Why do bilinguals code-switch? *Current Directions in Psychological Science*, 10(5):164–168.

Mikhail Korobov. 2015. sklearn-crfsuite.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001a. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Proceedings of International Conference on Machine Learning (ICML)*, pages 282–289.

John Lafferty, Andrew McCallum, and Fernando Pereira. 2001b. 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.

Suraj Maharjan, Elizabeth Blair, Steven Bethard, and Thamar Solorio. 2015. Developing language-tagged corpora for code-switching tweets. In *Proc. LAW IX at NAACL*.

Andrew McCallum and Wei Li. 2003. Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons. In *Proceedings of the seventh conference on Natural language learning at HLT-NAACL 2003-Volume 4*, pages 188–191. Association for Computational Linguistics.

Olagunju Robert Modupeola. 2013. Code-switching as a teaching strategy: Implication for english language teaching and learning in a multilingual society. *IOSR Journal of Humanities and Social Science*, 14(3):92–94.

Giovanni Molina, Fahad AlGhamdi, Mahmoud Ghoneim, Abdelati Hawwari, Nicolas Rey-Villamizar, Mona Diab, and Thamar Solorio. 2016. Overview for the second shared task on language identification in code-switched data. In *Proceedings of The Second Workshop on Computational Approaches to Linguistic Code Switching, held in conjunction with EMNLP 2016*.

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

Burr Settles. 2004. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications*, pages 104–107. Association for Computational Linguistics.

Prajwol Shrestha and Nepal Dhulikhel. 2014. Incremental n-gram approach for language identification in code-switched text. *EMNLP 2014*, page 133.

Miikka Silfverberg, Teemu Ruokolainen, Krister Linden, Mikko Kurimo, et al. 2014. Part-of-speech tagging using conditional random fields: Exploiting sub-label dependencies for improved accuracy. In *ACL (2)*, pages 259–264.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steve Bethard, Mona Diab, Mahmoud Gonheim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirshberg, Alison Chang, and Pascale Fung. 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. EMNLP 2014, Conference on Empirical Methods in Natural Language Processing*, Doha, Qatar.

Nan Ye, Wee S Lee, Hai L Chieu, and Dan Wu. 2009. Conditional random fields with high-order features for sequence labeling. In *Advances in Neural Information Processing Systems*, pages 2196–2204.

Liang-Chih Yu, Wei-Cheng He, and Wei-Nan Chien. 2012. A language modeling approach to identifying code-switched sentences and words. *CLP 2012*, page 3.