Proceedings of the 2023 International Conference on Machine Learning and Automation
DOI: 10.54254/2755-2721/34/20230286

Improving machine translation and post-editing for Chinese tourism texts using transformer-based models

Yuan Feng[1,†], Qiwei Hu[2,4,†] and Siran Yang[3,†]

[1]The University of Manchester, Manchester, M139PL, United Kingdom
[2]School of Media Technology, Communication University of China Nanjing, Nanjing, 210013, China
[3]Jiangxi Hanvos High School, Nanchang, 330000, China

[4]1910821111@mail.sit.edu.cn
†All the authors contributed equally.

Abstract. As the digital age and globalization continue to evolve, the demand for accurate machine translation of tourism texts has increased substantially. This paper investigates how to improve the quality of machine translation (MT) and machine translation post-editing (MTPE) of Chinese tourism texts for non-native speakers. A review of the machine translation literature reveals a significant progression in translation methods from rule-based to corpus-based, statistical, and finally to the current neural machine translation (NMT) models. Despite its advanced capabilities, NMT requires large amounts of parallel data for training, which often presents challenges. This study proposes the use of Transformer-based models for MT and MTPE to improve translation quality. A dataset was curated from online sources, mainly Chinese tourism websites. The methodology involved pre-processing the data, performing machine translation using the Transformer model, and post-editing the results. The experiment demonstrated an increase in the BLEU score, suggesting an improvement in translation quality. However, challenges such as the handling of synonyms and geographical nouns were encountered, indicating the need for further research and model optimization.

Keywords: machine translation, neural machine translation, transformer model, tourism texts, machine translation post-editing, Chinese-English translation, BLEU score.

1. Introduction

As globalization intensifies, people increasingly traverse the globe, turning to the Internet for relevant tourism information. Local citizen-authored tourism resources are particularly sought-after for their rich, up-to-date insights. However, language barriers often hinder non-native speakers from accessing useful information from these target countries. In this context, translation plays a crucial role in enabling people to accurately comprehend tourism texts, thereby holding immense practical value.

Neural Machine Translation, an emergent approach to machine translation, has shown remarkable advances. Despite this, it still necessitates vast amounts of data to achieve accurate generalizations that reflect human intelligence.

Machine translation, an early offshoot of Natural Language Processing (NLP), employs computers to translate text from a source language to a target language. The evolution of machine translation has

© 2023 The Authors. This is an open access article distributed under the terms of the Creative Commons Attribution License 4.0 (https://creativecommons.org/licenses/by/4.0/).

DOI: 10.54254/2755-2721/34/20230286

been significant, from the initial rule-based methods to the now-dominant statistical machine translation, traversing through stages of initial recovery and prosperity.

Neural Machine Translation tasks involve transforming a sentence from a source language into a sentence in a target language using a translation model. Although NMT models generally require substantial amounts of parallel data for training, such high-quality data is often scarce. The production of parallel synthetic data is time-consuming and costly, posing a significant problem, especially for low-resource languages or domains.

The increasing demand for practical machine translation systems indicates their significant market potential, particularly in the tourism sector. However, conventional machine translation systems often fall short in effectively translating tourism texts. Therefore, the study of the specific characteristics of tourism texts to enhance the quality of machine translation in the tourism sector is of vital importance.

2. Literature review

Over the past half century, machine translation research has garnered recognition and expanded its applications despite occasional challenges [1]. Encouragingly, numerous commercialized systems have entered the market, including PC translation products both domestically and internationally. Concurrently, the prevalence of PCs and the demand for internet browsing has initiated a trend of machine translation products making their way into millions of homes.

During a memorable encounter with Mr. Nagao Zhen at an academic conference in Singapore in the summer of 1996, he expressed that machine translation was beginning to turn a profit, an announcement that visibly brought him relief [2]. This sentiment resonates deeply with those who spend a lifetime conducting research that initially seems to yield no return, continuously convincing others of the value in such an investment. The pressure faced is immense, understood fully only by those who experience it firsthand.

Japan's situation might be a unique case, with reported investments in machine translation research from 1978 to 1993 totaling $200 million. Currently, the annual sales of machine translation software in Japan reach about 500,000 sets, most priced between $100 to $1000 per set. Therefore, pessimistic views about the research and development of machine translation are seldom found today [3]. However, it's essential to acknowledge the pervasive dissatisfaction and disappointment users express regarding the translation quality of machine translation systems, sometimes quite vehemently [4].

In terms of theoretical or technical breakthroughs, some might point to “rule-based” or “linguistic methods”, “corpus-based” or “corpus and statistical methods”, or “empiricism” or “rationalism”. Such concepts indeed sparked debates in the early 1990s in machine translation and other natural language processing domains [5]. However, it was quickly realized that integrating linguistic methods with corpus and statistical methods yielded better results than opposition. The breakthroughs discussed here refer to other pivotal issues that might instigate technological change [6].

To date, most practical machine translation systems use a sentence as their processing unit, implying their analysis and production are confined to an isolated sentence range. The referred context pertains to this isolated sentence, not a paragraph or several successive sentences [7]. This narrow context proves difficult to analyze, even in syntactic analysis, as it fails to provide sufficient information to guarantee analysis accuracy. This limitation is primarily responsible for the poor translation quality of machine translation systems, which is predominantly due to analysis failures or errors in ambiguity discrimination.

3. Methodology

The evolution of machine translation has seen significant advancements since its inception in the 1950s. It progressed from rule-based machine translation to corpus-based machine translation, inclusive of statistical machine translation, further advancing to neural machine translation. Prominent structures in machine translation applications encompass the likes of Transformer, BERT, bi-LSTM, etc.

As machine translation evolved, the quest for enhanced machine translation quality led to the emergence of machine translation post-editing. Defined as fine-tuning machine translation results with minimal human intervention, MTPE seeks to rectify typical machine translation errors such as word

DOI: 10.54254/2755-2721/34/20230286

substitutions, word form changes, word order changes, additions, and deletions. This approach has evolved into rapid and complete MTPE to minimize the corrective effort [8].

Presently, automatic MTPE tools exist in medical and business domains. Concurrently, the growth of machine translation has amplified the demand for accurate machine translation of Chinese historical sites, poetry, and other cultural terminologies [9]. The focus of this essay is primarily on improving the coherence of sentences in Chinese-to-English translation. Given the increasing interest in Chinese tourist sites among foreign visitors and the growing demand for precise guide scripts to better present Chinese culture and history, this text highlights the MTPE procedure for improving guide script translations.

The core structure of the selected language model is the transformer with attention mechanism, featuring classical neural network components such as encoders, decoders, and self-attention layers. The basic transformer model serves as the training model in the machine translation stage, exemplifying raw machine translation. The Transformer's principle revolves around the attention mechanism, allowing efficient processing of data sequences without traditional recurrent or convolutional layers. The self-attention mechanism is at its heart, enabling the model to weigh the importance of various words in a sequence while processing each word. This feature enables greater focus on relevant words and their relationships, effectively capturing long-range dependencies [10].

The Transformer comprises an encoder and a decoder stack, each with multiple layers. The encoder processes input sequences, while the decoder generates output sequences during tasks like machine translation. The self-attention mechanism is deployed multiple times in parallel, referred to as multi-head attention, facilitating the model to capture different types of information and learn diverse representations. As the Transformer doesn't deploy sequential layers like RNNs, positional encoding is added to input embeddings to provide positional information. Each layer within the Transformer contains feed-forward neural networks, introducing non-linear transformations to the attention output. For stable learning of deep architectures, residual connections are used with layer normalization. In the decoder stack, masking is applied to prevent future positions from being attended to during the self-attention step, ensuring the model attends only to past positions, a crucial aspect for autoregressive generation tasks.

4. Experiments

4.1. Data preparation

The experiments were built up based on online sources from the Internet, mainly the scripts of the main tourism sites in China from their websites. The input is the guide scripts in Chinese language. For the machine translation part, the output should be the results of raw machine translation in comparison with the official translation in English language. For the MTPE procedure, the input should be the results of raw machine translation in English language while the output should be the result of MTPE in English in comparison with the official translation in English language.

4.2. Key steps

For MTPE, there existing a variety of methods to improve the quality of the machine translation results. Several efforts have been done in the Korean language and Arabic language. Scientists build up language models to identify the grammar errors of the machine translation results and offer automatic correction with rule-based models. Previous work has done analysis to figure out the underlying errors from raw machine translation. We set up a MTPE model to follow up the experiments and applying the model into the area of tourism.

4.3. Procedures

The experiment is structured into three main sections: data pre-processing, machine translation, and Machine Translation Post-Editing of the machine translation results.

DOI: 10.54254/2755-2721/34/20230286

4.3.1. Data Pre-processing. This stage involves data gathering from the Internet, focusing on tour guide scripts from typical tourism cities in China. The gathered data is then modified into a specific format that is compatible with the transformer model input. Additionally, stop words are removed to enhance the quality of the dataset.

4.3.2. Raw machine translation. In this stage, the transformer model, a classic in machine translation, is used to process the data for raw machine translation. Bilingual Evaluation Understudy (BLEU) and Translation Error Rate are set as evaluation standards. The process involves training and testing existing machine translation models with the tourism data, and the corresponding BLEU score is obtained. This score is then tracked against the increase in training steps.

4.3.3. Machine translation post-editing. This critical part of the experiment involves applying Neural Machine Translation (NMT) to the machine translation results. However, this time, the input data is the machine translation results in the English language, with the expectation that the output will be the best fit for the translation work. An MTPE model is constructed to address fixed and repeated errors and terms arising from the results of machine translation. The model is then trained and tested, and the resulting BLEU score from the MTPE model is recorded.

4.3.4. Expected results. The goal of this report is to test and substantiate the statement while constructing a simple MTPE model in the tourism sector. An increase in the BLEU score following the post-processing of the machine translation data is anticipated (Figure 1).

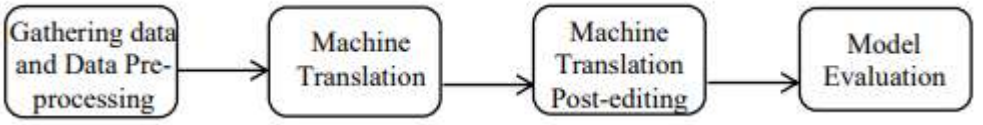

Figure 1. The Outline of Experiment Set ups (Photo/Picture credit: Original).

5. Experiment results and data analysis

The study began with an exploration of the machine learning model. During this exploration, a key finding was the remarkable accuracy of machine learning. Conventionally, manual judgement is used to evaluate the accuracy of machine translations, consuming significant resources. However, the use of the penalty mechanism of the BLEU algorithm allows for optimal length matching, and the short penalty value is calculated on a sentence-by-sentence basis before averaging.

The ability to execute the machine learning model was an initial requirement. Upon running the model, it was noted that the speed of machine learning improved significantly after multiple iterations.

Data was then introduced in a. txt format. Two text files were placed in the 'data' folder; one served as the training set for the machine learning model, while the other contained the text that required translation. The results of the machine translation and the corresponding human translation were displayed in the output field after translation.

Each machine learning iteration automatically saved the translated content to the 'log' folder. Comparison of these files revealed substantial improvements in accuracy when contrasted with the initial machine translation results.

While manual comparison was the primary method employed, the BLEU algorithm was also utilized. This algorithm, which employs an N-Gram matching mechanism, specifically assesses the similarity of N-phrases between the translation and the reference. The BLEU algorithm offers numerous advantages: it is rapid, low-cost, easy to understand, language-agnostic, and widely employed. However, it is also deficient in some areas. For instance, it fails to recognize synonyms and does not adequately account for sentence meaning and structure. Geographical nouns are also poorly handled, and identical words often result in high scores.

DOI: 10.54254/2755-2721/34/20230286

Even with these shortcomings, the BLEU algorithm's score will change relative to the value of N adjusted for the desired level of accuracy. A balance must be struck when selecting the value of N, as too small a value is not ideal. An intermediate value for N increases operational efficiency, reduces the burden on the CPU, and most importantly, enhances accuracy.

6. Conclusion

This study demonstrates that the integration of Neural Machine Translation within the Machine Translation Post-Editing process can significantly enhance the accuracy and coherence of machine translation outcomes. NMT, with its capability to process and generate human-like text, proves to be a valuable tool in refining the rough translations yielded by machine translation systems. It effectively addresses common problems encountered in machine translation, including misinterpretation, incorrect grammar, and lack of naturalness, thus improving the overall quality of the translation.

Looking forward, assessing the quality of machine translation using metrics such as BERTScore, BLEU, and TER will enable a more comprehensive and accurate evaluation. This strategy would involve comparing translations from different models, such as BERT, bi-LSTM, etc., to ascertain their respective effectiveness. BERT, with its capability of understanding the context of each word in a sentence, might provide more accurate translations especially for nuanced and culturally sensitive texts. On the other hand, Bi-LSTM, with its ability to process sequences in both directions, could potentially enhance the coherence of translated sentences by taking into account information from future states.

Further improvements in machine translation quality could be achieved by focusing on syntactical and semantic aspects of language. Deep learning models could be trained to better understand and reproduce complex linguistic structures and to infer meaning from context. Additionally, there is a need to devise sophisticated methods for handling referential words, which is a common challenge in machine translation. Such words often require understanding of the broader context to translate correctly, a task that current models struggle with.

Therefore, future research should focus on developing machine translation models that can better understand the syntactic and semantic rules of languages, effectively manage referential words, and incorporate context-awareness to produce high-quality translations. These advancements will not only enhance the translation quality but will also contribute to the wider field of Natural Language Processing.

References

[1] Xu C, Song C, et al, Improving Machine Translation and Post-editing for Chinese Tourism Texts Using Transformer-Based Models. [J]. IEEE Transactions on Industrial Electronics .2019.

[2] Li R, Jiang Z, Wang L, et al. Enhancing Transformer-based language models with Commonsense Representations for Knowledge-driven Machine Comprehension [J]. Knowledge-Based Systems, 2021(4). DOI:10.1016/j.knosys.2021.106936.

[3] Yang M , Liu S , Chen K ,et al.A Hierarchical Clustering Approach to Fuzzy Semantic Representation of Rare Words in Neural Machine Translation[J].IEEE Transactions on Fuzzy Systems, 2020, 28(5):992-1002.DOI:10.1109/TFUZZ.2020.2969399.

[4] Yang B , Wong D F , Chao L S ,et al.Improving tree-based neural machine translation with dynamic lexicalized dependency encoding[J].Knowledge-Based Systems, 2019, 188:105042.DOI:10.1016/j.knosys.2019.105042.

[5] Zhuang Y, Xu C, Song C, et al.Improving Current Transformer-Based Energy Extraction From AC Power Lines by Manipulating Magnetic Field [J]. IEEE Transactions on Industrial Electronics, 2019, PP(99). DOI: 10. 1109/TIE. 2019. 2952795.

[6] Kumar P, Pathania K, Raman B. Zero-shot learning based cross-lingual sentiment analysis for sanskrit text with insufficient labeled data [J]. Applied Intelligence, 2022, 53(9): 10096-10113. DOI:10. 1007/s10489-022-04046-6.

[7] Li Y , Li J , Zhang M .Deep Transformer modeling via grouping skip connection for neural machine translation[J].Knowledge-based systems, 2021(Dec.25):234.

DOI: 10.54254/2755-2721/34/20230286

[8] Su J , Zhang X , Lin Q ,et al.Exploiting reverse target-side contexts for neural machine translation via asynchronous bidirectional decoding[J].Artificial Intelligence, 2019, 277:103168.DOI:10.1016/j.artint.2019.103168.

[9] Nassiri K, Akhloufi M. Transformer models used for text-based question answering systems [J]. Applied Intelligence, 2022, 53(9): 10602-10635. DOI: 10. 1007/ s10489 – 022 – 04052 - 8.

[10] Tang S J , Holle J , Lesslar O ,et al.Improving quality of life post-tumor craniotomy using personalized, parcel-guided TMS: safety and proof of concept[J].Journal of neuro-oncology, 2022, 160(2):413-422.DOI:10.1007/s11060-022-04160-y.