Word-Level Language Identification and Predicting Codeswitching Points in Swahili-English Language Data

Mario Piergallini, Rouzbeh Shirvani 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
mchouikha@howard.edu

Abstract

People in developing countries tend to use a lot of words from English vocabulary. This will make them to switch back and forth between English and their native language. In this work we study English-Swahili codeswitch behavior. We talk about data collection method which consist of interview and online forum data. Then we talk about the methods that we used in order to identify the language of the collected codeswitch data. At the end, some classifier is used to predict the codeswitch behavior of the people.

1 Introduction

Language technology has progressed rapidly in many applications (speech recognition and synthesis, parsing, translation, sentiment analysis, etc.), but efforts have been focused mainly on large, high-resource languages and on monolingual data. Many tools have not been developed for low-resource languages nor can they be applied to mixed-language data containing codeswitching. In many cases, dealing with low-resource languages requires the ability to deal with codeswitching. For example, it is quite common to codeswitch between the lingua franca and English in many former English colonies in Africa, such as Kenya, Zimbabwe and South Africa (Myers-Scotton, 1993b). Thus, expanding the reach of language technologies to users of these languages may require the ability to handle mixed-language data, depending on which domains it is intended for.

Codeswitching produces additional challenges for basic NLP tasks due the simple fact that monolingual tools cannot be applied to other languages, but beyond that, codeswitching also has its own peculiarities and can convey meaning in and of itself. Codeswitching can be used to increase or decrease social distance, indicate something about a speaker's social identity or their stance towards the subject of discussion, or to draw attention to particular phrases (Myers-Scotton, 1993b). Sometimes, of course, it may simply indicate that the speaker does not know the word in the other language, or is not able to recall it quickly in this instance. Computational approaches to discourse analysis will require tools specific to codeswitching in order to be able to make use of these social meanings.

Multiple theories propose grammatical constraints on codeswitching (Myers-Scotton, 1993a), and computational approaches may contribute to providing stronger evidence for or against these theories (Solorio and Liu, 2008). These grammatical constraints also can inform the social interpretation of codeswitching. If a codeswitch occurs in a position that is less expected, it may be more likely to have been used for effect. Similarly, when a codeswitch occurs in a less likely context based on features of the discourse, this also affects the interpretation. The longer a discussion is carried out in a single language, the more likely it would seem that a switch indicates a change in the discourse. For example, Carol Myers-Scotton (1993b) analyzes a conversation where a switch to Swahili and then to English after small talk in the local language adds force to the speaker's rejection of a request. This type of switch could also be precipitated by a change in conversation topic, task (e.g. pre-class small talk transitioning into the beginning of lessons), location,

etc. By contrast, in conversations where participants switch frequently between languages, each individual switch carries less social meaning. In those situations, it is the overall pattern of codeswitching that conveys meaning (Myers-Scotton, 1993b). A model should be able to see this pattern and adjust the likelihood of switches accordingly. Being able to predict how likely a switch is to occur in a particular position may thus provide information to aid in the social analysis of codeswitching behavior.

In this paper, we will be introducing two corpora of Swahili-English data. One is comprised of live interviews from Kenya, while the other was scraped from a large Tanzanian/Kenyan Swahili-language internet community. We will be analyzing codeswitching in both data sets. Human-annotated interviews and a small portion of human-annotated internet data are used to train a language identification model, which is then applied to the larger internet corpus. The interview data and this automatically-labeled data are then used in training a model for predicting codeswitch points.

There are few NLP tools for Swahili and we could find no prior computational work on Swahili that addressed codeswitching. Additionally, existing corpora in Swahili are monolingual, so the creation of two sizable corpora of mixed Swahili-English data will be valuable to research in this area.

2 Prior Research

2.1 Language Identification

Until recent years, most work on automatic language identification focused on identifying the language of documents. Work on language identification of very short documents can be found, for example, in Vatanen et al. (2010). But language identification at the word level in codeswitching data has begun to receive more attention in recent years, particularly with the First Workshop on Computational Approaches to Codeswitching (FWCAC). The workshop had a shared task in language identification, with eight different teams submitting systems on the four language pairs included (Spanish-English, Nepali-English, Mandarin-English and Modern Standard Arabic-Egyptian Arabic) (Solorio et al., 2014). Additionally, prior to this workshop, some work had been done on word-level language identification in Turkish-Dutch data (Nguyen and Doğruöz, 2013) and on language identification on isolated tokens in South African languages (Giwa and Davel, 2013), both with an eye towards analyzing codeswitching.

Most, if not all, of the previous approaches to word-level language identification utilized character n-grams as one of the primary features (Nguyen and Doğruöz, 2013; Giwa and Davel, 2013; Lin et al., 2014; Chittaranjan et al., 2014; Solorio et al., 2014). Those focused on intrasentential codeswitching also utilized varying amounts of context. Nguyen and Doğruöz (2013) and all but one system submitted to the shared task at FWCAC used contextual features. A number of other types of features have been utilized as well, including capitalization, text encoding, word embeddings, dictionaries, named entity gazeteers, among others (Solorio et al., 2014; Volk and Clematide, 2014). Significant variation in the difficulty of the task has been found between language pairs. More closely related languages can be more difficult if they also share similar orthographic conventions, as was found with the MSA-Egyptian Arabic language pair (Solorio et al., 2014). In the FWCAC shared task, notable declines in system performance were found when introduced to out-of-domain data.

2.2 Codeswitch Point Prediction

There has been significantly less work done on the task of predicting codeswitch points. We could only find two articles that deal precisely with this task, Solorio and Liu (2008) and Papalexakis, Nguyen and Doğruöz (2014). The two groups take fairly different approaches to feature design and performance evaluation, while groups use Naïve Bayes classifiers. Solorio and Liu also explore Voting Feature Intervals.

Solorio and Liu look at English-Spanish codeswitching in a relatively small conversational data set created for the study. They use primarily phrase constituent position and part-of-speech tagger outputs as features. The word, its language and its human-annotated POS were also used. These are tested both with and without the features for the previous word. Initial evaluation was done using F1-scores, but as noted in the paper, codeswitching is never a forced choice. As such, the upper-bound

Table 1: Data set Stats

	Interviews	JamiiForums
# Utterances/Posts	10,105	220,434
# Words (tokens)	188,188	16,176,057
Avg. words/item	18.6	73.4
% English words	84.5%	45.8%
% Swahili words	15.4%	54.1%
% Mixed words	<0.1%	<0.1%
% Other words	<0.1%	<0.1%

on this task should be relatively low. To get around this issue, Solorio and Liu came up with a novel approach to test performance by generating artificial codeswitched sentences and had them scored for naturalness by bilingual speakers and achieved scores not far from the natural examples. This approach seems well-justified but requires significant human input.

Papalexakis et al. use simpler features focused on the context of the word. These include the language of the word and the two previous words, whether there was codeswitching previously in the document, the presence of emoticons in the previous two words and the following words, and whether the word is part of a common multi-word expression. These features are applied on a large data set from a Turkish-Dutch internet forum. The language of tokens in this data was labeled automatically using the system in Nguyen and Doğruöz (2013). They find that these features are useful except for perhaps the emoticon-based features.

3 Data Sets

The two data sets we use in this paper come from very different domains. The first is comprised of live interviews, and as such is spoken conversation. The second is from a large internet forum, and so is casual written data with use of emoticons and other behaviors specific to computer-mediated communication. The use of data from two linguistic domains also provides a test of the robustness of our model.

Some descriptive statistics about the two data sets can be seen in Table 1.

3.1 Kenyan Interviews

The interviews in this data set were conducted in Kenya. The participants were students at a Kenyan university and the interviewers were a combination of students and professors at the same university. Most of the participants were interviewed twice, once by a student and once by a professor. This provides two social contexts, one in which the participant and interviewer have the same social status, and one in which the interviewer has a higher social status. In our examination of the data, some differences can be seen in the codeswitching behavior across these two conditions, but this is beyond the scope of this paper.

The interviews were transcribed, translated and annotated for language by native speakers of Swahili who are fluent in English. Words were labeled as either English, Swahili, mixed or other. Some words were labeled as Sheng, which is a mixed language variety in Kenya that is seen as a sort of urban street slang (Mazrui, 1995). Nevertheless, most words labeled as Sheng originate in either Swahili or English and were relabeled accordingly. Words were not labeled as named entities or ambiguous, in contrast to the FWCAC shared task (Solorio et al., 2014). Words that might have been labeled that way were instead labeled according to the context - a proper name was labeled as English if it was surrounded by English, or Swahili if it was surrounded by Swahili. Such words that occur at language boundaries were labeled with the following words or the words within the same sentence (if it occurred at the end of a sentence). There were relatively few instances of such words occurring at language boundaries. This was done partly to simplify the determination of where codeswitch points occur.

While spellings and punctuation are mostly standard, periods were used to mark pauses not merely ends of sentences. False starts and interruptions are also transcribed, so some words end part of the way through.

3.2 JamiiForums Internet Data

The internet data comes from a large Tanzania-based internet forum named JamiiForums[1]. It was scraped by a script and only comprises a small fraction of the entire forum. Due to changes in the forum software and increased security at the board, scraping was interrupted. As we already had a large amount of data, we did not feel it necessary to continue imme-

[1] https://www.JamiiForums.com

diately[2]. Since our interview data came from Kenya, we did, however, scrape the entire Kenyan subforum first.

During scraping, full URLs, embedded images and email addresses were replaced with placeholder terms. Bare hostnames[3] were left alone since they can double as the name of an organization or website. JamiiForums emoticons were replaced with the name of the emoticon as defined by the hover text or image file name. Text within quotation boxes was separated from text in the main body of the post. None of these are used in our analysis, although prior work has explored whether emoticons have any influence on codeswitching behavior (Papalexakis et al., 2014). However, given that users do not always format their posts correctly, some improperly formatted forum code will inevitably have been included in our data.

Language labels for 22,592 tokens of the JamiiForums data were annotated by a native English speaker. These were annotated according to the same rules as the interview data. Annotation was done after applying the initial language identification model to the forum data with only disagreements being labeled by the annotator. This significantly increased the speed that annotation could be done.

4 Language Identification Task

4.1 Methodology

For the language identification task, we experimented with a few different features before settling on the final set. The first feature used were from character *n*-grams (unigrams through trigrams), filtered to exclude *n*-grams that occurred less than 25 times. The symbol # was appended to the beginning and end of the word to enable the *n*-gram features to capture prefixes and suffixes. Additionally, we used a capitalization feature, a dictionary feature and a regular expression feature. The capitalization feature categorized words by whether the first letter only was capitalized and if so, whether it occurred at the beginning of a sentence. Otherwise, words were categorized as either all lower case, all upper case, being comprised of numbers or symbols, and finally words which did not match any of those patterns were labeled as "other". The dictionary-type features were generated using the English and Swahili models using the TreeTagger tool (Schmid, 1994). They were binary features based on whether the word was recognized by the English tagger or the Swahili tagger. The final feature we explored was a regular expression designed to match Swahili phonology. Since Swahili orthography is so regular and native Bantu vocabulary conforms strictly to certain phonological constraints, it was possible to write a regular expression that matched $>95\%$ of Swahili words, with the primary exceptions being words borrowed from Arabic. We found that the Swahili regular expression was redundant with the use of character *n*-grams. Additionally, the English TreeTagger was highly overinclusive, marking many Swahili words as recognized, while the Swahili TreeTagger was underinclusive, making those features relatively weak. So we settled on using only the *n*-gram features along with the capitalization feature.

We then used the LIBLINEAR algorithm (Fan et al., 2008) with L2-regularization to generate context-free predictions over the words from the interview data. This context-free model was then used to expand the feature vector for each word. In addition to the original features, the generated probabilities for each class (English, Swahili, mixed, other) on the previous and following word were added to the feature vector. This achieved a high performance within our training set over a 10-fold cross-validation.

Next, we applied this model to a subset of the JamiiForums data. These labels were used to aid in annotating a portion of the forum data (JF Small). The 6,118 words annotated were then added to the training set and the resulting model was applied to an additional 16,475 words which were then hand-labeled (JF Large). The final model used all of the annotated data and was applied to the full 16+ million word JamiiForums data set.

4.2 Results & Discussion

The results of the various iterations of the model are summarized in Table 2. As you can see, the language probability scores of the word context im-

[2]We are working on getting the complete forum data, see Section whatever

[3]For example, Pets.com

Table 2: Performance of Word-Level Language Identification Models

Train / Test Set Context		Interview 10-fold CV None	Interview 10-fold CV Word±1	Interview JF Small None	Interview JF Small Word±1	Intvw & JF Small JF Large None	Intvw & JF Small JF Large Word±1
English	Precision	94.2%	99.4%	41.6%	87.6%	90.1%	99.2%
	Recall	99.0%	99.7%	95.9%	96.6%	96.5%	98.8%
	F1 Score	96.5%	99.5%	58.0%	91.9%	93.2%	99.0%
Swahili	Precision	92.1%	97.9%	98.1%	99.0%	83.7%	95.3%
	Recall	67.0%	97.2%	62.4%	96.2%	64.1%	97.7%
	F1 Score	77.6%	97.5%	76.3%	97.6%	72.6%	96.5%
Accuracy		94.0%	99.3%	69.7%	96.5%	89.0%	98.4%
Cohen's Kappa		0.74	0.98	0.40	0.92	0.66	0.96

prove performance significantly. Error analysis suggests that it primarily reduces the errors on named entities and numbers. Since we consider named entities and numbers as belonging to the language they're embedded in, it makes sense that these can only be correctly labeled using information about the context. But it also reduces errors on other words. For example, "wake" can be a word in both English and Swahili and context is necessary to disambiguate the language.

Overall, performance within our training set was highly accurate. The greater test was applying it to the out-of-domain forum data. As expected, performance decreased noticeably, with the context-dependent model going from 99.3% to 96.5% accuracy, and 0.98 Cohen's Kappa to 0.92. Nevertheless, this performance compares favorably to the performance of the systems in the FWCAC shared task on the out-of-domain "surprise" data (Solorio et al., 2014). There are several potential explanations for this. One obvious hypothesis is that the Swahili-English language pair is simply easier to distinguish than the language pairs in the shared task. English and Swahili are quite distinct phonologically; for example, Swahili words of Bantu stock universally end in vowels, so a final consonant is a strong indicator that a word is not Swahili. Another potential explanation is that our language label set was different and so the fact that we did not attempt to label named entities or ambiguous words explains the difference in performance. A final hypothesis is that using fewer features made our model more robust across domains. These explanations are difficult to disambiguate without direct comparisons of systems on similar data.

Error analysis on the JF Small set suggested that many of the errors were simply due to out-of-vocabulary *n*-grams. Our interview data included very few numerals and no symbols such as '&', since transcribers were instructed to write only the words as spoken. But these characters are common in written communication. Rather than adjusting our feature set, we decided to add this annotated data to the training set and see how this improved performance. Adding the JF Small set to the interview data and testing on the JF Large set cut the error rate by over half and brought the Cohen's Kappa up to 0.96, almost as high as the performance within the training set. The accuracy of over 98% made us feel confident in applying this model to the full JamiiForums set, which would be used for the codeswitch point prediction task, discussed below.

5 Predicting Codeswitch Points

In studying the codeswitching behavior one of the interesting task as discussed in Section 1 could be predicting whether the person is going to switch or not. Consider example (1) original sentence is the sentence that we are interested to predict the codeswitch on each word. The prediction works this way that we go from the first word up to the last word in the sentence and at each word we predict whether next word is going to be in different language than the current word or not. For example if the current word is "Okay", then we are going to predict if the word after "Okay" is in different language(Swahili) or not. We do the same for all the words inside each post/utterance. If we are at the location of the word "important", we are going to predict whether or not we are going to have switch

Table 3: Example of processing and labeling of a sentence

Raw Text	Manze niko na unenge ile deadly leo tunamanga nini.									
Tokenized	manze	niko	na	unenge	ile	deadly	leo	tunamanga	nini	.
Identify Languages	Swahili	Sw	Sw	Sw	Sw	English	Sw	Sw	Sw	.
Segment Switch Point	manze niko na unenge ile					deadly	leo tunamanga nini			.
	Sw				Sw→En	En→Sw	Sw			.

Table 4: Classification features for codeswitch points

Feature #	Feature Name	Description
1	lang_i	Language of word_i
2	lang_{i-1}	Language of word_{i-1}
3	lang_{i-2}	Language of word_{i-2}
4	match($i-1$)	Are lang_i and lang_{i-1} the same?
5	match($i-2$)	Are lang_i and lang_{i-2} the same?
6	# same lang words	# of words of lang_i in $\text{words}_{[0..i]}$
7	# diff lang words	# of words *not* of lang_i in $\text{words}_{[0..i]}$
8	log # same lang words	log_2(feature 6)
9	log # diff lang words	log_2(feature 7)
10	% same lang words	% of words of lang_i in $\text{words}_{[0..i]}$
11	Previous codeswitch	Did a codeswitch occur before word_i?

for the word after "important" which is "kujua".

(1) Okay, *na unafikiria ni* important *kujua* native language?
(**Translation**: Okay, *and do you think it is* important *to know* native language?)

We look at the prediction task as a classification task. On each word level we would like to be able to say if there will be a switch or not. We can look at "switch" as 1 and "not-switch" as 0. The classification algorithm that is going to be used for this task is naive Bayes.

In order to predict the label for each word we need to have a set of features for each word. The set of features that we define for each word are shown in Table. We also assumed that features of each word are independent from each other, that's one of the reasons that we chose naive Bayes over other classification algorithm. Moreover, in order to support our assumption we tried other classification algorithms and among which naive Bayes did the best job that we are going to see the results. We did not considered punctuation as word and just ignored them.

As it can be seen in Table 4, there are totally eleven features. For the features 6 up to 10 that do not take discrete values we defined some bins and if the feature for each word is in a specific bin, that bin is considered as the discrete value for the features. This will make the classification task easier since we are using naive Bayes algorithm.

Table 5: Data set Stats

	Interviews	JamiiForums
# Switch	8,508	922,547
% switch from English	4,217	463,475
% switch from Swahili	4,232	458,465

The classification is done on two different data sets and each data set is divided into training and testing set based on the ten fold cross validation criteria. Two data set as we described before are interview and JamiiForums data. Data set Statistic is shown in Table 6. After training the naive Bayes classifier on the training data, the performance of the classifier is tested. The result for the two data sets are pretty much similar to eachother that we our analysis is that since both data sets are english-swahili codeswitch data set, it will give us similar results. The result is pretty much similar on two data set and this supports that our lang ID algorithm works well. Comparing to previous works in codeswitch prediction(Solorio and Liu, 2008), codeswitch prediction improved the F1 score by 8 percent. Still, we believe this can be improved by using some other features like part of speech of each word as an additional feature.

Table 6: Prediction Results

	Interviews	JamiiForums
Accuracy	97.6%	96.9%
Accuracy(Random guess)	95.4%	94.2%
Precision	26.6%	27.4%
Recall	56.4%	51.3%
F1 Score	36.1%	35.7%
Cohen's Kappa	31.7%	30.6%

6 Discussion

The reason that we start with codeswitch prediction is that we can make the foundation for our next analysis about social meaning of the text. In order to have a better understanding of the codeswitching behavior of the text we need to know the codeswitching habit of the text that we are dealing with. There are several important factors describing the codeswitching behavior including sociolinguistic and socioprogmatic aspects of codeswitch. Suppose people from different cultures and countries form a group, so codeswitching is probably a habit of these communities. By studying the codeswitching pattern and predicting the codeswitching points we will have a better understanding of the codeswitch phenomena. In other words we can say whether the relationship between both languages is symmetric or asymmetric. (I need to put some references about this statement). As we discussed earlier, we are going to extract social meaning from codeswitching pattern. In other words, we want to achieve a better understanding of codeswitching and the reason that people switch their languages. What are the grammatical influences of the languages on the codeswitch? In order to answer these questions we first need to know what the positions of the codeswitching within a text are. In other words we want to try to predict if the person who is talking is going to have a switch on the next word that is coming out of his mouth.

7 Conclusion

Having larger amount of data could help us to understand the social meaning behavior around the codeswitching points as well as improving the precision for predicting codeswitch point. Thats part of the reason we started to crawl JamiiForums to increase the amount of data that we have. Social meaning is a broad category, certain social behaviors are unlikely to occur in an interview. A wide range of social behavior would require a wide range of types of data. We would like to be able to answer questions such as: how unexpected is a speakers behavior? What are they trying to achieve with their language choice? Are they simply unable to find the right word in the other language, are they increasing or decreasing social distance, expressing identity, or are they responding to topic changes, etc.? Can we determine who is the more powerful or influential speaker in a conversation? Can we determine what group people belong to?

Acknowledgments

Do not number the acknowledgment section.

References

Gokul Chittaranjan, Yogarshi Vyas, Kalika Bali, and 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. Association for Computational Linguistics.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

Oluwapelumi Giwa and Marelie H. Davel. 2013. N-gram based language identification of individual words. In *Proceedings of the Twenty-Fourth Annual Symposium of the Pattern Recognition Association of South Africa*, pages 15–22. Association for Computational Linguistics.

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. Association for Computational Linguistics.

Alamin Mazrui. 1995. Slang and codeswitching: The case of Sheng in Kenya. *Afrikanistische Arbeitspapiere*, 42:168–179.

Carol Myers-Scotton. 1993a. *Duelling Languages: Grammatical Structure in Codeswitching*. Oxford University Press, Oxford, UK.

Carol Myers-Scotton. 1993b. *Social Motivations for Codeswitching: Evidence from Africa*. Oxford University Press, Oxford, UK.

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.

Evangelos E. Papalexakis, Dong Nguyen, and A. Seza Doğruöz. 2014. Predicting code-switching in multilingual communication for immigrant communities. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 42–50. Association for Computational Linguistics.

Helmut Schmid. 1994. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Thamar Solorio and Yang Liu. 2008. Learning to predict code-switching points. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 973–981. Association for Computational Linguistics.

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 codeswitched data. In *Proceedings of The First Workshop on Computational Approaches to Code Switching*, pages 62–72. Association for Computational Linguistics.

Tommi Vatanen, Jaakko J. Väyrynen, and Sami Virpioja. 2010. Language identification of short text segments with n-gram models. In *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC'10)*, pages 3423–3430. European Language Resources Association (ELRA).

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.