Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/37/20230465

Machine translation of classical Chinese based on unigram segmentation transformer framework

Zhuonan Ju[1,†], Yixuan Xin[2,4,†] and Mingda Ye[3,†]

[1]College of Liberal Arts, Nanjing University of Information Science & Technology, Nanjing, 210044, China
[2]School of Materials Science and Engineering, Wuhan University of Technology, Wuhan, 430070, China
[3]SCHOOL OF Sports Engineering, Beijing Sport University, Beijing, 100091, China

[4]xinyx@whut.edu.cn
[†]These authors contributed equally.

Abstract. In the translation work of Chinese ancient books, traditional manual translation is difficult and inefficient. As an important field of natural language processing, machine translation is expected to solve this problem. Due to the rapid development of NLP technology, prior works mainly follow the pipeline of Transformer when dealing with the machine translation task, which can extract the high-quality feature representation with its self-attention mechanism. The great success of Transformer has inspired the direction of our ancient text translation work. In this paper, we screen the Unigram word division by exploring and comparing, and propose a solution for the translation of ancient literary texts. Specifically, we adopt the evaluation of BLEU value and achieve the BLEU values of 43.4 and 40.03 for short and long sentences respectively. When compared with the results of Baidu Translation, our BLEU values increase by 8.12 and 5.18. Additionally, our translation results are more in line with the original text than Baidu Translation, demonstrating the potential and advantage of the model in bridging the ancient and modern Chinese era rift.

Keywords: Machine Translation, Classical Chinese Translation, Transformer, Unigram, BLEU.

1. Introduction

China is one of the four major ancient civilizations in the world. Ancient texts that have gathered thousands of years of wisdom are a unique cultural heritage of the Chinese land. With the continuous development of human civilization, the ancient Chinese language has been influenced by various factors and evolved in the change of time, and it has become more difficult for us to understand the ancient Chinese language nowadays. For traditional manual translation, translation of ancient books is difficult and time-consuming, although it can achieve high-quality and more elegant translations, the cultural level of the translator is very demanding, lots of labor and time will be taken into the cost. As the artificial intelligence technology grows fast, machine translation has been able to efficiently solve most of the routine tasks of translators, reducing the repeated labor of translators and saving a lot of time.

Machine translation is a research branch of natural language processing, which first appeared in the 1940s, and its function is to translate one natural language into another natural language using computers

DOI: 10.54254/2755-2721/37/20230465

[1]. Machine translation can be categorized into following types. (1) Rule-based Machine Translation [2] (RBMT). RBMT methods predefined grammatical and lexical rules to convert source language text into the target language, which usually requires a large number of manual rules and is therefore more difficult for complex language structures and polysemous word processing. (2) Statistical Machine Translation [3] (SMT). SMT methods are used to find the mapping relationship between source and target languages based on a great number of bilingual parallel datasets. Commonly used methods include phrase translation modeling and language modeling, but they may have limitations in dealing with long-distance dependencies and fluency. (3) Neural Machine Translation [4] (NMT). NMT methods use neural networks to learn the mapping relationship between source and target languages. Among them, the Transformer model introduces the self-attention mechanism [5], which has achieved great success in the field of NMT, and is able to better deal with long-distance dependencies and contextual information, and the quality of translation has been improved tremendously, completely replacing SMT as the mainstream machine translation technology.

The field of machine translation currently mainly concentrates on the research between different languages, and has achieved great results in many bilingual translation research, such as Chinese-English, Chinese-Russian, English-German and other different languages, however, the research on the translation between Classical Chinese language and the modern language remains relatively scarce. So far, only a few scholars have had related studies [6][7]. In this paper, according to Transformer has a better translation mechanism than GRU and LSTM [8], based on the Transformer machine translation framework, the optimal Unigram is selected as the Transformer model's segmentation method by exploring different segmentation methods, which greatly improves the model's translation performance.

2. Methods

2.1. Summarisation of proposed model

We constructed the following model under the Transformer basic framework [5][9]. As shown in Figure 1, before feeding the corpus into Inputs, the data is preprocessed, we use Tokenizer to segment each sentence pair, and feed the segmented corpus into the model training, which avoids frequent accessing of hard disk when training our model, and improves the training efficiency. At the same time, the corpus is processed by Tokenizer to build a Vocabulary, and each word is converted into a word vector by Word embedding, which is used to convert the word vectors predicted by the model into different words again in Output Probabilities to get the final results.

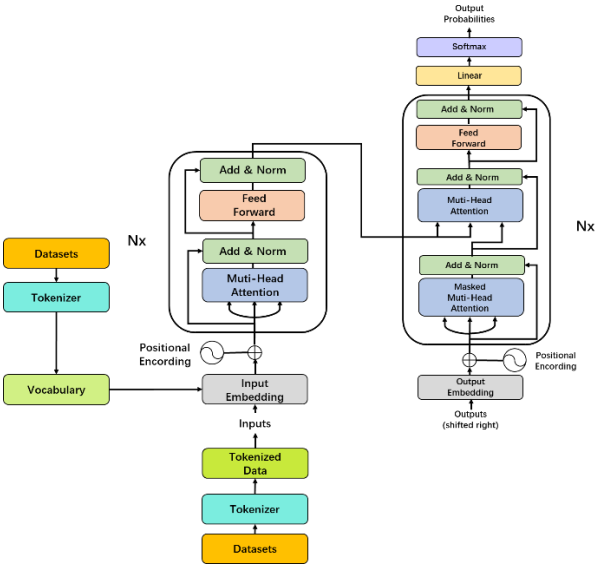


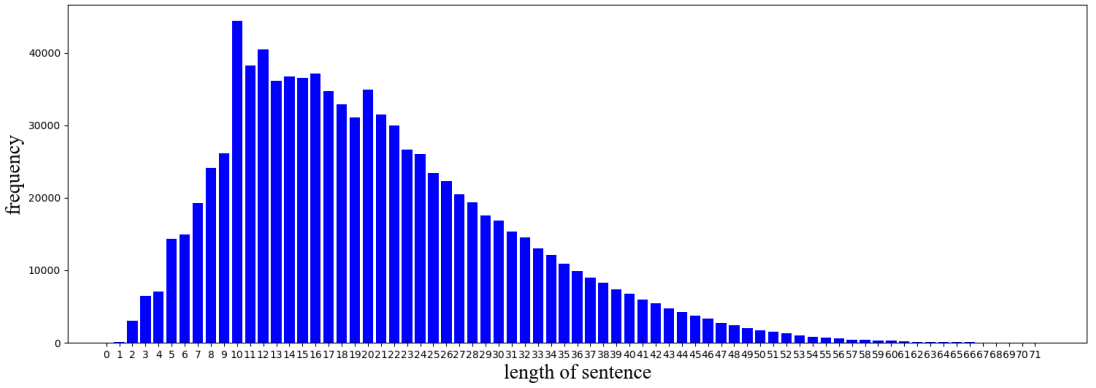
Figure 1. Transformer model with embedded tokenizer.

DOI: 10.54254/2755-2721/37/20230465

2.2. Datasets

Our dataset originates from a joint collection by the Natural Language Processing Laboratory at Northeastern University and the NiuTrans, and the collected corpus data is sourced from the Internet. The raw data crawled is chapter-level aligned bilingual data, which is processed into sentence-level aligned bilingual (parallel) data after a script for clause splitting and alignment, totaling 972,467 sentences. The core alignment idea uses the normalized edit distance algorithm with the length ratio indicator. The bilingual data contains a total of 97 books.

Considering that the max_padding parameter is set to 72 within the program, i.e., sentences exceeding the length of 72 will be automatically truncated, resulting in semantic incompleteness, which will have an impact on the translation accuracy. Therefore, we select the sentences in the dataset that do not exceed 72 characters and exclude the pairs of sentences whose the modern translation's length is less than that of the original ancient text. Finally, 904,419 parallel pairs of sentences were obtained. The average length of sentences in the original ancient text after data cleaning is 20.28 characters, and that in modern translation is 29.87 characters. The distribution of sentence lengths is as follows in Figure 2.



(a) Sentence Length Distribution of the ancient text.

(b) Sentence Length Distribution of the modern translation.

Figure 2. Distrubution of sentence lengths.

2.3. Analysis of different tokenizers

In the Transformer model, the input corpus needs to be constructed into a vocabulary before it is sent to the model for training, so that the corpus can be converted into word encoding. The vocabulary construction process involves the selection of a tokenization method for the input text. In our research, we identified several Chinese tokenization methods through paper review, including N-gram, Jiayan, and Jieba. Given that different tokenization methods can impact both training efficiency and accuracy

DOI: 10.54254/2755-2721/37/20230465

of the model, the following analysis provides a brief overview of the effects of these different tokenization methods.

2.3.1. Introduction to the Three Tokenization Methods. N-gram [10][11] involves grouping N characters together and splitting a sentence into segments of N characters each. It is primarily used for calculating the probability of a sentence.

$$P(w_1 w_2 \cdots w_n) = \prod_{i=1}^{n} P(w_i | w_1 w_2 \cdots w_{i-1}) \quad (1)$$

Jiayan [12] segmentation is an NLP toolkit specialized in processing Classical Chinese. Its segmentation approach employs unsupervised, dictionary-free N-gram grammar and Hidden Markov Models for automatic tokenization of Classical Chinese. The toolkit generates Classical Chinese lexicons through vocabulary construction and uses directed acyclic word graphs, sentence maximum probability paths, and dynamic programming algorithms for segmentation.

Jieba [13], or "结巴" in Chinese, is currently one of the best Python-based Chinese word segmentation libraries. It efficiently utilizes a prefix dictionary to perform a graph scan, generating a Directed Acyclic Graph (DAG) of all possible word compositions within the sentence's Chinese characters. The library employs dynamic programming to identify the maximum probability path based on word frequencies, thus identifying the optimal segmentation combination. For out-of-vocabulary words, Jieba employs an HMM model based on Chinese character composition ability and employs the Viterbi algorithm.

2.3.2. Theoretical analysis. For the translation task of Classical Chinese, there is an ancient Chinese saying that "to chant one word is to twist off several stems of beard", which reflects the concise nature of Classical Chinese writing where each character holds a distinct meaning. Hence, for our N-gram selection, we opted for the Unigram approach.

$$P(w_1 w_2 \cdots w_n) \approx \prod_{i=1}^{n} P(w_i) \quad (2)$$

The Unigram model scatters the probability of the sentence across individual words, aligning closely with the concise nature of Classical Chinese, where each character carries a distinct meaning.

The following example in Table 1, taken from the sentence "其于寿夭何如"(What is the relationship between it and the length of people's life) in the classical text "黄帝内经(五常政大论篇)"(Huang Di Nei Jing: Discussion of the law of Five Elements), illustrates the aforementioned points. In the case of the Classical Chinese sentence "其于寿夭何如", Jiayan separates "其于" (between it) into individual words, while Jieba does not split them. The reason behind this lies in the fact that Jieba's tokenization model treats "其于" as a single phrase during training, which aligns with the characteristics of modern text and is not suitable for Classical Chinese. Similarly, in the translation of the modern sentence "它对于人的寿命长短有什么关系"(What is the relationship between it and the length of people's life), the phrase "关系" (relationship) needs to be treated as a single unit. However, Jiayan's incorrect separation of it stems from Classical Chinese linguistic habits. Based on the analysis above, to accommodate the distinct linguistic characteristics of both Classical Chinese and modern text, the Unigram approach proves to be superior.

Table 1. Comparison of Jiayan and Jieba Tokenization effects.

The text to be tokenized.	Jiayan	Jieba
其于寿夭何如	['其', '于', '寿夭', '何', '如']	['其于', '寿夭', '何如']
它对于人的寿命长短有什么关系	['它', '对', '于', '人', '的', '寿命', '长短', '有', '什么', '关', '系']	['它', '对于', '人', '的', '寿命', '长短', '有', '什么', '关系']

DOI: 10.54254/2755-2721/37/20230465

3. Experiment

3.1. Evaluation

When considering to evaluate the quality of machine translation, we use the consistent BLEU (Bilingual Evaluation Understudy) [14]. The evaluation criterion is to compare with the original corpus translation results. The so-called understudy is to evaluate each output result of machine translation instead of manual labor. Bleu score i.e., given a machine-generated translation, a score is automatically calculated to measure the quality of machine translation. The value ranges from 0 to 100, the translation quality improves as it approaches closer to 100.

$$BLEU = BP \times exp(\frac{1}{n}\sum_{i=1}^{N} P_n) \tag{3}$$

It can also be written as:

$$BLEU = BP \times exp(\sum_{i=1}^{N} w_n log P_n) \tag{4}$$

where BP is the brevity penalty factor, which penalizes a sentence for being too short in length and prevents the training results from leaning towards short sentences.

$$BP = \begin{cases} 1 & if\ c > r \\ exp(1 - \frac{r}{c}) & if\ c \leq r \end{cases} \tag{5}$$

There's also P_n, which is based on the n-gram accuracy,

$$P_n = \frac{\sum_{n-gram \in y} CounterClip(n-gram)}{\sum_{n-gram \in y} Counter(n-gram)} \tag{6}$$

3.2. Design of experiments

Based on the theoretical analysis in Section 2.2, in order to validate the influence of different tokenizers on model training effectiveness, we designed the following four ablation experiments in Tabel 2. These experiments respectively investigate the effects of Unigram tokenization, Jiayan tokenization, Jieba tokenization, and their combined effects.

Table 2. Design details of ablation experiments.

Experiment	Tokenization of Classical Text	Tokenization of Modern Language Translation	batch_size	accum_iter
Experiment 1	Unigram	Unigram	80	10
Experiment 2	Jiayan	Unigram	20	40
Experiment 3	Unigram	Jieba	6	100
Experiment 4	Jiayan	Jieba	5	160

3.3. Data splitting and experimental setting

The dataset we used is the corpus after data cleaning in Section 3.1. Considering time constraints, we utilized approximately half of the corpus for training, totaling 399,770 parallel sentences. Following a ratio of 95% for training and 5% for validation, we conducted model training and validation. The model utilizes CrossEntropyLoss as the loss function and Adam as the optimizer. All the experiments are conducted by a RTX 4090 GPU with 24GB memory. The batch_size and accum_iter vary in different experiments while other relevant settings are depicted as follows.

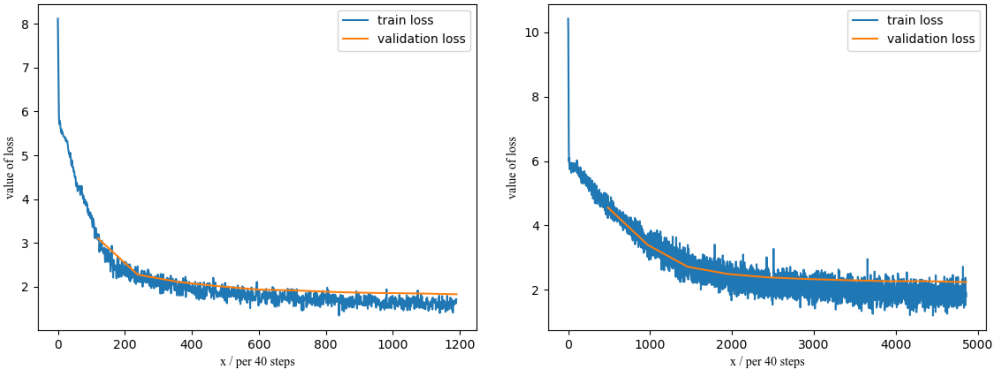
DOI: 10.54254/2755-2721/37/20230465

Table 3. Argument configuration.

Argument	Value	Explanation
num_epochs	10	Number of training rounds
warmup	600	Customize the number of warm-up steps for attenuation
d_model	512	Number of encoder-decoder Hidden nodes
N	6	Encoder Decoder Layers
heads	8	Number of attention heads
dropout	0.1	Discard rate
train_rate	0.95	Corpus for training
valid_rate	0.05	Corpus for validation
d_ff	2048	Number of hidden nodes in the feedforward layer
label_smoothing	0.1	smoothness
max_padding	170	Maximum sentence length

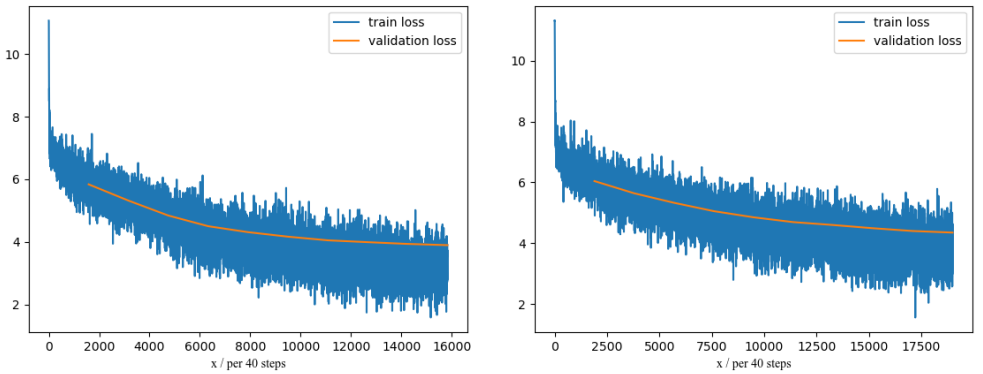
3.4. Result analysis

Figure 3 illustrates the curves depicting the decrease of train loss and validation loss during the experimental process. From the graph, it can be observed that both types of loss for all four cases gradually decrease as the training progresses, although with differing magnitudes and degrees of stability. The Unigram configuration yields the lowest train loss and validation loss, reaching 1.345 and 1.83 respectively in the fewest training epochs, while maintaining the highest stability. Additionally, it outperforms Baidu Translate significantly in terms of BLEU evaluation for translating both short and long sentences. A detailed comparison is provided in Table 4.



(a) Unigram-Unigram (b) Jiayan-Unigram

DOI: 10.54254/2755-2721/37/20230465



(c) Unigram-Jieba (d) Jiayan-Jieba

Figure 3. Loss decline curve.

Table 4. BLEU comparison.

Tokenization method	Types of Classical Text Sentences	BLEU
Jiayan - Jieba	Long(Longer Than 20 Characters)	10.97
	Short(Shorter Than 20 Characters)	20.12
Jiayan – Unigram	Long	20.10
	Short	17.32
Unigram - jieba	Long	11.64
	Short	15.04
Unigram – Unigram (best)	**Long**	**40.03**
	Short	**43.40**
Baidu	Long	34.85
	Short	35.28

Based on the analysis from the experiments, it is evident that the Unigram-Unigram tokenization approach yields the best results. We applied this tokenization approach in subsequent training and expanded the dataset to cover the entire corpus. After training for 99 epochs, the results are as follows. Table 5 presents a comparison between the Transformer machine translation using the best-performing Unigram tokenization from the aforementioned exploratory experiments and Baidu Translate. It's apparent that the machine translation outperforms Baidu Translate in both long and short sentence tasks.

Table 5. Model performance.

Text Types	Content
Classical Chinese	是故百病之始生也,必先于皮毛。
Chinese	因此说,百病的发生,一定是先从皮部开始。
Baidu Translation	所以,各种疾病的产生原因,一定要先在皮毛。
Machine Translation	**因此百病的开始,一定要先从皮毛上发出。**
Classical Chinese	其留于筋骨之间,寒多则筋挛骨痛。
Chinese	若病邪留滞在筋骨之间,寒气盛了,就会筋挛骨痛。

DOI: 10.54254/2755-2721/37/20230465

Table 5. (continued).

Baidu Translation	他留于筋骨之间,寒多则筋挛、骨痛。
Machine Translation	**如果把它留在筋骨之间,寒冷太多就筋骨疼痛。**

4. Conclusion

The automatic translation of ancient Chinese has emerged as a prominent research area in machine translation, while the exploration of translation between Classical Chinese and Modern Chinese remains relatively scarce. In order to alleviate the above issue, we put forward a machine translation model based on Transformer using Unigram as Tokenizer. Specifically, we process the corpus through Tokenizer and build the vocabulary. Then, each word is transformed into a word vector via word embedding and predicts the probability of corresponding translation results. Considering the difference in the training efficiency and accuracy of the model by different word segmentation methods, we further quantitatively made a comparison of the effects of different tokenizers on the translation results. The effectiveness of the proposed model is verified by a large number of experiments, these results demonstrate that our method is able to achieve accurate translation from ancient Chinese to modern Chinese.

References

[1] Li X and Hao X 2021 English Machine Translation Model Based on Artificial Intelligence *Journal of Physics: Conference Series* **1982**

[2] Zhou L 2016 Machine Translation Based on Translation Rules for Processing Natural Language *Proceedings of 2016 6th International Conference on Machinery,Materials,Environment,Biotechnology and Computer(MMEBC 2016)* 488-91

[3] Vogel S, Och F J, Tillmann C, Nießen S, Sawaf H and Ney H 2000 Statistical Methods for Machine Translation *Verbmobil: Foundations of Speech-to-Speech Translation* 377-93

[4] Bengio Y, Ducharme R, Vincent P and Janvin C 2003 A Neural Probabilistic Language Model *J. Mach. Learn. Res.* **3** 1137-55

[5] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L and Polosukhin I 2017 Attention is All you Need *Advances in Neural Information Processing Systems* **30**

[6] Liu Z 2022 Ancient-Modern Chinese Machine Translation Models Based On Transformer *East China Normal University* **11** 103

[7] Zhou C and Liu Z 2022 Ancient Text Machine Translation Method Based on Semantic Information Sharing Transformer *Technology Intelligence Engineering* **8** 114-27

[8] Chung J, Gulcehre C, Cho K H and Bengio Y 2014 Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling *ArXiv https://doi.org/10.48550/arXiv.1412.3555*

[9] Huang A, Subramanian S, Sum J, Almubarak K and Biderman S 2022 The Annotated Transformer *http://nlp.seas.harvard.edu/annotated-transformer/*

[10] Zhou D, He W and Gang C 2011 Research on Tibetan Text Classification Based on N-Gram Model *2011 13th IEEE Joint International Computer Science and Information Technology Conference(JICSIT 2011)* **02**

[11] Kim N S, Baldwin T and Kan M-Y 2010 Evaluating N-gram Based Evaluation Metrics for Automatic Keyphrase Extraction *The 23rd International Conference on Computational Linguistics Proceedings of the Main Conference* **1**

[12] Cui D, Liu X, Chen R, Liu X, Li Z and Qi L 2020 Named Entity Recognition in Field of Ancient Chinese Based on Lattice LSTM *Computer Science* **47** 18-22.

[13] Zeng X 2019 Technology Implementation of Chinese Jieba Segmentation Based on Python [J]. *China Computer & Communication* **31** 38-39+42.

[14] Papineni K, Roukos S, Ward T and Zhu W J 2002 BLEU: A Method for Automatic Evaluation of Machine Translation *Association for Computational Linguistics* 311-8