# Mixed Embedding of XLM for Unsupervised Cantonese-Chinese Neural Machine Translation (Student Abstract)

**Ka Ming Wong, Richard Tzong-Han Tsai**

Dept. of Computer Science and Information Engineering, National Central University, Taiwan
{kamingwong, thtsai}@g.ncu.edu.tw

## Abstract

Unsupervised Neural Machines Translation is the most ideal method to apply to Cantonese and Chinese translation because parallel data is scarce in this language pair. In this paper, we proposed a method that combined a modified cross-lingual language model and performed layer to layer attention on unsupervised neural machine translation. In our experiments, we observed that our proposed method does improve the Cantonese to Chinese and Chinese to Cantonese translation by 1.088 and 0.394 BLEU scores. We finally developed a web service based on our ideal approach to provide Cantonese to Chinese Translation and vice versa.

## Introduction

Cantonese (a.k.a. Yue Chinese) is one of the varieties of Chinese. Chinese has the second most number of speakers in the world. In the meanwhile, Cantonese, itself alone, is also ranked in the 19$^{th}$ top languages[1] by population in the world. Although these two languages are very similar and share the same writing system, they are not equivalent. They have slightly different grammar, syntax, pronunciation and use of words. As there is a large population speaking these languages, there is a need to develop a translation task for these languages in order to reduce the cultural gap and facilitate communication. However, there is a major problem that is mostly faced when having a Natural Language Processing (NLP) task corresponds to Cantonese. The resources of this language are very rare, especially corpora. For training a neural machine translation model, corpora, especially parallel corpora, plays a very important role. Both Cantonese monolingual and parallel corpora are scarce. There is nearly no parallel corpus available online. Although the WMT(Workshop on Machine Translation) has similar tasks, the language pairs are not like Cantonese-Chinese, with a shared writing system and highly relevant.

In this work, we focus on improving Cantonese to Chinese(vice versa) machine translation. As there is nearly no parallel corpus for this language pair, we decided to train our model based on UNMT (Lample et al. 2018b). We also adopt the idea from Wan et al. (2019), combining with a modified pre-trained language model and layer to layer attention (He et al. 2018).



[1]https://www.ethnologue.com/guides/ethnologue200

## Related Work

Artetxe et al. (2018) and Lample et al. (2018a) both publish a method of unsupervised neural machine translation (UNMT). Both of them don't need parallel corpus to train a model. Only monolingual data is required which is favourable for our task.Wan et al. (2019) published a method, Dialect Neural Machine Translation, which is using UNMT to train a translation model for Cantonese and standard Chinese. They performed 2 main modifications for training the model. The first one is to concatenate the shared cross-lingual embedding with the monolingual embedding of the source and target language as input. The second one is to apply He et al. (2018)'s layer to layer attention mechanism. Under these modifications, it can provide a better translation and improve the fluency of the output sentence of this language pair.

Pre-trained language models have been proven which can provide a huge impact in many NLP tasks, including UNMT. Lample et al. (2018b) also pointed out that a powerful language model could be able to help to construct better and more fluent sentences in UNMT. One of the most famous pre-training methods, BERT(Devlin et al. 2019), has shown its power and robustness in many tasks. However, BERT is mostly focused on monolingual language. For cross-lingual tasks, Lample and Conneau (2019), have published XLM(Cross-lingual Language Model Pretraining). XLM is a language model that supports cross-lingual training. XLM is very similar to BERT, it contains token embedding, position embedding and language embedding. Then trained together as a masked language model (MLM).

## Modified XLM with UNMT

In this work, We have modified the structure of the token embedding of XLM. Figure 1 shows the structure of our modified XLM. We separate the token embedding into two parts. The first part is the original token embedding which is shared and trained together with both source and target languages. The second part is the private token embeddings. Each of the languages is individually trained with source or target language only. Lastly, we concatenate the private and

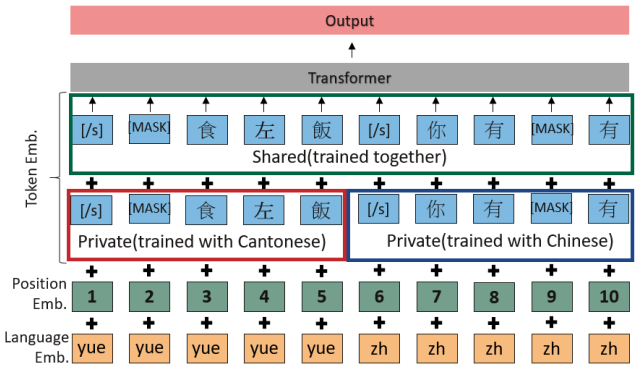


Figure 1: Modified XLM

shared token embeddings according to the words. The whole setup is then trained as an MLM by using Wikipedia in both languages for training data.

After training a language model(modified XLM), we then train a UNMT model by using this pre-trained language model to initialise the encoder of the model. The UNMT model is also modified according to He et al. (2018)'s concept(layer to layer attention). We train our model with the same data set that was used in pre-training XLM. We slightly improve the translation result in comparison with using the original XLM in UNMT. Table 1 shows the results of our experiment. The results are calculated with BLEU score. The testing data set are 2000 human translated parallel sentences. As Cantonese is a variant of Chinese, it is very similar to it. We assume that the worst case of the model is not going to translate anything but just output the original sentence. The non-translated set could be seen as the bottom line of the result. The baseline is using the original XLM with the non-modified UNMT model. The proposed set is the proposed method we introduced. The results show that our proposed method does improve Cantonese to Chinese(vice versa) translation by 1.088 and 0.394 in BLEU score.

## The Online Translator

To more practical use of our work, we have developed a web service to demonstrate our model. We created a web page for the instant online translator. Figure 3 shows the layout of our web page. We have created a translation page for both Cantonese to Chinese and vice versa. We also provide several sample sentences for users to try. Finally, we added a feedback function. We believe that the feedback will be very useful for us to tune the translation model in the future.

| Model | Testing Set | |
|---|---|---|
| | yue-zh | zh-yue |
| Non-Translated | 51.49 | 51.27 |
| Baseline | 68.504 | 70.596 |
| Proposed | **69.592** | **70.99** |

Table 1: Experiment Results

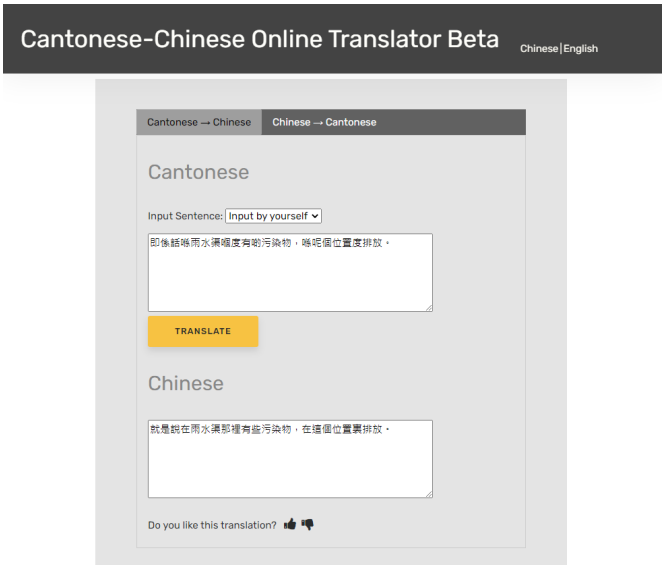


Figure 2: Web Application Layout

## Conclusion

In this paper, we have contributed a method to improve the translation of Cantonese-Chinese. Concatenation of private and shared embedding of XLM with combining layer to layer attention has been proven beneficial to Cantonese-Chinese translation. We also develop a web application of an online translator for these two languages. We hope that the feedback we get could help improve our translation model and also more work related to Cantonese could be done in the future.

## References

Artetxe, M.; Labaka, G.; Agirre, E.; and Cho, K. 2018. Unsupervised Neural Machine Translation. arXiv:1710.11041.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv:1810.04805.

He, T.; Tan, X.; Xia, Y.; He, D.; Qin, T.; Chen, Z.; and Liu, T.-Y. 2018. Layer-Wise Coordination between Encoder and Decoder for Neural Machine Translation. In Bengio, S.; Wallach, H.; Larochelle, H.; Grauman, K.; Cesa-Bianchi, N.; and Garnett, R., eds., *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Lample, G.; and Conneau, A. 2019. Cross-lingual Language Model Pretraining. arXiv:1901.07291.

Lample, G.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018a. Unsupervised Machine Translation Using Monolingual Corpora Only. arXiv:1711.00043.

Lample, G.; Ott, M.; Conneau, A.; Denoyer, L.; and Ranzato, M. 2018b. Phrase-Based & Neural Unsupervised Machine Translation. arXiv:1804.07755.

Wan, Y.; Yang, B.; Wong, D. F.; Chao, L. S.; Du, H.; and Ao, B. C. H. 2019. Unsupervised Neural Dialect Translation with Commonality and Diversity Modeling. arXiv:1912.05134.