Biweekly Milestone Report: Identifying Context-Based Clusters of Mandarin-English Code-Switching 07-400 Fall 2022

Anna Cai
Mentor: Professor Alan Black, LTI
https://www.andrew.cmu.edu/user/annacai/07400.html

February 18, 2022

1 Major Changes

There have not been any major changes to the goals or implementation of the project.

2 Progress Report

Since the last status update, I have analyzed and compared the tf-idf vectors (unigram and bigram) across provided categories (speaking environment, gender, nationality). I found that tf-idf did not seem to be a particularly useful measure of word importance; although it accounts for words in common across documents, words that are overall extremely common are still biased towards high tf-idf values. Even though I filtered out stopwords and function words, I was not able to completely ignore common words with little information about the content or context of the conversation, and they ended up saturating the ranks of highest tf-idf values. A more extensive list of stopwords might mitigate this, but I believe it could be more useful to move on to different keyword extraction methods.

I have also trained classification models on the sets of categories for two reasons: to see whether the classes can be differentiated at all, and to find what exactly differentiates them. For the first, I fine-tuned multilingual BERT on each of the category sets, with the consideration that if one of the best out-of-the-box models could not achieve significant performance, then it would be likely that the classes did not have any meaningful differences. The BERT models all ended up performing extremely well, indicating that there most likely are significant differences between the classes. However, as BERT is notoriously non-transparent, I also tried to train some simple classification models using the features I had so far too see if they could indicate which features best explained the class differences. Unfortunately, my logistic regression models using POS distributions and count/tf-idf vectors did not perform very well, indicating that I would need better features in order to approach BERT's performance.

3 Milestone Evaluation

I believe that I have been able to meet my milestone goal for this week.

4 Surprises

There have not been any major surprises in the project so far.

5 Future Plans

For the next status update, I aim to continue testing simple classification models and trying to explain their performance through various methods (gradient analysis, feature selection, decision tree branching), which will help indicate the most important and differentiating features. I hope to also explore more keyword or feature extraction methods to help design feature vectors that better represent the data. Additionally, if I have time, I would like to use MUSE's bilingual dictionary to analyze the differences in distributions of words in different languages with similar meanings.

6 Milestone Revisions

I will not make any revisions to my milestone goals this time.

7 Resources

So far, I have all the resources I need. However, it would be nice to have access to bilingual or multilingual word embeddings.