Evaluation on Chinese to English Translation Language Model Performance on Context-Rich Text

Yujia Zhang yuz232@ucsd.edu
Dina Chen dic004@ucsd.edu
Ruixin Qiao rqiao@ucsd.edu
Nitya Davarapalli ndavarapalli@ucsd.edu

1 Introduction

Simplified Chinese is used by a significant portion of the global population. It is the official script of the People's Republic of China, Singapore, and is widely used in mainland China, Malaysia, and other Chinese-speaking communities around the world. According to estimates, mainland China has a population of over 1.4 billion people. Additionally, Singapore, with a population of over 5.8 million people, also uses Simplified Chinese as one of its official languages.

As such, there is a significant need to bridge the language barrier between Chinese and English to facilitate effective communication and collaboration between Chinese and English-speaking individuals. Chinese literature, films, art, and other cultural products have gained significant international popularity. Translating Chinese works into English allows people worldwide to access and appreciate Chinese culture, promoting cross-cultural understanding and dialogue. Among all the culture media, movies play a significant role as a cultural medium in a fast-paced society. Movies provide a platform for the representation of different cultures, traditions, and lifestyles. They showcase diverse characters, settings, and stories that reflect the richness and complexity of human experiences around the world.

Given the significance of the task, we propose to evaluate a couple machine translation models. MT(Machine translation0 can significantly enhance productivity and efficiency in handling large volumes of text. It can quickly generate translations for vast amounts of content, reducing manual effort and enabling organizations to process and analyze multilingual data at scale. Machine translation is a complex and challenging task that drives advancements in NLP techniques. Developing accurate and effective machine translation systems requires addressing various linguistic phenomena, such as grammar, syntax, semantics, and context understanding. The research and development in machine translation contribute to advancements in other NLP areas as well. We hope with the deep dive analysis of the MT models' performance and behavior on the movie subtitles, we can given some insights to improve the current technique and therefore foster a better translation result in the future.

2 Related work

In this related work section we will discuss five related papers regarding Chinese to English translation. The OpenSubtitles dataset we are using has been used in several research papers. The dataset is a large collection of user contributed subtitles in various languages for movies and TV programs. The first paper we will discuss explores the use of user-contributed subtitles in the OpenSubtitles dataset for finding alternative translations for tasks such as training paraphrase systems and creating multi-reference test suites for machine translation (Tiedemann, 2016). The paper discusses the challenges related to translation differences caused by misspellings, incomplete or corrupt data files, wrongly aligned subtitles, and alternative punctuations. The author presents a methodology for recognizing and classifying alternative subtitle translations using language-independent techniques that involve time-based alignment, lexical re-synchronization techniques, BLEU score filters, and edit distance metrics (Tiedemann, 2016). The paper resulted in a large number of sentence-aligned translation alternatives for over 50 languages and highlights the creation of cross-lingual links between subtitle files and the identification of possible errors in the collection for future cleanup.

Another paper discusses the novel approach to

automatically construct parallel discourse corpus for dialogue machine translation (Wang et al., 2016). The method the authors used involves crawling and collecting parallel subtitle data and monolingual movie script data from the internet which is then used to extract information such as speaker and discourse boundaries and projected onto the subtitle data to map monolingual discourse to bilingual texts. The paper evaluates the mapping results and explores the integration of speaker information into the translation process. The results show that the proposed method achieves high accuracy in speaker and dialogue boundary annotation and improves translation quality, demonstrating the effectiveness of speaker information in dialogue machine translation (Wang et al., 2016).

A third paper focuses on improving the benchmark for evaluating Chinese-to-English machine translation systems (Hadiwinoto and Ng, 2018). The current benchmark for Chinese-to-English translation has certain limitations. The GTC-EN benchmark that the authors introduce incorporates different styles of Chinese writing and data from various sources. The authors worked to establish a common benchmark in order to encourage Chinese-to-English MT experiments.

In the context of chat translation, one approach to improve translation quality is called Scheduled Multi-task Learning (SMTL) (Liang et al., 2022). The authors utilize a scheduled multi-task learning framework with an additional in-domain pre-training stage and a gradient-based scheduled multi-task learning strategy. Experiments on Chinese-English and English-Chinese demonstrated the effectiveness and generalizability of the framework, with significant improvements of translation quality on BLEU and TER metrics (Liang et al., 2022). The paper resulted in two large-scale in-domain paired bilingual dialogue datasets.

Another related task is Chinese-English news translation tasks. One potential architecture is a multilayer encoder-decoder architecture with attention mechanism (Wang et al., 2017). The authors sought to use ensemble and reranking techniques and to improve the named entity translation problem. For English-Chinese and Chinese-English, the system discussed in the paper improved by 3.1 to 3.5 BLEU over baseline systems by using the following techniques of a deep NMT model, ensemble of diverse deep NMT models, reranking n-best lists with NMT variant models, n-gram language models, and entity tagging and translation model (Wang et al., 2017).

We have discussed a few different machine translation models for specific tasks. We plan to keep these approaches in mind as we explore the use of recurrent neural networks and transformers for Chinese to English translation.

3 Approach

The benchmark model we want to use for the Chinese to English translation task is Recurrent Neural Network. RNNs are able to capture the dynamics of sequences via recurrent connections, which can be thought of as cycles in the network of nodes. Thus, in the context of machine translation, RNNs can be used to encode the source language sentence and generate the target language translation based on the many-to-many process. Nowadays, many transformers have surpassed the performance of RNN-based approaches. Since RNNs have a simpler architecture compared to more advanced models like transformers, it is easier for us to understand and interpret. And there will also be fewer computational resources required to do evaluation on the dataset. So we can choose RNNs as a baseline before exploring more advanced architectures.

Other than the baseline RNN, we are mostly going to focus the evaluation on Transformer based MT models. We prepare to pre-process the dataset we chose to run on the following MT models: T5, M2M100, XLM, MarianMT. By comparing and contrasting the implementation methods in the corresponding paper and source, we will first come up with a couple hypothesis regarding the layer implementation's effectiveness in the translation task. Then we will test out the hypothesis with the MT evaluation metrics like BLEU score and TER (Translation Edit Rate), combining with some linguistic aspect of assessments (including accuracy, fluency, grammatical correctness,idiomatic expressions, terminology, tone and cultural adaptation), we will analyze which model best capture the context of the dataset.

3.1 Schedule

Each member will look into a different model and its related papers, and work on analysis and final report together. Tentative schedule below:

1. Acquire and pre-process data (1 week)

2. Experiment with models using our own data (each member runs one model) (3 week)

3. Analyze and compare 4 models (1 week)

4. Work on final reports (1 week)

4 Data

The dataset we are using is a collection of translated movie subtitles.

- source:
  the dataset can be downloaded from https://opus.nlpl.eu/OpenSubtitles2018.php

- citation: http://www.opensubtitles.org/
  P. Lison and J. Tiedemann, 2016, OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles. In Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)

- Statistic: dataset contains 11,203,286 Chinese-English pair, 92.29M words.

5 Tools

We are going to mainly use the Hugging Face's Transformers library and import the required modules from it. It is an open-source library provided by Hugging Face that focuses on state-of-the-art natural language processing (NLP) models, particularly transformer-based models. The Transformers library offers easy access to a wide range of pre-trained models, including popular architectures like T5, M2M100, NLLB, XLM, MarianMT, etc. We plan to choose 4 pre-trained models and directly load them for the further machine translation task. Since it also includes utilities for handling tokenization, we will use the different tokenizers according to the specific model to encode text into numerical representations suitable for input to transformer models.

Based on the OpenSubtitles dataset, we will first use the training data to train or fine-tune the models after preprocessing. For the deep learning framework, we are going to use Pytorch as the backend for models. We may need GPUs to ensure the enough computational resources in the experimental part. Then, in order to evaluate the models' performance, we also need to use the Evaluate library from Hugging Face. It involves dozens of evaluation methods, so we can choose the available metric for different models.

6 Artificial Intelligence (AI) tools Disclosure

- Did you use any AI assistance to complete this proposal? If so, please also specify what AI you used.
  - Yes, we used the ChatGPT AI tool to search for resources.

If you answered yes to the above question, please complete the following as well:

- If you used a large language model to assist you, please paste *all* of the prompts that you used below. Add a separate bullet for each prompt, and specify which part of the proposal is associated with which prompt.
  - Why is Chinese to English translation task important(Introduction)
  - Media's importance in culture exchange(Introduction)
  - why machine translation is an important task in NLP(Introduction)
  - Some transformer based MT models (Approach)
  - RNNs in machine translation (Approach)
  - How can we evaluate models from the Transformers library? (Tools)

- **Free response:** For each section or paragraph for which you used assistance, describe your overall experience with the AI. How helpful was it? Did it just directly give you a good output, or did you have to edit it? Was its output ever obviously wrong or irrelevant? Did you use it to generate new text, check your own ideas, or rewrite text?
  - Mostly ChatGPT gives a good idea on what direction we should probe a question, so we use the result Chatgpt generates to prompt more response for greater details on a question, but we still need to edit based on the result.

References

Hadiwinoto, C. and Ng, H. T. (2018). Upping the ante: Towards a better benchmark for chinese-to-english machine translation. In *arXiv*, page 1805.01676.

Liang, Y., Meng, F., Xu, J., Chen, Y., and Zhou, J. (2022). Scheduled multi-task learning for neural chat translation. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4375–4388, Dublin, Ireland. Association for Computational Linguistics.

Tiedemann, J. (2016). Finding alternative translations in a large corpus of movie subtitles. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3518–3522.

Wang, L., Zhang, X., Tu, Z., Way, A., and Liu, Q. (2016). Automatic construction of discourse corpora for dialogue translation. In *arXiv*, page 1605.06770.

Wang, Y., Cheng, S., Jiang, L., Yang, J., Chen, W., Li, M., Shi, L., Wang, Y., and Yang, H. (2017). Sogou neural machine translation systems for WMT17. In *Proceedings of the Second Conference on Machine Translation*, pages 410–415, Copenhagen, Denmark. Association for Computational Linguistics.