Identification and Analysis of Conversational Codeswitching Triggers

August 31, 2018

Final Progress Report

Student Co-PIs:

1. Jocelyn Alvarado
   joalvar1@student.uiwtx.edu
   Department of Mathematics and Statistics
   University of the Incarnate Word

2. Rouzbeh Shirvani
   rouzbeh.asghari@gmail.com
   Department of Electrical Engineering and Computer Science
   Howard University

3. Taylor Williams
   taylorhallwilliams@gmail.com
   Department of Bioengineering
   University of California, San Diego

4. Dr. Chen Chen
   carachenchen@gmail.com
   Department of Earth, Atmospheric, and Planetary Sciences
   Purdue University

5. Dr. Yanina Shkel
   yshkel@princeton.edu
   Department of Electrical Engineering
   Princeton University

PROJECT SUMMARY

In language, codeswitching occurs when a speaker uses two or more languages in the context of one conversation. Speculation on motivation for switching is multifaceted; it is possible that a person may switch languages to hide certain information from listening native speakers, to better express themselves because certain words cease to exist in a given language, or to accommodate the person with whom they are speaking. Very few scholars have investigated prediction methods related to codeswitching. Though it has been proposed that conversation dynamics affect factors of codeswitching, even fewer articles exist on the mathematical quantification of the hypothesized correlations between two individuals' switching frequency.

In this project, the team aimed to better understand the mechanisms of codeswitching by using Natural Language Processing (NLP) techniques. In particular, the team was interested in investigating and developing algorithms to predict when codeswitching will occur based on several conversation features. The team had access to an expert-annotated Swahili-English data set during the Center for Science of Information (CSoI) data science workshop in Purdue, as well as throughout the year. The preliminary analysis was done during the CSoI workshop using R for basic data manipulation. The team carried out a frequency analysis of codeswitching triggers and used data visualization to reveal the dynamics of codeswitching on a conversational level. The preliminary results suggested that certain words have better predictive value for codeswtiching. Moreover, there was a clear empirically observed correlation between the codeswitching behavior of the two speakers in a conversation.

The team built on these initial findings and applied *supervised*, as well as *unsupervised*, learning algorithms to the codeswitching data. First, the team focused on *classification*: a supervised learning task where the goal is to learn to accurately classify a data point's label given its set of features. The first problem that the team encountered is that most classification algorithms use features which are numerical vectors, while many of the important features in the codeswitchign data (e.g. words, parts of speech, language spoken) are categorical. The NLP techniques which are used to tackle this problem are known as *word embedding* and the team investigated a number of word embedding techniques. The team focused on the simplest technique called *one-hot encoding* and implemented it together with a *Naive Bayes* classification algorithm for prediction of codeswitching at an *utterance level*. That is, the predictor used the features associated with an utterance of speaker A to predict if speaker B's response will contain a codeswitch. The Naive Bayes is a good baseline algorithm because it is simple and known to work well on NLP tasks. The obtained results demonstrate that the features associated with an utterance of speaker A do indeed have predictive value for whether speaker B will codeswitch, and the Naive Bayes algorithm has good performance. Secondly, the team looked at unsupervised learning task and applied *topic modeling* techniques to the codeswitching data. The team found that there is significant overlap between most frequent codeswitching word and the topic of the conversation. The team used python and SKLEARN -- a scientific library that has different supervised and unsupervised machine learning algorithms -- for this analysis.

TEAM MEETINGS

In addition to our monthly group skype meetings, we found it effective to meet in smaller groups of 2-3 between meetings to check in and work collaboratively on challenges we met while completing individual milestones. We also met in the San Francisco Bay area on May 18-22 for focused team work. Dates of our monthly skype meetings and members of the team that were present for those meetings are listed in the table below.

Date of Meeting	***Attendees***
10/27/2017	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
11/17/2017	*Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
12/20/2017	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
1/24/2018	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel*
2/27/2018	*Chen Chen, Rouzbeh Shirvani, Taylor Williams*
4/3/2018	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
4/27/2018	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
5/02/2018	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
5/18/2018-5/22/2018 (in person meeting)	*Jocelyn Alvarado, Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*
08/19/2018	*Jocelyn Alvarado, Yanina Shkel, Taylor Williams*
08/26/2018	*Chen Chen, Rouzbeh Shirvani, Yanina Shkel, Taylor Williams*

BUDGET

The team requests that the Center fund one in-person team meeting which included domestic air travel and lodging costs. The team also asked for funding for three members to travel to a conference.
The team was awarded $6000 for these projected expenses.

To date, the team estimates that **$2,199.39 was spent** to attend the in-person meeting in the San Francisco Bay Area on May 18-22, 2018. Individual expenses and respective amounts requested are outlined in Table 1.

	Amount reimbursed (or requested)
Jocelyn Alvarado	$366.60
Rouzbeh Shirvani	$460.33
Taylor Williams	$159.96
Dr. Chen Chen	$0.0
Dr. Yanina Shkel	$1,212.5
Total	$2,199.39

The team is exploring the possibility of attending a conference in 2019, and would like to request that the **remaining funds** of **$3800.61** to be extended until next year.

OUTCOMES

The team's primary goals were to gain experience completing data analysis research in an interdisciplinary, collaborative environment, as well as to develop innovative approaches to NLP. The team accomplished these goals by:

- Meeting to discuss ideas, possible approaches, and technical challenges. The team met on Skype approximately once a month, and in person once throughout the year. In addition, the team found it very effective to have ad hoc interim meeting with 2-3 members at a time to check in and discuss ongoing issues.
- Reading literature on codeswitching as well as learning about broader techniques for Natural Language Processing (NLP). The team learned about basic NLP techniques such as word embedding, and Naive Bayes classification.
- Coding in Python: in particular, the team spent the in-person meeting collaboratively tackling a series of coding challenges, as well as learning to use Machine Learning python library SKLEARN.

In addition, the team has the following outcomes:

- The team put together a 13-page technical report summarizing progress and results to date. The results of python implementation of the Naive Bayes Classification of codeswitching are included in Chapter 4 of the report, and the results of python implementation of Topic Modeling on codeswitching data are included in Chapter 5.
- Topic Modeling on codeswitching data is also Chapter 5 in Rouzbeh Shirvani's thesis
- Jocelyn Alvarado is planning to present the team's work classification with Naive Bayes at an undergraduate conference during Fall 2018. At this time Jocelyn is planning to attend at least one of the following two conferences:
    - TUMC at SFA (November 2-3)
    - the SIAM Louisiana-Texas Section at LSU (October 5-7).
- Dr. Yanina Shkel and Rouzbeh Shirvani presented a talk at the CSoI Virtual Brownbag Seminar on May 17th.

FUTURE DIRECTIONS

For the Naive Bayes classification the team used words contained in an utterance and if the current utterance contains a switch point as features to predict if the next utterance will contain a switch point. The presence/absence of a switch point is a reasonable feature; As observed in the preliminary work, the frequency of switching of one speaker in the conversation is often correlated with the frequency of switching of the other speaker. Moreover, the words contained in the utterance are related to the grammatical structure of the utterance, as well as to the topic of the conversation. Thus, they could reasonably predict codeswitching too. The team is currently investigating other features that could be useful for prediction. For example, using current codeswitch together with the speaker's identity, and parts of speech could be useful features.

In addition, the team is considering using information theoretic model selection techniques for feature selection. The team is also considering repeating this analysis with more data, as well as more sophisticated machine learning techniques. In the later case, the question of word embedding for bilingual data would also need to be addressed in more detail.