Technical Report on Ancient Chinese Machine Translation Based on mRASP Model

Wenjing Liu 18851091773@163.com

Jing Xie* Xie_Hugh@njucm.edu.cn

School of Health Economics and Management, Nanjing University of Chinese Medicine， Nanjing，210046， China

Abstract: Objective This paper aims to improve the performance of machine translation of ancient Chinese classics, which can better promote the development of ancient books research and the spread of Chinese culture. **Methods** Based on the multilingual translation machine pre-training model of mRASP, the model was trained by fine-tuning the specific language pairs, namely a2m, and a2e, according to the two downstream tasks of classical Chinese translation into modern Chinese and classical Chinese translation into English, using the parallel corpus of ancient white and white and ancient English parallel corpus of Pre-Qin+ZiZhiTongJian, and the translation performance of the fine-tuning model was evaluated by BIEU evaluation index. **Results** The BIEU4 results of the three downstream tasks of 24_histories_a2m、Pre-Qin+ZiZhiTongJian_a2m、 Pre-Qin+ZiZhiTongJian_a2e were 17.38, 13.69 and 12.90 respectively.

1 Introduction

Ancient Chinese classics are an important part of Chinese traditional culture. How to mention the automatic translation effect of ancient books is an important topic in the study of ancient Chinese classics. Improving machine translation performance from ancient Chinese to modern Chinese can better promote the development of ancient book research. At the same time, improving the machine translation technology from ancient Chinese to English can also promote the promotion of Chinese traditional culture in the world. EvaHan 2023 is the second international evaluation of ancient Chinese information processing. This evaluation task is the machine translation of ancient Chinese, including two sub-tasks: translating ancient Chinese into modern Chinese; ancient Chinese into English.

*Corresponding author: Jing Xie Ph.D., Associate Professor, School of Health Economics and Management, Nanjing University of Chinese Medicine. Main research directions : natural language processing, intelligence analysis and evaluation based on information technology

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 48–54
September 5, 2023, Macau SAR, China.

This evaluation is divided into two modes: open mode and closed mode. The team chooses the open mode and selects the multi-language translation pre-training model-mRASP released by ByteDance in 2020. Based on this pre-training model, the specific language pairs are fine-tuned, that is, a2c and a2e, to realize the translation of classical Chinese into modern Chinese and English.

2 Multilingual translation model mRASP

2.1 The design motivation of mRASP

mRASP is a recent multilingual translation byte-beating AI Lab at EMNLP2020-Multilingual Random Aligned Substitution Pre-training (mRASP) (Lin et al, 2019). It aims to implement BERT in the field of machine translation and proposes a universal machine translation model. At present, it has become a new successful paradigm of NLP to pre-train the model with a large amount of easily available data and to fine-tune the model with a small amount of labeled data in specific application scenarios to achieve the model available in actual scenarios. For example, after pre-training on large-scale plain text, BERT (Devlin et al, 2018) can achieve good results with a small amount of fine-tuning on multiple natural language processing tasks. However, in multilingual machine translation, the paradigm of pre-training and fine-tuning has not yet achieved universal success. The previous NLP pre-training methods such as BERT and GPT(Radford et al, 2018) have a large gap between the training objectives and the translation focus and are not easy to use directly. MRASP proposes a new idea, which uses a large number of bilingual parallel corpora that have been accumulated in multiple languages to combine and train a unified model, and then fine-tune based on this, so that the pre-training and fine-tuning objectives are as close as possible, to give full play to the role of the pre-training model.

For machine translation, the translation ability is transferred to different languages, so that the information between different languages can be used to each other, thus mentioning the effect of machine translation. Based on this consideration, the design method of mRASP is to design a general pre-training model to learn the commonality of conversion between languages, and then it is easier to migrate to a new translation direction. The design of mRASP follows two basic principles: first, the goal of pre-training is the same as that of machine translation,

and it is necessary to learn the language conversion ability; second, learn the universal representation of language as much as possible. For cross-language sentences or words, if the semantics are close, the representation in the hidden space should also be close.

2.2 mRASP model framework

The mRASP follows a common pre-training-fine-tuning framework. In the pre-training stage, mRASP uses multi-lingual parallel data as the main goal of pre-training. The data sets of 32 open language pairs are put into the same model for joint training and then fine-tuned according to the specific language pairs. The neural network structure uses Transformer, plus a language token to identify the source language and the target language. To ensure that sentences and words of different languages can be embedded into the same space, sentences with the same meaning should be corresponding to the same vector representation in both Chinese and English, and the random substitution alignment technique RAS is introduced to create richer context, It makes the words with similar meanings in different languages closer in the vector space. This method can connect the semantic space between different languages, which greatly improves the final translation effect.

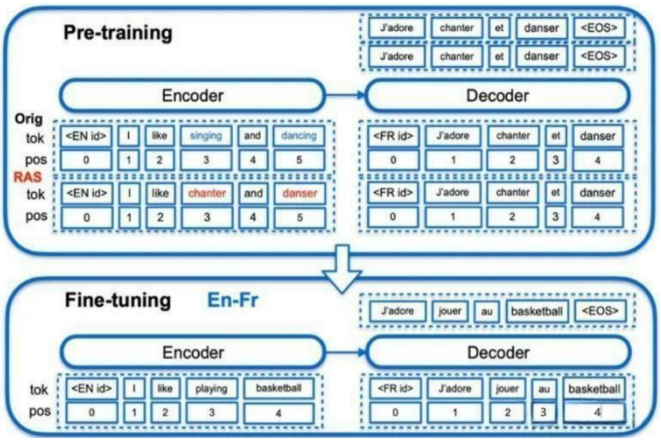

Figure 1 mRASP pre-training model framework

2.3 The role of the mRASP model

The mRASP model confirms for the first time that the multilingual machine translation model can be applied to improve the machine translation model with rich corpus resources. For the first time, the concept of ' zero-shot translation ' in multilingual neural machine translation is extended to 'exotic translation' and divided into four scenarios. In addition, mRASP can even

improve the translation quality of exotic languages which has never appeared in the pre-training corpus. The four scenarios of mRASP extended 'exotic translation ' are as follows:

Exotic Pair: Although both the source language and the target language have been pre-trained, they are separated in the pre-training stage ;

Exotic Source: The source language is not pre-trained, and the target language is pre-trained ;

Exotic Target: The source language is pre-trained, and the target language is not pre-trained ;

Exotic Full: Neither the source language nor the target language was pre-trained.

3 Downstream language pair fine-tuning experiment based on mRASP model

3.1 Source of corpus

This experiment uses the data provided by the second international evaluation of automatic analysis of ancient Chinese EvaHan (Ancient Chinese Machine Translation), including the ancient and white parallel corpus of China's twenty-four histories, the pre-Qin classics, and the ancient and English parallel corpus of 'Zi Zhi Tong Jian'.All the corpora in the experiment are divided into training sets and verification sets according to the ratio of 9:1.The specific expected basic data statistics are shown in Table 1.

Table 1 Basic data statistics of the training corpus

Data set	The number of ancient characters	Number of characters in translation
24_history ancient white parallel corpus	9,583,749words	12,763,534words
Pre-Qin classics and 'Zizhi Tongjian' ancient English parallel corpus	618,083words	838,321words

3.2 Data preprocessing

Data preprocessing for the corpus is an important part of machine translation. The quality of corpus processing determines the effect of the machine translation system training model to a certain extent. According to the data requirements of the mRASP model, the training set and the validation set are preprocessed at the same time. The main preprocessing steps are as follows:1)Data filtering and cleaning;2) The joint BPE sub-vocabulary is used for word

segmentation;3)Binarize the data using the fairseq-preparing command.

3.3 Experimental model parameters and evaluation index

In this paper, Transformer is used as the baseline model, which consists of 6 encoder layers and 6 decoder layers. The hyperparameters of the training model in this experiment are shown in Table 2.

Table 2 The main super parameter settings of the experiment

Super parameter	value
batch_size	512
Learn rating	0.0001
label_smoothing	0.1
dropout	0.3
Update frequency	10
warmup_init_lr	1e-07
fp16	True

In this experiment, BIEU, a commonly used indicator,was used to evaluate the performance of the fine-tuned pre-trained model.BLEU (Bilingual Evaluation Understudy), an automatic evaluation method proposed by IBM researchers in 2002, is currently the most widely used automatic evaluation index (Papineni et al, 2002), by using n-gram matching to evaluate the similarity between the machine translation result and the reference answer, the closer the machine translation is to the reference answer, the higher its quality is determined. The larger the n value is, the larger the matching fragment considered in the evaluation is. The calculation of BLEU first considers the matching rate of the n-gram in the reference answer in the machine translation to be evaluated, which is called n-gram Precision. The calculation method is as follows :

$$P_n = \frac{count_{hit}}{count_{output}}$$

Among them， $count_{hit}$ indicates the number of n-gram hits in the machine translation in the reference answer， $count_{output}$ denotes the total number of n-grams in machine translation. To avoid the same word being repeated calculation, The definition of BLEU is defined by truncation $count_{hit}$ and $count_{output}$.

3.4 Experimental results

Based on the mRASP model, the experiment of translating classical Chinese into modern Chinese and classical Chinese into English is carried out. The research data set is based on the whole sentence. The specific experimental results are shown in Table 3.

Table 3 Experimental evaluation results

Machine Translation Tasks	BIEU-4
24_histories_a2m	17.38
Pre-Qin+ZiZhiTongJian_a2m	13.69
Pre-Qin+ZiZhiTongJian_a2e	12.90

4 Conclusion

This group uses the mRASP multi-language translation machine pre-training model and fine-tunes the specific language pairs of the model according to the two downstream tasks of classical Chinese translation into modern Chinese and classical Chinese translation into English. Firstly, the parallel corpus of specific language pairs is cleaned, segmented, and binarized. Secondly, the generated data is fine-tuned on the mRASP pre-training model, and the translation performance of the fine-tuned model is evaluated by the BIEU evaluation index. Since the source language classical Chinese is not pre-trained in mRASP, while modern and English are trained as target languages, the model training of this language pair belongs to the scenario of 'Exotic Source', and the BIEU value obtained is small. In future work, we can continue to use the mRASP2 (Pan et al, 2021) pre-training model for fine-tuning training. mRASP2 combines 32 language datasets and generates a total of 64 directional translation pairs. On the Transformer model of multilingual translation, a comparative learning task is added at the top of the encoder (Encoder) end to further improve the translation performance.

References:

Zehui Lin, Xiao Pan, Mingxuan Wang, Xipeng Qiu, Jiangtao Feng, Hao Zhou, and Lei Li,J,(2019).Pre-training Multilingual Neural Machine Translation by Leveraging Alignment Information. DOI: https://doi.org/10.48550/arXiv.2010.03142

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, J. (2018). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). Improving language understanding by generative pre-training.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Xiao Pan, Mingxuan Wang, Liwei Wu, and Lei Li, J. (2021). Contrastive Learning for Many-to-many Multilingual Neural Machine Translation.arXiv:2105.09501