Identification and Analysis of Conversational Codeswitching Triggers

March 1, 2018

Mid-year 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

SUMMARY

So far we have been familiarizing ourselves with current codeswitching and data analysis literature, implementing the analyses using Python, working to frame the correct questions, and exploring which machine learning techniques would help us best tackle our problem. We have considered the following three directions:

1) Autoregressive models for prediction
2) Minimum Description Length (MDL) based models and other unsupervised learning approaches
3) Clustering approach using word embedding

TASK PROGRESS

In our proposal, we originally identified the following tasks as crucial steps in our pursuit of bilingual codeswitching analysis:

Task 1: Literature Review
Task 2: Data Acquisition
Task 3: Data Mining
Task 4: Model Selection and Identification
Task 5: Model Validation
Task 6: Algorithm Design

At this point we have made significant progress on Tasks 1, 3, and 4, with notable exploration of Task 2 in terms of the data we already have collected and other options for data sources.

Task 1: Literature Review - From this step, we verified the need for quantification of causal codeswitching. We also concluded that our analysis should be completed using Python, as most language processing scholars utilize Python as a coding platform.

Task 2: Data Acquisition - We have not acquired any new data, but we have considered expanding our database to include Swahili/English conversation data from Twitter or from other instant messaging platforms. In addition, we learned how to work with the data we do have, by extracting key features such as the root, part of speech, and language of the word.

Task 3: Data Mining - In terms of utilizing the data we have, we were able to convert the work we did at the Center's Data Analysis Workshop from R to Python. We also discussed the most valuable aspects of our data and the many options we have regarding analysis approaches.

Task 4: Model Selection and Identification - Given the above three tasks, we have made some progress preliminarily for this task. Moving forward, we plan to focus primarily on this task for the quantitative analysis of our project.

Task 5: Model Validation - We have yet to complete this task, due to the lack of solid model selection and identification that will come as a result of Task 4 completion.

Task 6: Algorithm Design - Though we have not quite reached this task, we have discussed unsupervised learning approaches, autoregressive approaches, and clustering formulation as all potential algorithms that could be of use down the line.

REFINED FUTURE TASKS

Moving forward, we plan to delve into the quantitative aspect of our analysis. Namely, we aim to direct our efforts towards a clustering approach in order to understand the deeper motivation behind codeswitching. In addition, we plan to explore conversation dynamics through a similar approach. Both paths will incorporate word embedding, which is a word representation that effectively maps a word into a vector space for ease of quantitative analysis in the form of clustering or principal component analysis. An ongoing challenge with word embedding is that not all words -- both English and Swahili -- exist in the database, so its application to our project is valuable but rather limited. As a result, we are considering multiple word embedding packages such as Google news and Polyglot as possible word-to-vector-space mapping approach. However, this platform only has English words, which, again, limits the extent to which we can extract useful information from our dataset(/s).

Another direction we are considering involves conversation content; we hope to predict the context of a conversation based solely on the codeswitched words, in congruence with current hypotheses in language processing literature. Development of the clustering algorithm will expedite the implementation of this analysis approach.

MEETING ATTENDANCE

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. 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*