Project Milestone Report:
Identifying Context-Based Clusters of Mandarin-English Code-Switching
07-300 Fall 2021

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

December 10, 2021

1 Progress Report

Since the proposal, I have read several papers (cited below) and begun analyzing the SEAME corpus.

In my analysis of the SEAME corpus so far, I have compared the Mandarin and English unigrams and bigrams across the conversation and interview sets. I found that while over 50% of words (unigrams) in the conversation set are English, less than 40% of words in the interview set are. However, the proportion of Mandarin-to-English and English-to-Mandarin bigrams are approximately the same for both of the sets (each type making up approximately 10% of the total bigrams in either set). I also visually compared the most common unigrams and bigrams across the sets and languages and found them to be similar, likely because they simply included the most common words.

Additionally, I used part-of-speech (POS) taggers to evaluate the distribution of function vs content words. Since state-of-the-art context-based POS taggers do not perform reliably on code-switched data and I only needed a rough estimate, I simply tagged each unigram without context and considered noun, verb, adjective, and adverb types to be content words (and all other types to be function words). For English unigrams, I used NLTK's built-in POS tagger, while for Mandarin unigrams, I used the Stanford Tagger for Chinese. I found the ratio of function to content words (unigrams) to be similar for both Mandarin and English across the conversation and interview sets (about 20-30% function, 70-80% content). The distribution of the function and content pairs for the bigrams was also similar for both languages across both sets, but interestingly, the code-switched bigrams had very different distributions that were reflected in both sets. Mandarin-English bigrams were less likely to be function-function and content-function and more likely to be content-content and function-content than monolingual bigrams; on the other hand, English-Mandarin bigrams were much less likely to be function-content and much more likely to be content-function. Still, the content-content type made up about half or more of the bigrams in each case, reflecting the high counts of content words. As language tends to have around equal amounts of content and function words, this could be due to noise in the dataset or inaccuracy of the taggers, so I intend to investigate this by first cleaning the data further. I also intend to continue the analysis of the distribution of function and content words across monolingual vs code-switched sentences (as labeled in the dataset).

2 Reflection

There have not been any major changes to the goals of my project since my proposal.

I was largely able to meet the first technical milestone described in the original proposal. I ended up jumping straight into the distributional analysis of the corpora in order to get a sense of the data I would be working with, so I did not spend as much time as I would have liked continuing the literature search. To make up for this, I intend to read more papers by the beginning of next semester.

While my mentor and I were originally not optimistic about the amount of code-switching in the CALLHOME dataset, after analyzing its distribution of English words, we concluded that it did not have enough English usage for this project. As CALLFRIEND was developed by the same organization under similar circumstances, we believe that it will also not turn out to be useful for this project. Currently, we are looking into a code-switching dataset developed by HKUST, but have not been able to acquire it yet. It is likely that I will continue the project using just SEAME, though I would have liked to have data in a greater variety of contexts.

As of now, I do not have any revisions to my biweekly 07-400 milestones.

So far, I have all the resources needed for this project, barring the HKUST dataset as described above.

References

[1] Ying Li, Yue Yu, and Pascale Fung. A Mandarin-English code-switching corpus. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2515–2519, Istanbul, Turkey, May 2012. European Language Resources Association (ELRA).

[2] Ruowei Liu. Exploring the prosodic and syntactic aspects of mandarin-english code switching. 2021.

[3] Jung-Ying Lu. Code-switching between mandarin and english. *World Englishes*, 10(2):139–151, 1991.

[4] Dau-Cheng Lyu, Tien-Ping Tan, Eng Chng, and Haizhou Li. Mandarin–english code-switching speech corpus in south-east asia: Seame. volume 49, pages 1986–1989, 01 2010.

[5] Dau-Cheng Lyu, Tien-Ping Tan, Eng-Siong Chng, and Haizhou Li. Mandarin—english code-switching speech corpus in south-east asia: Seame. *Lang. Resour. Eval.*, 49(3):581–600, sep 2015.

[6] Koustav Rudra, Shruti Rijhwani, Rafiya Begum, Kalika Bali, Monojit Choudhury, and Niloy Ganguly. Understanding language preference for expression of opinion and sentiment: What do Hindi-English speakers do on Twitter? In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1131–1141, Austin, Texas, November 2016. Association for Computational Linguistics.

[7] Sunayana Sitaram, Khyathi Raghavi Chandu, Sai Krishna Rallabandi, and Alan W. Black. A survey of code-switched speech and language processing. *CoRR*, abs/1904.00784, 2019.