# Classical Chinese (文言文) to English Translation

**Emily Huang**
University of Michigan
MS in Information
emihuang@umich.edu

**Ping-Lun Lai**
University of Michigan
MS in Data Science
alanlai@umich.edu

**Yonnie Chan**
University of Michigan
MS in Data Science
yonniech@umich.edu

## 1 Problem Statement

The project aims to address the challenge of translating Classical Chinese (文言文) into English. Classical Chinese is an ancient written form that differs significantly from all modern Chinese dialects, making it challenging for even native Chinese speakers to understand. Additionally, discrepancies in word meanings between Classical and Modern Chinese, due to linguistic nuances and historical context, increase the complexity for accurate translation, even with existing tools like Google Translate. The complexity increases when translating these texts into English for non-Chinese speakers.

Take the first sentence in the Romance of the Three Kingdoms, one of the Four Great Classical Novels of Chinese literature, for instance –

> 話說天下大勢，分久必合，合久必分。周末七國分爭，併入於秦。及秦滅之後，楚、漢分爭，又併入於漢。

As we see in Figure 1, a direct interpretation of the term **"周末"** yield *over the weekend* in Google Translate, but it actually signifies *the last reign of the Zhou Dynasty* in its original context.

The project's primary goal is to democratize access to classical Chinese literature and

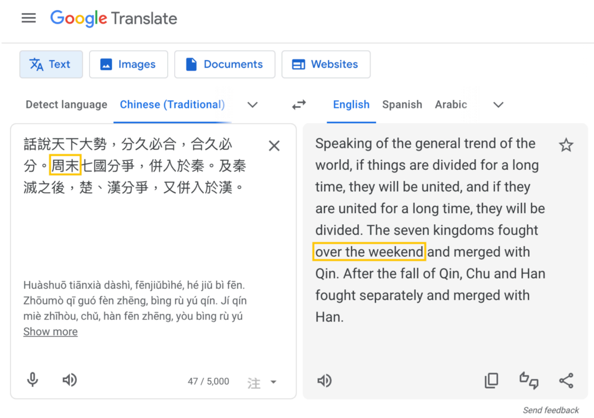


Fig. 1: Screenshot of Google Translate mistranslating the term "周末"

by developing a model to accurately translate these Classical Chinese texts into English. The project seeks to make these texts more comprehensible and accessible for a more diverse and broader audience, aiming to dismantle knowledge barriers across different languages and cultures.

## 2 Previous work

Up to this point, our survey reveals that while there are scholars dedicated to the translation of Classical Chinese to Modern Chinese, and a substantial body of work on Modern Chinese to English translation, there remains an unaddressed gap in the field—specifically, the translation of sentences or paragraphs from Classical Chinese to English. In addition, previous work focuses more on tagging. In this research, we would use two separate works as our baseline: one translating Classical Chinese to Modern Chinese, and another translating Modern Chinese to English.

### 2.1 Classical Chinese to Modern Chinese translation

The *WYWEB: A NLP Evaluation Benchmark For Classical Chinese*(Zhou et al., 2023) includes nine NLP tasks in classical Chinese, covering tasks like sentence classification, sequence labeling, reading comprehension, and machine translation. The study evaluates current pre-trained language models, highlighting their challenges with this benchmark. Additionally, supplementary datasets and tools are introduced to support further advancements in classical Chinese NLU, ultimately enhancing the capabilities of pre-trained models for this language.

### 2.2 Modern Chinese to English translation

The *Upping the ante: Towards a better benchmark for Chinese-to-English machine translation*(Hadiwinoto and Ng, 2018) presents a benchmark for evaluating Chinese-to-English machine translation systems. They have developed a novel MT system trained on a large-scale dataset, which surpasses reported results on NIST OpenMT test sets from major conferences and journals in computational linguistics and artificial intelligence over the past 11 years.

## 3 Datasets

### 3.1 Corpus Datasets for Classical Chinese

**Data Introduction** The Classical Chinese (Ancient Chinese) to Modern Chinese Parallel Corpus[1] is a comprehensive parallel corpus combining Classical Chinese (Ancient Chinese) and Modern Chinese texts. This corpus is sourced from the internet, with the original data obtained at the chapter level as bilingual-aligned data. It undergoes processing using scripts for sentence segmentation, alignment, and conversion into sentence-level parallel data. The dataset covers a total of 327 Chinese classical literature. The bilingual data comprise 97 pieces of literature, with a total of 972,467 pairs of the corpus.

**Data Preprocessing** In our approach to creating a seq2seq dataset, we separated each sentence from both Classical Chinese and its corresponding Modern Chinese translation into the list. (The result is presented in Figure 2). A significant challenge arose from the absence of punctuation in Classical Chinese sentences, whereas Modern Chinese relies on punctuation for coherent comprehension.

To address this, our dataset intentionally introduces variability by randomly removing punctuation from select sentences during the training process. This deliberate choice not only accommodates the inherent differences in punctuation usage between Classical and Modern Chinese but also enhances the model's robustness in handling diverse lin-

[1]Corpus dataset on Github:https://github.com/NiuTrans/Classical-Modern/tree/main

| | source | target |
|---|---|---|
| 0 | 假之以便，唆之使前，断其援应，陷之死地。 | 故意露出一些破绽，以引诱敌人深入我方，趁机切断他的后援和前应，最终陷他于死地。 |
| 1 | 遇毒，位不当也。 | 这就如《易经》 噬嗑 卦中说的，咬坚硬的腊肉而伤了牙齿一样，敌人为贪求不应得的利益，必招致后患。 |
| 2 | 宁伪作不知不为，不伪作假知妄为。 | 宁肯装作无知而不采取行动，不可装作假聪明而轻易妄动。 |
| 3 | 静不露机，云雷屯也。 | 要在心里暗暗谋划计策，外表不露任何声色，就像迅猛激烈的云雷在冬时隐藏地下一样地平静。 |
| 4 | 频更其阵，抽其劲旅，待其自败，而后乘之。曳其轮也。 | 多次变动敌人的阵容，把他的兵力调开，等待他自己败阵，然后用谋进攻他，好比拖住了车轮，车子就不能运行了。 |

Fig. 2: Corpus dataset

guistic patterns. This nuanced dataset stands as a testament to our commitment to providing a comprehensive resource for advancing research in machine translation and language understanding between Classical and Modern Chinese.

### 3.2 Corpus Datasets for English

**Data Introduction** We selected three classical Chinese literary works—The Analects of Confucius (論語)[2], Hung Lou Meng (紅樓夢)[3], and Tao Teh King (道德經)[4]—obtained from Project Gutenberg [5]. Our comprehensive approach involved leveraging web crawling on the HTML webpages to procure both the original Chinese texts and their corresponding English translations. This process ensures a rich and diverse dataset that spans multiple genres, providing a valuable resource for training and evaluating models in classical Chinese to English translation.

**Data Preprocessing** Upon collecting the texts, our next crucial step involves meticulous data cleaning to eliminate any inconsistencies. This includes the removal of duplicate white spaces, unwanted punctuation or symbols, and the organization of each line to delineate separate sentences or chapters. Given that discrepancies such as missing content or inconsistent chapter alignment between the Chinese and English versions can occur, a manual comparison is undertaken to ensure their alignment. The subsequent phase entails generating several accurately paired Chinese texts along with their corresponding English translations. Finally, we undertake a thorough comparison and merging of these Chinese and English datasets, culminating in the creation of a comprehensive and reliable resource for Chinese-to-English translation. This meticulous process ensures the quality and integrity of the dataset, laying a robust foundation for the development and evaluation of translation models.

[2]eBook of the Analects of Confucius:https://www.gutenberg.org/cache/epub/3330/pg3330-images.html(English),https://www.gutenberg.org/cache/epub/23839/pg23839-images.htm/(Chinese)

[3]eBook of Hung Lou Meng:https://www.gutenberg.org/cache/epub/9603/pg9603.html,https://www.gutenberg.org/cache/epub/9604/pg9604-images.html(English),https://www.gutenberg.org/cache/epub/24264/pg24264-images.html(Chinese)

[4]eBook of Tao Teh King:https://www.gutenberg.org/cache/epub/216/pg216-images.html(English),https://www.gutenberg.org/cache/epub/7337/pg7337-images.html(Chinese)

[5]eBook library:https://www.gutenberg.org/

## 4 Approaches

### 4.1 Traditional Chinese to Modern Chinese

In this project, we used Baidu Translate[6], few-shot prompting on GPT-4, and an encoder-decoder transformer language model with different pre-trained models to look for a better performance.

**Few-shot Prompting on GPT-4** We tried several different prompts to get the translation. So far, the best prompt we have is as follows:

> 你是文言文领域的专家，请根据以下输入的文言文段落或句子提供对应的现代中文翻译：

In the context above, we told GPT-4 that you are an expert in the field of classical Chinese. Please provide the corresponding modern Chinese translation based on the classical Chinese paragraphs or sentences entered below. Interestingly, if we remove the expert part but only ask for translations, the result will be slightly worse. A simple example is shown in Figure 3. Though didn't translate the word " **相**", the *with expert* result had the word in the input, while the *without expert* one did not.

**Encoder-decoder Transformer** Following the description in "Attention Is All You Need"(Vaswani et al., 2023), Figure 4 is the model architecture we used.

The encoder-decoder will then be like Figure 5. To focus on fine-tuning and not get bothered by the entire model structure, we used the pre-trained tokenizer and model on HuggingFace[7] with the help of PyTorch Lightning[8].

| model | training dataset |
|---|---|
| bert-base-chinese | |
| sikubert | Siku Quanshu |
| sikuroberta | Siku Quanshu |
| AnchiBert | 39.5 M of words in several books |

Table 1: Best parameter set

Before running into the training process, we need to create a data collator to prepare the input batches. Figure 6 shows how we make it. Using tokenizers to tokenize both classical Chinese and modern Chinese, we match them into the batch. By default, in PyTorch, the **CrossEntropyLoss** function ignores indices with the value '-100'. Thus, the value '-100' is assigned to the labels where the padding tokens are present, which is set for the model to not be penalized for its predictions on padding tokens since they don't represent real data.

The tokenizer and pre-trained models we have used are listed in Table 1. According to AnchiBert' s training description, the dataset it used is about 6 times larger than Siku Quanshu[9]. There are other Chinese-related pre-trained models on HuggingFace. Here, we focus on these four due to the restriction of training time for one epoch takes 4 hours. Besides, different models can have different usage, which increases the complexity of writing a unified code for all the models.

The best-performed model we get is AnchiBert, which is mentioned in Evaluation. After hyper-parameter tuning, the best parameter set we have so far is shown in Table 2.

[6] Baidu Translate:https://fanyi.baidu.com/

[7] HuggingFace:https://huggingface.co/

[8] PyTorch Lightning:https://lightning.ai/docs/pytorch/stable/

[9] Siku Quanshu:http://www.chinaknowledge.de/Literature/Science/sikuquanshu.html

| prompt | with “expert” | without “expert” |
|---|---|---|
| input | 郑之神巫相壶子林，见其征，告列子。 | |
| output | 郑国的神巫相壶子林，看到了某种预兆，于是去告诉了列子。 | 郑国的神巫，名叫壶子林，见到了某种征兆，于是告诉了列子。 |

Fig. 3: Prompting example

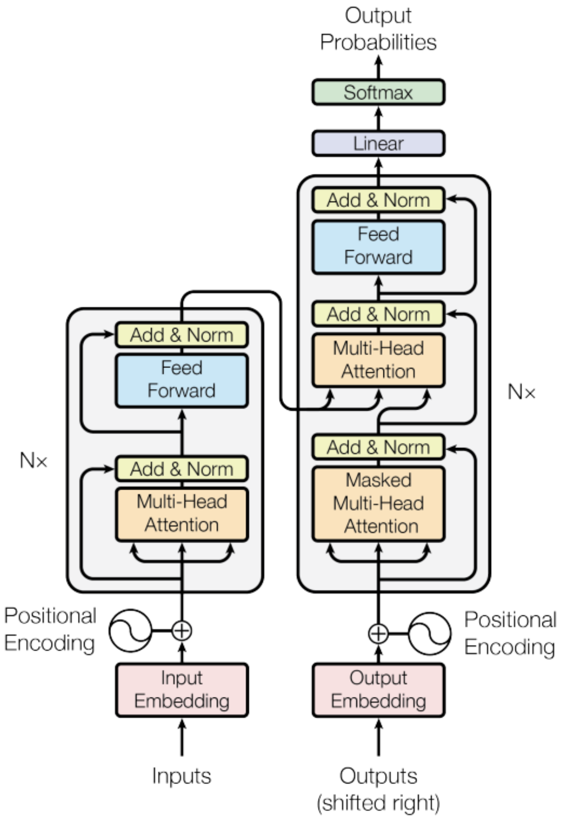


Fig. 4: The Transformer – model architecture

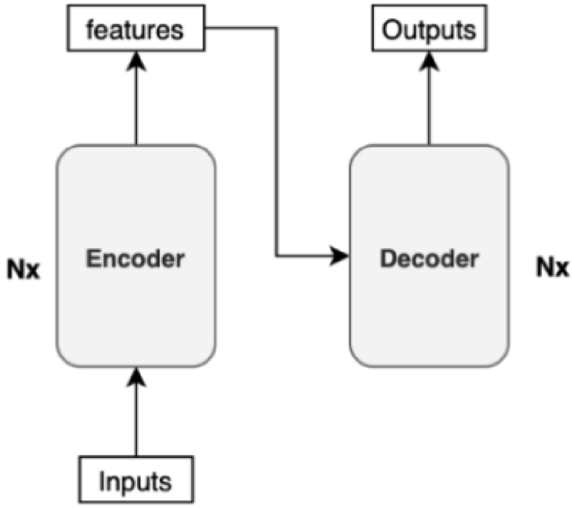


Fig. 5: The Transformer – model architecture

| Parameter | Value |
|---|---|
| epochs | 2 |
| batch_size | 24 |
| max_len | 128 |
| lr | 0.001 |
| weight_decay | 0.001 |

Table 2: Best parameter set

## 4.2 Modern Chinese to English

In this project, since Chinese-to-English translation has been studied intensively, we planned on using Google Translate[10], few-shot prompting on GPT-4, and fine-tuning the T5 model. For the following paragraphs, we selected 50 sentences in our dataset to see their performances.

**Google Translate** Many of the translations share the same issue as shown in Problem Statement. Using the same Chinese sentence we have in Figure 3 as an example, the translation result is in Figure 7. The term " 看相" is directly translated into *look at one's face*, whose meaning should be *read one's fortune*.

**Few-shot Prompting on GPT-4** The translation of GPT-4 was conducted using several different prompts. The results are encouraging, with every prompt yielding high-quality translations. Even the simplest prompts, such as “Translate to English,”have produced great results. Using the same example mentioned in the Problem Statement, both Chinese and

[10]Google Translate:`https://translate.google.com/`

```python
def collate(self, batch):
    batch_df = pd.DataFrame(list(batch))
    x, y = batch_df.source, batch_df.target
    if self.no_punkt:
        x = list(i if random.random()>0.5 else remove_all_punkt(i) for i in x)
    else:
        x = list(x)
    x_batch = self.tokenizer(
        x,
        max_length=self.max_len,
        padding='max_length',
        truncation=True,
        return_tensors='pt',
    )
    y_batch = self.target_tokenizer(
        list(y),
        max_length=self.max_len,
        padding='max_length',
        truncation=True,
        return_tensors='pt',
    )
    x_batch['decoder_input_ids'] = y_batch['input_ids']
    x_batch['labels'] = y_batch['input_ids'].clone()
    x_batch['labels'][x_batch['labels'] == self.tokenizer.pad_token_id] = -100
    return x_batch
```

Fig. 6: Data Collator

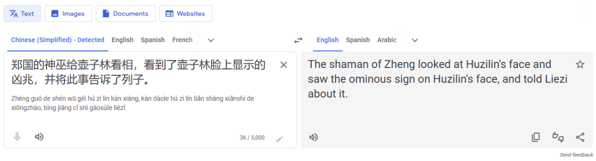


Fig. 7: Google translation of Chinese to English

English input texts were able to generate accurate results, as demonstrated in Figure 8, as demonstrated in Figure 8.

**T5 Fine-tuning** To get the fine-tuned result, we implemented the same step in the notebook[11] provided by HuggingFace, using the dataset we mentioned in Corpus Datasets for English. Following the notebook, the translation result we got is quite similar to what we have using GPT-4.

## 5 Evaluation

In the evaluation of our project, we utilize the BLEU (Bilingual Evaluation Understudy) score, a widely recognized metric for assessing the quality of machine-translated text against a set of high-quality reference translations.

The BLEU score quantitatively measures the similarity between a machine-generated translation and a set of professional-quality reference translations. It does this by comparing the n-gram overlap between the translated text and the reference text, where n-grams are contiguous sequences of 'n' words. A higher overlap indicates a higher-quality translation.

BLEU evaluates translation quality by calculating precision at the word level across different n-gram lengths (usually up to 4). It also applies a brevity penalty to discourage overly short translations. The final score is a number between 0 and 100, where higher scores denote better translations. According to the guide for evaluating models from Google Cloud Translation[12], BLEU score can be distinguished generally into 7 parts, as shown in Figure 9.

### 5.1 Traditional Chinese to Modern Chinese

After fine-tuning and few-shot prompting, the BLEU scores we get are listed in Table 3. The scores are comparative and reflect the models' ability to produce translations that are closer to a reference standard.

1. Performance Range: The BLEU scores range from 21.74 to 26.16, indicating a relatively close performance band among the models tested. This suggests that while there are differences in translation quality, all models have achieved a level of competency above a basic thresh-

[11] Notebook for T5 model translation: https://colab.research.google.com/github/huggingface/notebooks/blob/main/examples/translation.ipynb

[12] Google Cloud Translation: https://cloud.google.com/translate/automl/docs/evaluate

| | | | |
|---|---|---|---|
| prompt | You will be provided with a sentence in Chinese, and your task is to translate it into English. | Please translate the following text to English: | 將以下內容翻譯成英文: |
| input | 据说天下的形势，分离很久必然会合，联合很久必然会分离，周代末年七国纷争，都归于秦国，到秦朝灭亡之后，楚、汉两国争斗，又一起进入汉朝。 | | |
| output | It is said that the situation in the world, after a long period of separation, will inevitably unite, and after a long period of unity, will inevitably separate. At the end of the Zhou Dynasty, there were disputes among the seven states, all of which eventually belonged to the Qin state. After the demise of the Qin Dynasty, the Chu and Han states fought and together entered the Han Dynasty. | It is said that the trend of the world is such that what has been separated for a long time will inevitably unite, and what has been united for a long time will inevitably separate. At the end of the Zhou Dynasty, seven states were in conflict, all of which eventually fell to the Qin state. After the fall of the Qin Dynasty, the states of Chu and Han fought, and both eventually entered the Han Dynasty. | It is said that the trend of the world is such that what has been separated for a long time will inevitably come together, and what has been united for a long time will inevitably separate. At the end of the Zhou Dynasty, seven nations were in conflict, all of which eventually succumbed to Qin. After the fall of the Qin Dynasty, the states of Chu and Han contended, leading into the Han Dynasty. |

Fig. 8: Example of modern Chinese to English translation on different prompt

| BLEU Score | Interpretation |
|---|---|
| < 10 | Almost useless |
| 10 - 19 | Hard to get the gist |
| 20 - 29 | The gist is clear, but has significant grammatical errors |
| 30 - 40 | Understandable to good translations |
| 40 - 50 | High quality translations |
| 50 - 60 | Very high quality, adequate, and fluent translations |
| > 60 | Quality often better than human |

Fig. 9: BLEU score interpretation

old.

2. Implications for Model Development: The close scores between 'AnchiBert', 'Baidu Translate', and 'GPT-4' indicate that state-of-the-art models, whether they are specialized or more generalist like GPT-4, are pushing the boundaries of machine translation.

| Model | BLEU Score |
|---|---|
| bert-base-chinese | 21.74 |
| sikubert | 23.31 |
| sikuroberta | 22.45 |
| AnchiBert | 26.16 |
| Baidu Translate | 25.89 |
| GPT-4 | 26.03 |

Table 3: BLEU scores list

## 5.2 Modern Chinese to English Translation

According to the approach outlined in subsection 4.2, the results obtained from the three methods under consideration demonstrated very similar performance levels. Utilization of the BLEU score metric to evaluate their performance did not reveal any significant differences. Consequently, we devised an alternative ranking methodology to differentiate them more granularly.

The ranking system is defined as follows: a score of 1 is assigned for completely incorrect translations, 2 for translations that a reader could barely understand, 3 for translations that are generally correct but contain many mistakes, 4 for translations that are correct but have a few mistakes, and 5 for translations that are completely correct and accurately convey the meaning of the original text. Table 4 shows the average rank we calculated after considering the 50 sentences we selected.

Upon employing the average ranking system, a nuanced evaluation of the translation performance of T5, GPT-4, and Google Translate emerges. As illustrated in Table 4, T5 (av-

| Method | Rank |
|---|---|
| Google Translate | 3.76 |
| GPT-4 | 4.34 |
| T5 | 4.58 |

Table 4: Average rank list

erage rank of 4.58) achieved the highest average rank, indicating a relatively better performance compared to GPT-4 (average rank of 4.34) and Google Translate (average rank of 3.76). This outcome suggests that, on average, T5 outperformed its counterparts in delivering more accurate and contextually appropriate translations. The incrementally higher average rank implies a relatively better quality of translations produced by T5 across the evaluated sentences. However, T5, despite having the highest average rank, may have excelled in certain sentences but struggled in others, contributing to the overall average. Based on the evaluation results presented in Table 4, we would conclude that T5's performance is comparable to GPT-4.

## 6 Discussion

### 6.1 Result

Through rigorous evaluation and experimentation with various models, it is noticeable that models like ' AnchiBert' , ' Baidu Translate' , and ' GPT-4' exhibit commendable competency when doing Classical Chinese to Modern Chinese translation, with their BLEU scores between 25.89 and 26.16. This reflects the impressive capabilities of state-of-the-art models in pushing the boundaries of machine translation. However, the close scores suggest that while there are differences in translation quality, all models have achieved a baseline level of competency.

Moving to Modern Chinese to English translation, the average ranking system places T5 at the forefront, showcasing its relatively superior performance over GPT-4 and Google Translate. Despite the nuanced differences, these models collectively contribute to the advancement of translation technology.

### 6.2 Future Works

Moving forward, our research paves the way for future endeavors in refining and expanding the capabilities of translation models. Exploring additional language pairs and incorporating domain-specific knowledge can enhance the models' contextual understanding. Fine-tuning models based on domain expertise and leveraging advancements in pretrained models can contribute to more accurate and nuanced translations.

Furthermore, addressing specific challenges in Classical Chinese translation, such as handling missing content and historical context, can lead to tailored models that cater to the unique characteristics of Classical Chinese literature. Continuous collaboration with language experts and integrating user feedback will be integral to the iterative improvement of translation systems.

## 7 Conclusion

Our research tackles the challenging task of translating Classical Chinese texts into both Modern Chinese and English, with a focus on making these ancient literary works more accessible to a broader audience. In this research, we developed a comprehensive translation model, combining Classical Chinese to Modern Chinese translation, and Modern Chinese to English translation. Through a comprehensive evaluation using BLEU scores and average ranking, we have gained insights into the performance of various state-of-the-art models in both translation directions.

Our findings highlight the impressive competency of models like 'AnchiBert', 'Baidu Translate', and 'GPT-4' in Classical Chinese to Modern Chinese translation, showcasing their potential to bridge the gap between ancient and modern languages. In the realm of Modern Chinese to English translation, our results indicate that T5, GPT-4, and Google Translate exhibit comparable performance, with T5 demonstrating a slight edge in certain contexts.

As we look to the future, there is ample room for improvement and expansion of translation models, including the incorporation of domain-specific knowledge and addressing the unique challenges posed by Classical Chinese literature. Collaboration with language experts, iterative model refinement, and the incorporation of user feedback will be crucial for advancing the field of machine translation and making classical texts more accessible to a global audience.

## Contribution

**Emily Huang** Problem statement, gpt-4 prompt, limitation, UI for demo

**Ping-Lun Lai** Research, model development and training, evaluation, limitation

**Yonnie Chan** Data parsing, data preprocessing, discussion, conclusion

## Limitations

**Evaluation metrics** Evaluating the quality of translation models, especially for a language as nuanced and context-dependent as Classical Chinese, is indeed challenging due to the absence of universally agreed-upon translation standards.

**BLEU** While BLEU is an effective tool for quantitative evaluation, it has limitations. It primarily focuses on precision and doesn't account for the semantic accuracy of the translation. Additionally, BLEU scores can sometimes be misleading for very short or very long texts. Therefore, while BLEU scores provide a valuable quantitative measure of translation quality, they should be considered alongside other qualitative assessments.

**Linguistic nuances** One of the most significant challenges faced is the linguistic complexities, historical context, and nuanced meanings of Classical Chinese. Being able to accurately translate ambiguous phrases or words often heavily relies on context, however, our model might not capture these subtleties.

**Model generalization** There are many different genres of Classical Chinese literature, such as poetry, historical texts, and philosophical works, and each genre has wildly different writing styles, increasing difficulties when translating. The model' s ability to handle a diverse range of texts still needs to be further developed and evaluated.

**Culture context** Translating Classical Chinese into English involves more than just a literal translation of words. It requires a deep understanding of the cultural context, which can be challenging for a machine-learning model to grasp.

## Acknowledgement

For the complete code implementation, please refer to our GitHub repository: Classical-Chinese-to-English-Translator.

## References

Christian Hadiwinoto and Hwee Tou Ng. 2018. Upping the ante: Towards a better benchmark for chinese-to-english machine translation.

Kenton Lee Kristina Toutanova Jacob Devlin, Ming-Wei Chang. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding.

Z. Jiang, J. Wang, J. Cao, X. Gao, and L. Jin. 2023. Towards better translations from classical to modern chinese: A new dataset and a new method. In *Natural Language Processing and Chinese Computing. NLPCC 2023*, volume 14302 of *Lecture Notes in Computer Science*. Springer.

Chuanqiang Zhang Zhongjun He Hua Wu Zhi Li Haifeng Wang Ying Chen Qinfei Li Ruiqing Zhang, Xiyang Wang. 2021. Bstc: A large-scale chinese-english speech translation dataset.

Naoki Shibuya. 2021. Transformer's encoder-decoder. `https://kikaben.com/transformers-encoder-decoder/`. Accessed: 2023-12-13.

Nick Ryder Melanie Subbiah Jared Kaplan Prafulla Dhariwal Arvind Neelakantan Pranav Shyam Girish Sastry Amanda Askell Sandhini Agarwal Ariel Herbert-Voss Gretchen Krueger Tom Henighan Rewon Child Aditya Ramesh Daniel M. Ziegler Jeffrey Wu Clemens Winter Christopher Hesse Mark Chen Eric Sigler Mateusz Litwin Scott Gray Benjamin Chess Jack Clark Christopher Berner Sam McCandlish Alec Radford Ilya Sutskever Dario Amodei Tom B. Brown, Benjamin Mann. 2020. Language models are few-shot learners.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need.

Beilei Xiang, Changbing Yang, Yu Li, Alex Warstadt, and Katharina Kann. 2021. CLiMP: A benchmark for Chinese language model evaluation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2784–2790, Online. Association for Computational Linguistics.

Jian Guan Boxi Cao Zhengyan Zhang Chaojun Xiao Xiaozhi Wang Fanchao Qi Junwei Bao Jinran Nie Zheni Zeng Yuxian Gu Kun Zhou Xuancheng Huang Wenhao Li Shuhuai Ren Jinliang Lu Chengqiang Xu Huadong Wang Guoyang Zeng Zile Zhou Jiajun Zhang Juanzi Li Minlie Huang Rui Yan Xiaodong He Xiaojun Wan Xin Zhao Xu Sun Yang Liu Zhiyuan Liu Xianpei Han Erhong Yang Zhifang Sui Maosong Sun Yuan Yao, Qingxiu Dong. 2021. Cuge: A chinese language understanding and generation evaluation benchmark.

Bo Zhou, Qianglong Chen, Tianyu Wang, Xiaomi Zhong, and Yin Zhang. 2023. Wyweb: A nlp evaluation benchmark for classical chinese.