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 4, 2022

1 Major Changes

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

2 Progress Report

Since the first milestone report, I have analyzed and compared the function vs content word distributions of different subsets of the SEAME corpus, including between monolingual vs code-switched utterances (i.e. phrases or sentences), gender, and nationality (Malaysian vs Singaporean). Some observations I have noted are that on average, men use more English than women and Malaysians use less English than Singaporeans. In particular, Malaysian speakers' use of English is more likely to be with content words (as opposed to function).

I also computed the distribution of words per turn for each language (i.e. how many words speakers say in the language before switching again) and verified it against the histograms in the SEAME paper. While turns in Mandarin can be quite long, most English turns are only 1-2 words. This appears to support the matrix language-frame model of code-switching, where in this case, Mandarin is the dominant or matrix language and English is the embedded language.

Additionally, I filtered out words that are neither Mandarin or English and found them to largely consist of Singlish slang derived from Hokkien, with a few Malay words and references to Japanese pop culture. Based on the cultural backgrounds and age of the speakers, this observation is not surprising.

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 complete an analysis of content words in relation to different categories, which should help me design feature vectors to represent the data.

6 Milestone Revisions

I will make some minor revisions to my milestone goals. As the data is not labeled with parts of speech and out-of-the-box POS taggers are not reliable on code-switched data, rather than finding POS sequences, I will just use function vs content information. Also, since SEAME does not provide topic labels and is the only dataset available to me, rather than analyzing the distribution of grammatical patterns across provided topic groups, I will analyze the occurrence of most common/important content words in relation to language and grammar.

7 Resources

So far, I have all the resources I need.