Code-Switch Text Data Augmentation

Liang-Yuan Wu, National Taiwan University

Abstract

Code-switching, a language phenomenon that uses two languages within one sentence, is very common among bilingual speech communities. A major limitation to current work on code-switching can be boiled down to the lack of sufficient data. Therefore, downstream tasks, such as automatic speech recognition, have suffered greatly. To combat this challenge, we propose a novel method of generating effective code-switching data from monolingual sentences.

Keywords: Code-switching, data augmentation, sentence generation

Introduction

To solve the problem of scarcity[1] in code-switching data available to our community, we propose data augmentation methods for code-switching sentences from monolingual data. We hope that by adopting our method, we can provide logical and reasonable data for future works such as improving the accuracy of automatic speech recognition on code-switching utterances. Our main contributions include:

- code-switching data generation approach,
- and an improvement on code-switching language model.

Procedure

Code-switching highly depends on the habit of the speaker and the topic of the content. As a result, we proposed the method illustrated in Figure 1. We first collect the lecture data of the course Machine Learning of Prof. Hung-yi Lee, which mainly includes Mandarin but also some English terms. Then we split the ML data into two parts, the code-switching sentences(ML_{cs}) and the monolingual sentences(ML_{mono}). The ML_{cs} dataset is used for training the transformer model to learn the structure of code-switching sentences, and the ML_{mono} dataset is used for generating code-switching sentences.

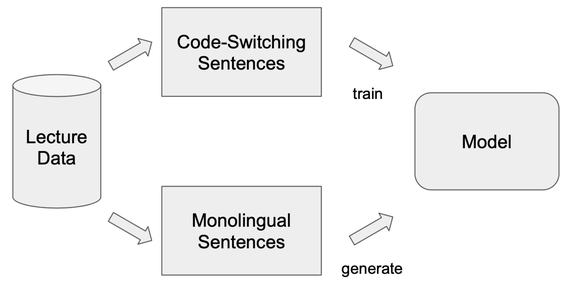

Figure 1: The pipeline of our proposed method

Methodology

The naive method of constructing a code-switching sentence is randomly selecting some words in the ML_{mono} dataset and translating them into English. Based on this idea, we build a baseline method for generating code-switching sentences with MUSE bilingual dictionaries($Data_{MUSE}$).[2]
The transformer models have achieved state-of-the-art results in many NLP tasks. We utilize MBERT and MT5 as the generation model.
MBERT($Data_{BERT}$)[3]

- Multilingual embeddings with EN and ZH
- Predicting masked English words

MT5($Data_{MT5}$)[4]

- Multilingual embeddings with EN and ZH
- Generating the code-switching sentence directly

Results

To evaluate the quality of the generated sentences, we train RNN language models with the code-switching ML_train dataset and the generated code-switching sentences using different generation strategies. We calculate the perplexity of different language models on the same test set of code-switching sentences. Table 1 shows the results of the performance of each language model.

Data	Perplexity
ML_{cs}	26.53
$ML_{cs} + Data_{MUSE}$	25.93
$ML_{cs} + Data_{BERT}$	26.16
$ML_{cs} + Data_{MT5}$	**25.76**

Table 1: LM performance with augmented training data

We can find that adding the augmented data can improve the language model's performance. The transformer-based methods, however, do not necessarily have better performance than the naive translating method.
The best method is using the MT5 model. We first transfer the code-switching sentences to monolingual, and fine-tune the MT5 model to generate the origin code-switching sentences. This method will teach the model not only changing the language of words, but also the proper position to code-switch. The comparison of the generated results are shown in Table 2.

Model	Sentence
Origin	跟一個二次微分比較大的峽谷
Naive	跟一個二次微分比較大的canyon
MBERT	跟一個二次微分比較大的mode
MT5	跟一個二次微分比較大的region

Table 2: Generated code-switching sentence example

Conclusion

- We propose a novel pipeline for code-switching data augmentation, utilizing the monolingual data to generate code-switching data.
- We introduce multilingual transformers into generating code-switching texts, with the help of word masking and conditional generation.
- We propose a new method of using the text-to-text transformer MT5 to generate reasonable and high quality code-switching sentences.
- We conduct evaluations of code-switching language models. Our proposed method reduce the perplexity by 2.8% relative to the baseline.

In the future, we will further survey the improvement on downstream ASR task with the help of code-switching text data augmentation.

Future Research

We will further evaluate the language models' performance on downstream ASR tasks, not only calculating the perplexity but also the word error rate. We will apply our approach to other code-switching datasets such as SEAME to compare the performance in different kinds of code-switching data. Besides, we have not designed a BERT-based model that can have better results in comparison with the naive baseline yet. We will also work on a better approach to generate code-switching sentences using MBERT.

References

[1] Ching-Ting Chang et al. Code-switching sentence generation by generative adversarial networks and its application to data augmentation. *CoRR*, 2018.

[2] Conneau et al. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*, 2017.

[3] Chi-Liang Liu, Tsung-Yuan Hsu, Yung-Sung Chuang, and Hung yi Lee. What makes multilingual bert multilingual?, 2020.

[4] Linting Xue et al. mt5: A massively multilingual pre-trained text-to-text transformer, 2021.