The Howard University System Submission for the Shared Task in Language Identification in Spanish-English Codeswitching

Mario Piergallini, Rouzbeh Shirvani, Gauri S. Gautam, and **Mohamed Chouikha**
Howard University Department of Electrical Engineering and Computer Science
2366 Sixth St NW, Washington, DC 20059
mario.piergallini@howard.edu,Rouzbeh.asgharishir@bison.howard.edu
gaurishankar.gaut@bison.howard.edu, mchouikha@howard.edu

Abstract

This paper describes the Howard University system for the language identification shared task of the Second Workshop on Computational Approaches to Code Switching. Our system is based on prior work on Swahili-English token-level language identification. Our system primarily uses character *n*-gram, prefix and suffix features, letter case and special character features along with previously existing tools. These are then combined with generated label probabilities of the immediate context of the token for the final system.

1 Introduction & Prior Approaches

The internet and social media have led to the emergence of new registers of written language (Tagliamonte and Denis, 2008). One of the effects of this is the emergence of written codeswitching as a common occurrence (Cárdenas-Claros and Isharyanti, 2009). The First Workshop on Computational Approaches to Codeswitching brought increased attention to this phenomenon. This paper is our submission for the shared task in token-level language identification in codeswitched data for the second such workshop. Our submission is for the Spanish-English language pair.

Our approach was informed particularly by the submissions to the previous shared task in language identification in codeswitched data. Most, if not all, of the previous approaches to word-level language identification utilized character *n*-grams as one of the primary features (Solorio et al., 2014). Nguyen and Doğruöz (2013) and all but one of the systems submitted to the previous shared task used some form of context, several of which used conditional random fields. A number of other types of features have been utilized as well, including capitalization, text encoding, word embedding, dictionaries, named entity gazetteer, among others (Solorio et al., 2014; Volk and Clematide, 2014).

	Train	Dev	Test
# Tweets	11,397	3,011	17,723
# Tokens	139,539	33,276	211,474
Avg. tokens/tweet	12.2	11.1	11.9
% English words	56.5%	50.5%	15.3%
% Spanish words	24.1%	26.0%	58.6%
% Mixed	<0.1%	<0.1%	<0.1%
% Ambiguous	0.2%	0.2%	<0.1%
% Named Entities	2.1%	2.2%	2.1%
% Foreign words	<0.1%	0.0%	<0.1%
% "Other"	16.9%	20.6%	23.9%
% "Unknown"	0.1%	0.4%	<0.1%

Table 1: Data set statistics

2 Data Description

Several thousand tweets were collected from Twitter and labeled by human annotators. Each token was labeled as being English, Spanish, ambiguous (words like *no* which are valid words in both languages and can't be disambiguated by context), mixed (tokens with elements from both languages), foreign (words from other languages), a named entity, "unknown" (tokens like "asdfhg") and "other". The "other" category includes numbers (unless they represent a non-numerical word, like <2> used for "to"), punctuation, Twitter @-mentions, URLs, emojis and emoticons. These tweets were divided

Proceedings of the Second Workshop on Computational Approaches to Code Switching, pages 116–120,
Austin, TX, November 1, 2016.

into train, development and test sets and released[1] to the participants in the shared task. Basic statistics about the train, development and test sets can be seen in Table 2. As can be seen, the proportion of English and Spanish is significantly different for the test set compared to the other two sets.

Systems were evaluated at the tweet level as well. For this purpose, tweets are considered as either monolingual or codeswitched. A codeswitched tweet must have tokens from at least two of the following categories: English, Spanish, mixed and/or foreign. All other tweets are considered monolingual.

3 Methodology

In another paper, also submitted to this conference, we experimented with a number of features for token-level language identification on mixed Swahili-English data (Piergallini et al., 2016). For this shared task, we modified our approach in a few ways due to the parameters of the task and also explored the use of a few new features. These are described below:

1) Word

2) Character *n*-grams (1- to 4-grams)

3) Word prefixes and suffixes (length 1 to 4)

For features 1-3, we filtered out words, *n*-grams, prefixes and suffixes that occurred less than 25 times for training our model. *N*-grams, prefixes and suffixes were also converted to lower-case at the three and four character length to reduce sparsity.

4) English-Spanish dictionary

The dictionary feature checks the token against the English and Spanish dictionaries used in the GNU Aspell package[2] and marked according to whether it was in one or both of the English or Spanish dictionaries, or neither.

5) English POS tag

6) Spanish POS tag

The part-of-speech tags were generated by the Stanford NLTK POS tagger (Toutanova et al., 2003). The Spanish tags were truncated at three characters to reduce sparsity.

7) Named entity tag

Tweets were labeled with the named entity recognition system described in Ritter et al. (2011). This system was developed for use on Twitter data.

8) Brown cluster and cluster prefixes

Brown clustering groups word types into a binary tree structure based on word context (Brown et al., 1992). Clusters tend to correlate with syntactical and semantic categories. They also correlate with language, since words of one language tend to co-occur with other words of the same language. To generate these clusters, we lower-cased all words and replaced all Twitter user names with "@username". We used 400 clusters based on the size of the data and the desire for some distinctions beyond basic word classes. Words that occur infrequently tend to be quite noisy in how they are clustered, so words that occurred less than 10 times were not given a cluster. To take advantage of the binary tree structure, we included features based on prefixes of the cluster. For example, in our clusters, nodes beginning with <0> were mostly Spanish words, while nodes beginning with <11> were mostly English words.

The remaining features are binary flags:

9) Is there a Latin alphabetic character?

10) Is there a Spanish-specific letter?

Spanish-specific characters are limited to accented vowels, <ü> and <ñ>. These are strong indicators of a word being Spanish, but they do not all occur equally frequently, so this feature reduces sparsity. For example, <ó> occurs approximately 40 times more frequently than <ü>. This is the most language-specific feature we use. These characters occur extremely infrequently in English text compared to Spanish text. A language-independent conceptualization of this would be whether the word contains a member of the relative complement of

[1]Data was released by providing tweet ID numbers. Participants scraped the text of the tweets themselves. Since Twitter users may delete or restrict access to their tweets, not all participants may have had the exact same subset of the full data.

[2]Available here: https://github.com/WojciechMula/aspell-python

the set of English letters in the set of Spanish letters. Such a feature would not be useful in the other direction since the 26 letters of the English alphabet are all used in Spanish, particularly in online usage (<w> and especially <k> are not limited to loanwords in internet Spanish writing).

11) Is there a number character?

12) Is the token a numerical expression?

Feature 12 is true for tokens which consist entirely of digits, mathematical symbols, and characters used for expressions of time ("12:00") or currency symbols (<$>), etc.

13) Is there an emoji Unicode character?

Since all tokens composed of emojis are labeled as "other", this feature does not rely on a particular emoji occurring in our training data to accurately classify tokens in the test data.

14) Does the token begin a tweet/sentence?

15) Is the first letter capitalized?

16) Are all of the other letters upper case?

17) Are all of the other letters lower case?

The last four features consider capitalization. These features was added particularly to account for named entities and abbreviations, acronyms, etc. which are typically capitalized or in all upper-case letters. Since words at the beginning of sentences are frequently capitalized, eliminating what is usually a distinction between proper and common nouns, feature 14 should reduce the weight towards labeling a word as a named entity.

Finally, we used logistic regression with L2-regularization to generate label probabilities on tokens using the various combinations of the first 14 features. The label probabilities of the previous and following tokens were then added to the feature vector for each token. Tokens at the beginning or end of a tweet were given all zero probabilities for the absent context. This was found to significantly improve performance in our work on Swahili-English codeswitching (Piergallini et al., 2016) and is simpler than CRF[3]. A second logistic regression model was then trained and applied to the final feature set.

[3]CRF using the same feature sets achieves improvements of only 0.05-0.2% on accuracy but is also much slower.

3.1 Results & Discussion

The results of various feature combinations on the development set are summarized in Table 3. Four of the labels are excluded from the table. None of our models ever predicted a token to be ambiguous, mixed or foreign because these categories were all very rare in the both the training and development data. Conversely, the other category was very easily predicted by even the baseline model and achieved F1 scores of about 99.8% for all configurations.

There is not a high variation in the accuracy based on the features used. What can be seen is that the addition of the label probabilities for the previous and following word consistently adds about 2% to the overall accuracy and improves performance on the English and Spanish categories. It seems that part-of-speech tags and Brown clusters are not especially helpful. It is possible POS tags they could be more useful with a coarser POS tag set, or that the Brown clusters could be more useful with different pre-processing. The use of the named entity recognizer does improve performance on the named entity category significantly, but it did not improve overall accuracy much.

For our predictions on the test data, we used features 2-7 and 9-14 with label probabilities on the word context. Results for our submitted predictions are summarized in Table 3.1. According to the released results, our system never correctly labeled a token as ambiguous or mixed. It also never labeled a token as foreign at all. There are two versions: one with the original test data, and one which excludes tweets which contained URLs. We overlooked URLs in designing our model since they never occurred in the training or development data, although our model likely would've labeled them correctly had they occurred in training. Nevertheless, we achieve an overall accuracy in line with other systems without correcting for this. When tweets containing URLs are excluded, we achieve the highest performance on several measures. Those measures which were highest among submitted systems are noted in bold.

To improve on our model, adding a feature or procedure for properly handling URLs would be the obvious first change to make. However, this does not account for all of the errors in our predictions.

Features Used		Baseline fts 1-3		+Dict +Binary fts 1-4, 9-14		+Brown clusters fts 2-4, 8-14		+POS/NER tags fts 2-7, 9-14		+Brown/NER fts 2-4, 7-14	
Label Prob.		none	w±1	none	w±1	none	w±1	none	w±1	none	w±1
English	P	93.1	95.7	93.4	96.1	93.4	96.1	94.0	96.1	93.5	96.4
	R	97.0	97.9	97.2	98.1	97.3	98.1	96.8	97.9	97.3	98.1
	F1	95.0	96.8	95.3	97.1	95.3	97.1	95.4	97.0	95.4	97.2
Spanish	P	91.6	94.0	92.7	94.4	92.7	94.4	91.7	94.2	92.8	94.5
	R	90.6	96.1	91.0	96.7	91.0	96.7	91.9	96.4	91.0	96.8
	F1	91.1	95.0	91.8	95.5	91.9	95.5	91.8	95.3	91.9	95.6
Named Entity	P	60.4	62.7	61.9	63.1	62.0	63.3	70.3	69.6	68.1	70.5
	R	26.6	29.7	33.5	32.3	33.1	32.6	38.4	39.9	38.7	41.6
	F1	37.0	40.3	43.5	42.7	43.2	43.0	49.7	50.7	49.3	52.3
Unknown	P	0	0	0	25.0	33.3	33.3	33.3	16.7	50.0	0
	R	0	0	0	0.8	0.8	0.8	0.8	0.8	0.8	0
	F1	–	–	–	1.5	1.5	1.5	1.5	1.4	1.5	–
Accuracy		93.8	95.7	94.1	96.0	94.1	96.0	94.3	96.0	94.3	96.2

Table 2: Word-level performance of language identification models on development set (given in percentages)

Token-level		Test	w/o URLs
Overall Accuracy		95.1%	**97.3%**
English	P	90.9%	93.6%
	R	92.9%	**94.1%**
	F1	91.9%	**93.8%**
Spanish	P	97.6%	**98.4%**
	R	**97.8%**	**98.4%**
	F1	**97.7%**	**98.4%**
Named Entity	P	48.9%	60.6%
	R	59.6%	59.9%
	F1	53.7%	60.3%
Other	P	**99.9%**	**99.9%**
	R	92.9%	99.3%
	F1	96.3%	**99.6%**
Unknown	P	1.3%	1.8%
	R	7.0%	8.0%
	F1	2.1%	2.9%

Tweet-level	Test	w/o URLs
Weighted F1	**89.0%**	**91.3%**

Table 3: Performance of the final system on the test data

Notably, our system does poorly with ambiguous, mixed and foreign words. This is largely due to there being very few instances of these categories. We also suspect that dealing with them would require some special approaches to account for their particular features. For example, a mixed language word would be expected to have some *n*-grams found in both English and Spanish, but logistic regression can't easily account for this type of pattern. A feature designed to represent the interaction between the English- and Spanish-like features of a mixed word would be required. It is also possible that some tokens were mislabeled. In our examination, it seemed that the ambiguous and mixed categories were not consistently distinguished.

It is also evident that our system does much worse on named entities than on other large categories. It could be that the tool we used did not have a comprehensive list of named entities (we missed "Orange Is the New Black", for example). It was also only trained on English. Our case features may also be more powerful when combined rather than made into separate binary features. There is an interaction between whether the first letter or all letters are upper or lower case and whether the word is at the beginning of a sentence, and the algorithm we used cannot capture that easily. This could potentially slightly improve performance on named entities. We would also note that English and Spanish do not consider the same types of words to be proper nouns, and this may be the cause for some inconsistencies in the annotations that we noticed.

4 Conclusion

In this paper, we described our system for Spanish-English token-level language identification. We achieved the highest performance on several measures using only the token's immediate context. We also found that POS/NE tagging tools and Brown clusters did not significantly improve overall accuracy over using simpler features, but it is possible refinements could make them more useful.

References

Peter F. Brown, Peter V. deSouza, Robert L. Mercer, Vincent J. Della Pietra, and Jenifer C. Lai. 1992. Class-based n-gram models of natural language. *Computational Linguistics*, 18(4):467–479, December.

Mónica Stella Cárdenas-Claros and Neny Isharyanti. 2009. Code-switching and code-mixing in internet chatting: Between 'yes,' 'ya,'and 'si'-a case study. *The Jalt Call Journal*, 5(3):67–78.

Dong Nguyen and A. Seza Doğruöz. 2013. Word level language identification in online multilingual communication. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 857–862. Association for Computational Linguistics.

Mario Piergallini, Rouzbeh Shirvani, Gauri Gautam, and Mohamed Chouikha. 2016. Word-level language identification and predicting codeswitching points in Swahili-English language data. In *The Second Workshop on Computational Approaches to Code Switching*. Association for Computational Linguistics. To appear.

Alan Ritter, Sam Clark, Mausam, and Oren Etzioni. 2011. Named entity recognition in tweets: An experimental study. In *EMNLP*.

Thamar Solorio, Elizabeth Blair, Suraj Maharjan, Steve Bethard, Mona Diab, Mahmoud Ghoneim, Abdelati Hawwari, Fahad AlGhamdi, Julia Hirshberg, Alison Chang, and Pascale Fung. 2014. Overview for the first shared task on language identification in codeswitched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 62–72. Association for Computational Linguistics.

Sali A. Tagliamonte and Derek Denis. 2008. Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, 83(1).

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - Volume 1*, NAACL '03, pages 173–180, Stroudsburg, PA, USA. Association for Computational Linguistics.

Martin Volk and Simon Clematide. 2014. Detecting code-switching in a multilingual Alpine heritage corpus. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 24–33. Association for Computational Linguistics.