Identifying Context-Based Clusters of Mandarin-English Code-Switching
07-300 Fall 2021

Anna Cai
https://www.andrew.cmu.edu/user/annacai/07400.html

November 12, 2021

1 Project Description

I will be working with Professor Alan Black in the Language Technologies Institute on identifying clusters in Mandarin-English code-switching based on social context.

Code-switching is a form of language mixing that occurs in communication between multilinguals where elements from two or more languages are combined, or multiple languages are frequently and fluently alternated, typically within sentences. This language mixing does not occur randomly and consistently follows grammatical rules that may not necessarily be obvious from the grammar of either language.

Just like how the structure of monolingual conversation between two speakers can be affected by social context such as subject matter and the relationship between the speakers, there is strong reason to believe that language patterns (e.g. syntax, word choice) in code-switched communication would also be affected by social context. In this case, some potential axes of social context would include subject matter, role of speaker in relationship, fluency of speaker in each language, and dialect. Our aim is to identify the clusters of social context with similar code-switching patterns; specifically, we hope to find clusters such that there are largely exclusive patterns for each cluster. While we will specifically be working with Mandarin-English code-switching since grammatical rules differ greatly between different languages and the Mandarin-English pair is less studied than others (e.g. Hindi-English, Spanish-English), we hope that some of the insights gained about the types of clusters will be generalizable to code-switching in other language pairs.

The identification of context-based clusters of code-switching patterns has many possible applications. As instances of code-switching generally occur in casual conversational settings, there is relatively little collected code-switched data compared to monolingual data, so language models tend to have difficulty understanding code-switched input. Being able to classify code-switched conversations into these clusters can provide models information about the social context and bigger picture without needing to understand every word. In addition, generated code-switched messages can be made to fit a cluster in order to convey the appropriate tone. This can be applied to targeted persuasion (such as public service announcements or advertisements) or dialogue systems to reach people using the most persuasive or natural method of communication.

2 Project Goals

2.1 75% Project Goal

- identify clusters using unsupervised topic clustering techniques on features incorporating syntactic and semantic information
- analyze and compare characteristics of cluster sets generated by different methods, and of clusters in each set

2.2 100% Project Goal

- identify clusters using unsupervised topic clustering techniques on features incorporating syntactic and semantic information
- analyze and compare characteristics of cluster sets generated by different methods, and of clusters in each set
- train and evaluate classifiers on the clusters
- brainstorm and apply additional methods of evaluating cluster quality
- compare clusters to literature on context-based variance in code-switching patterns for both Mandarin-English and other language pairs (particularly Spanish-English)

2.3 125% Project Goal

- identify clusters using unsupervised topic clustering techniques on features incorporating syntactic and semantic information
- analyze and compare characteristics of cluster sets generated by different methods, and of clusters in each set
- train and evaluate classifiers on the clusters
- brainstorm and apply additional methods of evaluating cluster quality
- compare clusters to literature on context-based variance in code-switching patterns for both Mandarin-English and other language pairs (particularly Spanish-English)
- apply style transfer techniques to the generation of code-switched data in each cluster
- collect additional code-switched data from video sharing sites (YouTube, Bilibili) and/or social media

3 Project Milestones

3.1 First Technical Milestone

For the first technical milestone, I will first gain a better understanding of the different contexts in which code-switching occurs (for any language pairs, but in particular Hindi-English and Spanish-English in addition to Mandarin-English) by reading relevant papers. I will also read about the characteristics of Mandarin-English code-switching in general. Additionally, I will search for additional datasets that could potentially be relevant to this project and look through some samples from each of the corpora I plan to use in order to get a sense of the data. If time permits, I will identify some potentially useful patterns and analyze their distributions in the corpora, as well perform a general distributional analysis of the corpora (most common words/phrases, statistics on use of each language, etc).

3.2 First Biweekly Milestone: February 1

By the first biweekly milestone, I hope to have completed a distributional analysis of the corpora and begun creating sets of syntactic patterns.

3.3 Second Biweekly Milestone: February 15

By the second biweekly milestone, I hope to have several sets of potentially useful patterns collected through various automated methods (e.g. most common part-of-speech sequences). I also hope to have investigated the distribution of these patterns in across the corpora and within provided topic and speaker categories.

3.4 Third Biweekly Milestone: March 1

By the third biweekly milestone, I hope to have started trying topic clustering techniques on the data and investigating methods of incorporating syntactic and semantic information into feature vectors.

3.5 Fourth Biweekly Milestone: March 15

By the fourth biweekly milestone, I hope to continue applying unsupervised clustering to different feature combinations and compare the characteristics (e.g. pattern distribution) of the resulting sets of clusters.

3.6 Fifth Biweekly Milestone: March 29

By the fifth biweekly milestone, I hope to have come up with methods of evaluating the quality of clusters and started training classifiers to predict the cluster of given samples.

3.7 Sixth Biweekly Milestone: April 12

By the sixth biweekly milestone, I hope to have compared the methods of clustering by applying the previously created evaluation methods and evaluated and compared the performance of the classifiers.

3.8 Seventh Biweekly Milestone: April 26

By the seventh biweekly milestone, I hope to investigate the generalizability of the clusters to other language pairs by comparing them to examples in literature and potentially applying our techniques to datasets of other language pairs.

4 Literature Search

While there is much research on code-switching in natural language processing, there is comparatively less research for the Mandarin-English language pair as opposed to pairs such as Hindi-English and Spanish-English. Additionally, many linguistics studies analyzing the features of code-switching do not verify their theories across large corpora. Below I have cited some papers for general inspiration [1] [8] [7], Mandarin-English code-switching [3] [2] [4], topic clustering [9], and style transfer [6] [5].

5 Resources Needed

We will be using Python and the deep learning library PyTorch. So far, we plan to include the publicly available SEAME, CALLHOME Mandarin Chinese, CALLFRIEND Mandarin Chinese, and Bangor Miami corpora in our investigation, though we will continue to search for potentially relevant datasets. If necessary, we will utilize the Language Technology Institute's GPU resources for computation-heavy tasks.

References

[1] A. Seza Doğruöz, Sunayana Sitaram, Barbara E. Bullock, and Almeida Jacqueline Toribio. A survey of code-switching: Linguistic and social perspectives for language technologies. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1654–1666, Online, August 2021. Association for Computational Linguistics.

[2] Zhu Hua. Duelling languages, duelling values: Codeswitching in bilingual intergenerational conflict talk in diasporic families. *Journal of Pragmatics*, 40(10):1799–1816, 2008.

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

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

[5] Aman Madaan, Amrith Setlur, Tanmay Parekh, Barnabás Póczos, Graham Neubig, Yiming Yang, Ruslan Salakhutdinov, Alan W. Black, and Shrimai Prabhumoye. Politeness transfer: A tag and generate approach. *CoRR*, abs/2004.14257, 2020.

[6] Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W. Black. Style transfer through back-translation. *CoRR*, abs/1804.09000, 2018.

[7] 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.

[8] 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.

[9] Christian Wartena and Rogier Brussee. Topic detection by clustering keywords. In *2008 19th International Workshop on Database and Expert Systems Applications*, pages 54–58, 2008.