# The Ups and Downs of Training RoBERTa-based models on Smaller Datasets for Translation Tasks from Classical Chinese into Mandarin Chinese and Modern English

**Stuart M. McManus** smcmanus@cuhk.edu.hk
**Leo Tam** 1155158173@link.cuhk.edu.hk
**Yuji Li** 1155157174@link.cuhk.edu.hk
**Songyu Liu** 1155191559@link.cuhk.edu.hk
**Shuyang Qiu** 1155157147@link.cuhk.edu.hk
**Daniel Ng** ngcheuknamdaniel@link.cuhk.edu.hk
**Letian Yu** Letian.Yu@link.cuhk.edu.hk

Chinese University of Hong Kong Digital History Lab, Chinese University of Hong Kong, Shatin, Hong Kong, China

**Abstract**

The paper presents an investigation into the effectiveness of pre-trained language models, Siku-RoBERTa and RoBERTa, for Classical Chinese to Mandarin Chinese and Classical Chinese to English translation tasks. The English translation model resulted in unsatisfactory performance due to the small dataset, while the Mandarin Chinese model gave reasonable results.

## 1. Introduction

Classical Chinese was the written lingua franca of East Asia for millenia. As with other learned languages (e.g. Latin, Greek, Arabic, Persian, etc.), texts were frequently translated into and out of Classical Chinese, thereby allowing the spread of ideas across linguistic and cultural borders, both within and beyond East Asia's fluctuating polities (Hung, 2005). For example, the extensive translation in multiple directions between Classical Chinese, Manchu and Mongolian with frequent trilingual documents in Qing represents veritable Rosetta Stones that contain information both about word equivalence and pronunciation, as it was common to transliterate terms between languages in the diverse empire (Chang, 2021). Outside the nation, the translation also came at the cost of much effort, since learning foreign languages in an age with few bilingual dictionaries, and fewer teachers, was a tall order. Nonetheless, scholars have highlighted the successes of Persian translators during the Ming dynasty, who prioritized consistency and adherence to the source texts' structure, often glossing identical words in the same manner across different texts (Green and Nile, 2019) while translators for Persian in Ming



*Proceedings of ALT2023: First Workshop on Ancient Language Translation*, pages 15–22
September 5, 2023, Macau SAR, China.

China were trained from childhood to guarantee high proficiency in the language. Similar cases appeared following the footsteps of Xuanzang (Felbur, 2022;Boucher, 1996) and Jesuit Figuists (Wei and Ling-chia, 2019) Finally, there stands the flood of translations from European languages and Japanese which served as a conduit for technical knowledge (and much else) following the Industrial Revolution.

In our own time, the leaps forward in Artificial Intelligence and Machine Learning development propels the rapid evolution of translation from a purely human activity to a machine-regulated one (Sommerschield et al, 2023). Indeed, Machine Translation may be the perfect solution to the problem that some target languages are less popular or difficult to learn (e.g. Classical Chinese) (Chang, 2021). However, the limited corpus, polysemy idiosyncrasy and complex semantic shifts when compared to Mandarin Chinese (Yang et al., 2020) pose significant hurdles to further developments in this area.

## 2. Related Work

Researchers have been working on Machine Translation at a feverish pace aroused by Neural Machine Translation (NMT) models like the transformer-based BERT in 2018 (Luong, 2016) which outperformed the former Recurrent Neural Network (RNN) and achieved astonishing success in Natural Language Processing (NLP) applications, including text understanding and thus Machine Translation (Rogers et al., 2020). Later, Liu et al (2019) released the more advanced derivative RoBERTa, which exceeded BERT thanks to its larger batch size, longer training process, dynamic masking pattern and so on, boosting its abilities in contextualized word processing and offering the potential of model fine-tuning. While optimizing general application of self-attention mechanism and neural network algorithm in Machine Translation (Qin, 2022), researchers also work to introduce models specialized for particular domain, such as the Siku-RoBERTa used in our study, which is pre-trained on “Siku Quanshu” for Classical Chinese-related translation (Tang, 2022). Other efforts were also paid for the Classical Chinese translation quality. For instance, Zhang et al (2022) developed an unsupervised algorithm to overcome the lack of sentence-aligned corpora, while another study adopted a distant-supervision-based method to solve the people name recognition issue in machine translation and other tasks (Zhang et al, 2021). In this paper, we describe a translation model developed on limited training datasets and pre-trained RoBERTa-based models. We apply this method both to the problem of Classical Chinese to Mandarin Chinese translation and Modern English translation. This allows us a comparative view of the impact of the training dataset size and other variables across languages.

## 3. Model

### 3.1 Key Features of the Model

Machine Translation is a sequence to sequence task which is usually trained by an encoder-decoder model. The input sequence is firstly passed through the encoder, which generates a contextualized representation for each input token. These representations are then passed to the decoder, which generates the output sequence. Siku-RoBERTa has 109M parameters and 50,265 vocabularies (Wang, 2022), while RoBERTa has 355M parameters and 29,791 vocabularies (Liu, 2019). We decided to use RoBERTa to be both encoder and decoder as it has more parameters and vocabularies which can generate a more detailed output. Compared with Siku-RoBERTa as encoder and RoBERTa as decoder, the decoder can understand more when the contextualized representation is generated by the same model. Meanwhile, training

cost can be reduced by using shared encoder-decoder technique, it can reduce the memory usage from 109M+355M=464M parameters to 355M parameters.

### 3.2 Tokenizer

We used the Siku-RoBERTa tokenizer to tokenize input (Classical Chinese) and RoBERTa for output (English). In table 1, we demonstrate that both tokenizers can effectively decrease the number of tokens in their pretrained languages. Theoretically, they have a better interpretation in their own language, which is important when considering the maximum tokens input and output length of a model. When the length is too high, it will increase the computation cost as more tokens are trained and generated in each sentence. Table 1 shows that the average of token length is below 100. Therefore, choosing 192 as length can reduce computation cost without losing too much information from long sentences.

| Data\Tokenizer | Siku_RoBERTa | RoBERTa |
|---|---|---|
| Total Input's tokens (Average) | 160726 (27) | 347764 (59) |
| Max No. of Input's | 254 | 512 |
| Total Output's tokens (Average) | 374169 (64) | 253175 (43) |
| Max No. of Output's | 544 | 370 |

Table1

However, the tokenizers use different tokenization schemes and have different special token_ids. For example, Siku-RoBERTa uses 101 as [CLS] while RoBERTa uses 0 as [CLS]. This discrepancy might slightly affect the outcome of the training, so we decided to change the special tokens of Siku-RoBERTa to RoBERTa's in the input.

## 4. Experiment

### 4.1 English Model

#### 4.1.1 Data

The EvaHan2023 competition provided 5,899 sentences of training data, about 160 thousand Chinese characters and 1 million English characters. Considering the data size is small, we preferred to train the whole dataset rather than splitting them into training and validation data. In spite of the risk of overfitting, a small dataset had a bigger effect on outcome.

#### 4.1.2 Training

We considered two possible approaches:

Approach 1: Splitting data into training data (90%) and validation data (10%).

Approach 2: 100% training data and use 20% of training data as validation data.

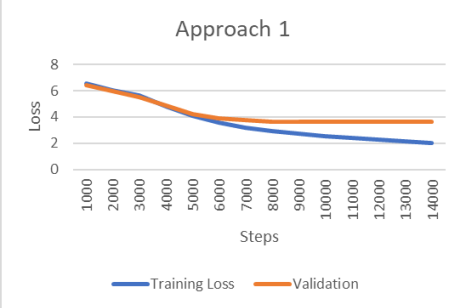


Figure 1

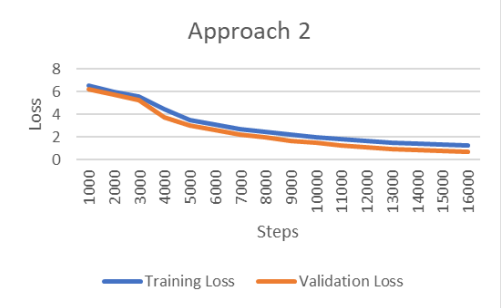


Figure 2

Approach 1 overfitted when the validation loss was around 3.6. Therefore, we used it as reference to predict the time model overfitted in Approach 2. We used a shared encoder-decoder model to lower the complexity of the model. Training Parameters: Learning rate=1.5e-5, Batch_size=8, Tie_encoder_decoder = true, epochs=20

**4.1.3 Comparison**

To investigate the relationship between data size and model performance, we also trained a model for translating from Classical Chinese to Mandarin Chinese. To keep it similar with the English model, we used Siku-RoBERTa as both the encoder and decoder. For the tokenizer, we used Siku-RoBERTa to tokenize input (Classical Chinese) and Chinese-RoBERTa_wwm_ext for output (Mandarin Chinese). The model maximum output length was 192. We used two approaches to process training data for better comparison.

Approach 1 : Using the English model's dataset. The dataset was splitted to 5,309 (90%) for training and 590 (10%) for validation.

Approach 2 : Using the Chinese model's dataset. The dataset was splitted to 305,957 (99.5%) for training and 1,537 (0.5%) for validation. A smaller proportion of validation data was used to shorten the training time.

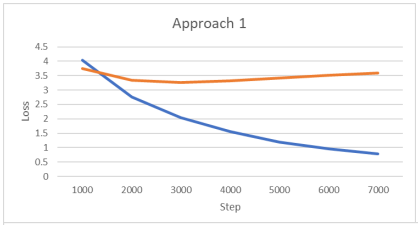


Figure 3


Figure 4

From figure 3, the Chinese model with small data size also resulted in overfit when validation loss is around 3.3. Comparing the score in table 2, we could see the Chinese model worked better than the English model when the data size was small. However, both of them resulted in a much worse performance than the model with larger datasets. It shows the importance of maximizing data size to train a good model.

| Model | EN Approach 1/ EN Approach 2 | M Approach 1 | M Approach 2 |
|---|---|---|---|
| BLEU | 6.1 | 9.0 | 45.4 |
| chrF | 0.005 | 0.20 | 0.64 |

Table 2

### 4.2 Mandarin Chinese Model

#### 4.2.1 Data and Model Architecture

During the experiment, the model was built on the Siku-RoBERTa model. The input was a sequence of Classical Chinese words, encoded using a WordPiece tokenizer. The encoded tokens were then input into the model, which generated a series of hidden states for each token. The final hidden state of the [CLS] token was used for translation. The model contained 9,583,749 characters of Classical Chinese text and 12,763,534 characters in the translation into Mandarin Chinese.

The output sequence was generated using a beam search algorithm, which considered multiple candidate solutions at each step to find the most probable output. The model was trained using the Adam optimizer with a learning rate of 1e-5 for 15 epochs, where batch_size was 16.

#### 4.2.2 Experiments and Results

The experiments conducted on the model aimed to evaluate its performance in the task of translating Classical Chinese to Mandarin Chinese. The dataset was split into training and validation sets, with a 90-10 split ratio.

The first experiment involved training the model on the training set and evaluating its performance on the validation set. The experimental results showed that the Validation Loss of the model could reach as low as 0.0078, and it achieved a BLEU score of 37.8 and a chrF score of 0.47 on the validation set.

The second experiment involved using the trained model to translate a test dataset consisting of Classical Chinese texts. The model used in this experiment was the best-performing training model from the first experiment, and was used to generate Mandarin Chinese translations for the test dataset using a beam search algorithm. We chose the Mandarin Chinese translations provided by Google Translate as the reference translation, achieving a BLEU score of 32.1 and a chrF score of 0.33.

| Experiment No. | 1 | 2 |
|---|---|---|
| BLEU | 37.8 | 32.1 |
| chrF | 0.47 | 0.33 |

Table 3

## 5. Limitation

In the English translation model, it is clear that the dataset is too small, which restricts the model learning ability. Many words in the test data are untrained and data argumentation cannot help solve this problem.

In the experiments on Classical Chinese and Mandarin Chinese translation models, we found that the BLEU score cannot accurately reflect the performance of the translation model because the same content can be expressed using different words, which greatly affects the credibility of the BLEU score as a quality evaluation metric for translation. Future research directions may focus on finding more reliable scoring methods.

## 6. Conclusion

In sum, we used the RoBERTa model to train the Classical Chinese to English translation model, and the Siku-RoBERTa model to train Classical Chinese to Mandarin Chinese model. After comparing the result with the Mandarin Chinese model, we found that the dataset for the English model is too small for obtaining good results. Therefore, we further looked into the Mandarin Chinese model. The Siku-RoBERTa is fine-tuned for the specific task of translation and achieves a reasonable BLEU score on the validation and test datasets. The experiments conducted on the model demonstrate the effectiveness of the Siku-RoBERTa pre-trained model for NLP tasks and highlight the importance of pre-training on large datasets for achieving state-of-the-art performance. The results of the experiments show that the model has the potential to be used for practical translation applications.

## References

Boucher, D. J. (1996). *Buddhist translation procedures in third-century China: a study of Dharmarakṣa and his translation idiom.* University of Pennsylvania.

Chang, K. (Kevin), Grafton, A., & Most, G. W. (2021). Recovering Translation Lost: Symbiosis and Ambilingual Design in Chinese/Manchu Language Reference Manuals of the Qing Dynasty. In *Impagination - Layout and Materiality of Writing and Publication* (pp. 323–350). Walter de Gruyter GmbH. https://doi.org/10.1515/9783110698756-012

Felbur, R., Meelen, M., & Vierthaler, P. (2022). Crosslinguistic Semantic Textual Similarity of Buddhist Chinese and Classical Tibetan. *Journal of Open Humanities Data, 8.*

Green, N. (2019). The Uses of Persian in Imperial China: Translating Practices at the Ming Court. In *The Persianate World* (pp. 113–130). University of California Press. https://doi.org/10.1515/9780520972100-007

Hung, E. (2005). Cultural borderlands in China's translation history. *Translation and cultural change: Studies in history, norms, and image projection, 61,* 43-64.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Luong, M. T. (2016). *Neural machine translation* (Doctoral dissertation, Stanford University).

Qin, Q. (2022). Design and application of Chinese English machine translation model based on improved bidirectional neural network fusion attention mechanism. *Wireless Communications and Mobile Computing, 2022.*

Rogers, A., Kovaleva, O., & Rumshisky, A. (2021). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842-866.

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., ... & de Freitas, N. (2023). Machine Learning for Ancient Languages: A Survey. *Computational Linguistics*, 1-44.

Tang, B., Lin, B., & Li, S. (2022, June). Simple Tagging System with RoBERTa for Ancient Chinese. In *Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages* (pp. 159-163).

王東波, 劉暢, 朱子赫, 劉江峰, 胡昊天, & 沈思等. (2022). Sikubert 與 sikuroberta:面向數字人文的《四庫全書》預訓練模型構建及應用研究(pp.7).圖書館論壇.

Wei, S. L. (2019). *Chinese theology and translation : the Christianity of the Jesuit figurists and their Christianized Yijing.* Routledge.

Yang, K., Liu, D., Qu, Q., Sang, Y., & Lv, J. (2021). An automatic evaluation metric for Ancient-Modern Chinese translation. Neural Computing and Applications, 33, 3855-3867.

Zhang, H., Zhu, H., Ruan, J., & Ding, R. (2021, May). People name recognition from ancient Chinese literature using distant supervision and deep learning. In *2021 2nd International Conference on Artificial Intelligence and Information Systems* (pp. 1-6).

Zhang, Z., Li, W., & Su, Q. (2019). Automatic translating between ancient Chinese and contemporary Chinese with limited aligned corpora. In *Natural Language Processing and Chinese Computing: 8th CCF International Conference, NLPCC 2019, Dunhuang, China, October 9–14, 2019, Proceedings, Part II 8* (pp. 157-167). Springer International Publishing.