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

March 25, 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 refined the feature importance analysis I did previously and in particular looked at the unique "important features" for each classification problem. For example, the terms "water polo" and "[playing] DOTA" were found to be useful for predicting gender, while "security guard" and "customer service" were found to be specifically indicative of nationality. However, while these observations may be interesting, they do not lend themselves to generalizable conclusions about any category.

My main focus recently has been applying unsupervised topic modeling and general clustering to define broad categories centered around topic that can be analyzed in terms of grammatical characteristics and compared. I tried applying NMF (non-negative matrix factorization), LDA (latent Dirichlet allocation), affinity propagation, and spectral clustering on full conversation (approx 1 hour) document sizes (the same granularity I previously used for classification). However, based on the words selected for each topic, this did not seem to yield meaningful topic clusters. I thought this could be due to the document size being far too large (as topics in conversation tend to change in a matter of minutes), so I tried applying the same algorithms to smaller conversation chunks of 100 lines (approx 5 minutes) and also retested the BERT classification baselines at this granularity as a sanity check. Unfortunately, while the generated clusters appeared to be slightly improved from a qualitative glance, after computing various clustering metrics (silhouette score, Calinski-Harabasz index, Davies-Bouldin index) and a distributional analysis of language use, I concluded that these clusters were not useful as they did not differ much in code-switching usage nor define actual topic areas.

In addition to the above, I also computed multilingual BERT sentence vectors for the entire corpus, which I may try to use to find topic clusters. Also, as I found an alternative to BERT for cross-lingual word embeddings (fastText aligned word vectors), I will likely use them when needed in the future instead as I found BERT difficult to translate to word-level vectors.

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 plan to experiment with using BERT sentence vectors for topic clustering, though I do not anticipate success with this method. I also hope to try methods of clustering given predefined topics and keywords, where I will attempt to come up with representative topics manually. I may also try previously tested methods again with even smaller granularity (a few lines).

Another approach to try is to “cluster” along the axis of language use or code-switching characteristics rather than topic - for example, to stratify the dataset by amount of English use and analyze the content words in each block, which could help to indicate potential topic and language associations.

6 Milestone Revisions

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

Although not explicitly included as part of a milestone, I have previously discussed using multilingual word embeddings to analyze the differences in language distribution of words with similar meaning. However, this goal has been mostly put off until now due to prioritization of other goals (topic clustering/identification), and will likely continue to stay low priority.

7 Resources

So far, I have all the resources I need. If there is time towards the end of the semester, I may search for monolingual conversation datasets in English and/or Chinese with similar conversation topics to analyze and compare to SEAME.