EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff

Dongbo Wang db.wang@njau.edu.cn
Litao Lin litaolin@njau.edu.cn
Zhixiao Zhao 2022114011@stu.njau.edu.cn
Wenhao Ye yewenhao@njau.edu.cn
College of Information Management, Nanjing Agricultural University, Nanjing, 210031, China
Kai Meng mengkai@njau.edu.cn
School of Marxism, Nanjing Agricultural University, Nanjing, 210095, China
Wenlong Sun 287971655@qq.com
School of Foreign Languages, Nanjing Tech University, Nanjing, 211816, China
Lianzhen Zhao buddy_zlz@163.com
School of Foreign Languages, China Pharmaceutical University, 211198, China
Xue Zhao 741081584@qq.com
College of Information Management, Nanjing Agricultural University, Nanjing, 210031, China
Si Shen shensi@njust.edu.cn
School of Economics & Management, Nanjing University of Science and Technology, 210094, China
Wei Zhang 292204510@qq.com
College of Information Management, Nanjing Agricultural University, Nanjing, 210031, China
Bin Li (Corresponding author: libin.njnu@gmail.com)
School of Chinese Language and Literature, Center of Language Big Data and Computational Humanities, Nanjing Normal University, 210097, China

Abstract

This paper presents the results of the First International Ancient Chinese Transalation Bakeoff (EvaHan), which is a shared task of the Ancient Language Translation Workshop (ALT2023) and a co-located event of the 19th Machine Translation Summit 2023 (MTS 2023). We described the motivation for having an international shared contest, as well as the datasets and modalities. The contest consists of two modalities, closed and open. In the closed modality, the participants are only allowed to use the training data, and the participating teams achieved the highest BLEU scores of 27.33 and 1.11 in the tasks of translating ancient Chinese to Modern Chinese and translating Ancient Chinese to English, respectively. In the open mode, contestants can use any available data and models. The participating teams achieved the highest BLEU scores of 29.68 and 6.55 in the Ancient Chinese to Modern Chinese and Ancient Chinese to English tasks, respectively.

Proceedings of ALT2023: First Workshop on Ancient Language Translation, pages 1–14
September 5, 2023, Macau SAR, China.

1. Introduction

As an important carrier of Chinese traditional culture, Ancient Chinese classics is of great value in historical and literary study. Through the translation of Ancient Chinese classics, excellent traditional Chinese culture can be passed on to contemporary readers and the international community, promoting cross-cultural communication and understanding. However, the fact that the morphology, syntax and lexical meaning of Ancient Chinese are quite different from those of Modern Chinese makes it difficult for Modern Chinese translation technology to achieve better results in Ancient Chinese translation.

Machine translation is mainly divided into methods based on statistics and rules, among which methods based on statistics are the main ones. At present, machine translation research mainly focuses on the interlingual translation of modern languages, while there are few studies on the translation of ancient languages. The translation of Ancient Chinese into Modern Chinese is a special kind of intralingual translation, and there are few related studies at present. The lack of parallel corpora between ancient and Modern Chinese is a key factor restricting the research of Ancient Chinese machine translation (Han et al., 2015). The advancement of neural network models such as Transformer (Vaswani et al., 2017) and BERT(Devlin et al., 2019) has spawned a batch of pre-trained language models for Ancient Chinese. Combined with prompt learning technology (Liu et al., 2021), the successful practice of GPT-like large-scale language models has brought new development opportunities for machine translation research in low-resource languages such as Ancient Chinese (Liu et al., 2023).

In the past period of time, there have been many machine translation evaluation competitions for different languages or domainized texts, such as WMT'22 Shared Task on Large-Scale Machine Translation Evaluation for African Languages (Srivastava & Singh, 2022). Many inspiring results have emerged, including machine translation models (Adelani et al., 2022; Kocmi et al., 2022) and translation effect evaluation methods (Freitag et al., 2022). However, there is still a lack of machine translation evaluation competitions for Ancient Chinese. In this context, we held the second EvaHan event (EvaHan2023): The International Ancient Chinese Transalation bakeoff[1]. EvaHan is a series of international evaluation focusing on the information processing of Ancient Chinese (Li et al. 2022). EvaHan2023 is a shared task of the Ancient Language Translation Workshop (ALT2023), which will be held as a co-located event of the 19th Machine Translation Summit 2023 (MTS 2023) in Macao SAR, China.

EvaHan

2023 is the second campaign devoted to the evaluation of Natural Language Processing (NLP) systems for Ancient Chinese with the following aims:

- To investigate the applicability of current MT techniques in Ancient Chinese translation.
- To examine the significant challenges in Ancient Chinese translation (e.g. word order and syntax errors).
- Provide a platform for the enthusiasts of machine translation in Ancient Chinese
- To facilitate machine translation research for Ancient Chinese and the exploration of forefront machine translation technology.

[1] https://github.com/GoThereGit/EvaHan

2. Task

EvaHan2023 consists of two translation tasks: Ancient Chinese to Modern Chinese (a2m) and Ancient Chinese to English (a2e).

- Ancient Chinese to Modern Chinese machine translation is the process of translating Ancient Chinese sentence in traditional Chinese characters to Modern Chinese in traditional Chinese characters or simplified Chinese characters.
- Ancient Chinese to English machine translation is the process of translating Ancient Chinese sentence in traditional Chinese characters to English.

All tasks require the original sentences to be automatically converted into target language sentences without human assistance. EvaHan2023 allows the teams to submit translation results in one or two of the above two target languages at the same time. Since Hong Kong, Macao and Taiwan regions of China use Modern Chinese writing characters in traditional form, while mainland of China uses simplified format, in the ancient-to-modern translation task, the participating teams are allowed to submit results in either traditional or simplified format. In the stage of translation quality evaluation, EvaHan2023 uses the text in the same language as the submitted results as a reference. Table 1 shows the forms of original sentences and three kinds of target language sentences.

Tasks	Source Language Sentences	Target Language Sentences
a2m (traditional format)	殘諂之吏，張設機網，並驅爭先，若赴仇敵。	殘暴諂媚的的執法官吏，張開羅網，設立陷阱，並駕齊驅，爭先恐後，好似追趕仇敵一樣。
a2m (simplified format)	殘諂之吏，張設機網，並驅爭先，若赴仇敵。	残暴谄媚的的执法官吏，张开罗网，设立陷阱，并驾齐驱，争先恐后，好似追赶仇敌一样。
a2e	殘諂之吏，張設機網，並驅爭先，若赴仇敵。	Cruel and slanderous officials have spread broad nets for me, and they encourage one another against us. It is as if they pursued an enemy.

Table 1: Examples of Modern Chinese Translation and English Translation

3. Dataset

The datasets of EvaHan2023 consists of two parts: training dataset and test dataset. The training dataset, with both Ancient Chinese source text and corresponding Modern Chinese reference translation and English reference translation, is provided for participating teams to train and validate their machine translation models. The test dataset is used to scoring and ranking the machine translation models performance of the participating teams, consisting of Ancient Chinese source text provided to participating teams and corresponding Modern Chinese reference translation and English reference translation remained by conference affairs before the submission deadline.

3.1. Data Format

All evaluation data are .txt files in Unicode (UTF-8) format, arranged by two fields of source language and target language to form a sentence level parallel corpus, as shown in Table 2 and Table 3.

Table 2 shows examples of the Ancient Chinese to Modern Chinese parallel corpus. The left column is the Ancient Chinese text, while the right column is the corresponding Modern Chinese (traditional Chinese format) texts.

Ancient Chinese	Modern Chinese (Traditional Chinese format)
后妃表	后妃表
后妃之制，厥有等威，其来尚矣。	后妃的制度，有它的等级威儀，它的由來很久遠。
元初，因其國俗，不娶庶姓，非此族也，不居嫡選。	元朝初年，因襲蒙古的習俗，不娶異姓，不是后族的，不處在可以選爲正妻的地位。
當時使臣為舅甥之貴，蓋有周姬、齊姜之遺意，歷世守之，因可嘉也。	當時的史臣以爲皇族后族的尊貴，原有周姬、齊姜的遺意，歷代都遵守它，本來是可以表彰的。

Table 2 : Examples of the Ancient Chinese to Modern Chinese (Traditional Chinese format) corpus

Table 3 shows examples of the Ancient Chinese to English parallel corpus. Sentences on the left side is in Ancient Chinese, and on the right side is in corresponding English.

Ancient Chinese	English
杜密素與李膺名行相次，起，對之揖，勸令從學。	Du Mi had shared in reputation with Li Ying, He stood up and bowed to him, then urged him to study.
濟陰黃允，以俊才知名。	Huang Yun of Jiyin was known for his outstanding talents.
兵士喜悅，大小皆出。	Officers and men were delighted, and they all went out to take part.

Table 3 : Examples of the Ancient Chinese to English corpus

3.2. Training data

Training data is excerpted from the *Twenty-Four Histories*(dynastic histories from remote antiquity till the Ming Dynasty), the Pre-Qin classics and *ZiZhi TongJian* (资治通鉴, Comprehensive Mirror in Aid of Governance). The *Twenty-Four Histories* is the general name of the twenty-four official histories of various dynasties in ancient China; the Pre-Qin classics are the historical materials of the Pre-Qin period (Paleolithic Period ~ 221 B.C.), which have an important position in ancient books, including history books and sub-books; *ZiZhi TongJian* is a chronological history book compiled by historians of the Northern Song Dynasty, covering sixteen dynasties from 403 B.C. to 959 A.D. over a span of 1362 years. The ancient Chinese classic texts in the corpus feature both diachronicity (i.e. spanning thousands of years) and synchronicity (i.e. covering the four traditional types of Chinese canonical texts *Jing* (经), *Shi* (史), *Zi* (子) and *Ji* (集)).

Descriptions about the overall parallel texts for machine translation are presented in Table 4.

Parallel Data	Source Data scale	Target Data scale
Ancient Chinese to Modern Chinese parallel texts of *Twenty-four Histories*	9,583,749 characters	12,763,534 characters

Ancient Chinese to English parallel texts of Pre-Qin canonical texts and *Zizhi Tongjian*	618,083 characters	838,321 words

Table 4 : Details of training data in EvaHan2023

3.3. Test Data

The test dataset for evaluation consists of 2,071 Ancient Chinese sentences with the corresponding translations in Modern Chinese and English. The Ancient Chinese sentences in test data is excerpted from the *HouShan TanCong(后山谈丛)* and *Jin Lou Zi(金楼子)*. The Modern Chinese and English translations of *HouShan TanCong* are firstly translated through Baidu's classical Chinese translation function, and then proofreaded and perfected by Chinese classical literature experts and English experts. *Jin Lou Zi*'s Modern Chinese translation comes from *Jin Lou Zi*'s translation and commentaries. The English translation was initially obtained through Baidu's classical Chinese translation function, and then proofread and perfected by three English experts.

4. Evaluation

4.1. Scoring Metrics

EvaHan2023 applies BLEU and CHRF to evaluate the quality of submitted translations.

BLEU: BLEU (Bilingual Evaluation Understudy) (Papineni et al., 2002) is an indic ator for automatically evaluating the quality of machine translation. By comparing the machine translation results with the reference translation, the degree of n-gram overlap and the number of matches are calculated to obtain the quality score of the machine translation. Generally, a higher BLEU score indicates that the machine translation result is closer to the reference translation.

Although BLEU has been widely used to evaluate the quality of machine translation, different studies have different optional parameter settings for BLEU (Post, 2018). SacreBLEU(Post, 2018) aims to solve the above problems. It is a toolkit for machine translation quality assessment, developed by Facebook AI Research, which provides a set of parameter setting schemes for standard data sets, and stipulates different language texts and word segmentation algorithm. In order to evaluate the translation quality of each model more reasonably, this study uses SacreBLEU as a specific evaluation tool, setting tokenizer to 'char', and the rest of the parameters remain at default values. EvaHan applies the offical SacreBLEU project (Post, 2017/2023) to do the evaluation.

CHRF: CHRF(Character n-gram F-score) (Popović, 2015), an indicator for evaluating the quality of machine translation systems, is mainly used to evaluate the similarity between machine translation output and reference translation.

The biggest difference beteen BLEU and CHRF is that CHRF evaluates the translation quality in units of words, while BLEU is a word-level translation quality evaluation method. Compared with BLEU, CHRF has the following advantages: First, it can better capture phenomena such as phrase matching and word reordering; second, it can better evaluate the performance of machine translation models in lower-resource language pair translation tasks; in addition, CHRF can also balance precision and recall by using different n-gram sizes. In this study, the word-level n-gram size of CHRF is set to 6, and the smoothing function is selected as "exponential decay". EvaHan applies the offical CHRF code (*ChrF - a Hugging Face Space by Evaluate-Metric*, n.d.) to do the evaluation.

4.2. Two Modalities

Each participant can submit runs following two modalities. In the closed modality, the resources each team could use are limited. Each team can only use the training data and the following pre-trained models listed in Table 5. Other resources are not allowed in the closed modality.

Pre-Trained Model	Language	Description
SikuRoBERTa[2]	Ancient Chinese	Ancient Chinese RoBERTa pre-trained on high-quality *Siku Quanshu*(四库全书) full-text corpus.
Chinese-RoBERTa-wwm-ext[3]	Modern Chinese	Modern Chinese pre-trained RoBERTa with Whole Word Masking strategy.
RoBERTa[4]	English	Pre-trained model on English with MLM objective.

Table 5 : Pre-trained models for closed modality

In the open modality, however, there is no limit on the resources, data and models. Annotated external data, such as the components, Pinyin of the Chinese characters, word embeddings, dictionaries, knowledge graphs, etc. can be employed. But each team has to state all the resources, data and models they use in each system in the final report and manual corrections for translaton results are not allowed.

4.3. Procedures

The open registration period for the competition is from February 15th to March 25th, 2023. The training data set will be available for download on April 1, 2023 and test dataset released on June 7, 2023. Deadlines for submission of translation results and technical reports are June 22 and June 31, 2023, respectively. The evaluation results of each team were returned on June 23, 2023. The deadline for submitting the Camera Ready version of technical reports is July 15, 2023. EvaHan2023 is held at the ALT2023 workshop co-located with the MTS2023 conference in Macao on September 5th, 2023.

5. Participants and Results

5.1. Participants

Table 6 gives the basic information of the participating teams and their submitted results. A total of 9 teams took part in EvaHan2023, submitting 18 translation results. Among them, 8 teams are from colleges and universities, and one is an individual team. All teams submitted translation results in Modern Chinese, two of which submitted translation results in Simplified Chinese. There are four teams that submitted English translation results.

Team \ Task			a2m (traditional)		a2m (simplified)		a2e	
ID	Name	Affiliation	C	O	C	O	C	O
1	BIT	Beijing Institute of Technology	1	0	0	0	1	0

[2] https://huggingface.co/SIKU-BERT/sikuroberta
[3] https://huggingface.co/hfl/chinese-roberta-wwm-ext
[4] https://huggingface.co/roberta-large

2	CUHK	The Chinese University of Hong Kong	1	0	0	0	1	0
3	ISTIC	Institute of Scientific and Technical Information of China	0	1	0	0	0	1
4	L&C	Individual	1	0	0	0	0	0
5	NJU	Nanjing University	1	0	0	0	0	0
6	NJUCM	Nanjing University of Chinese Medicine	0	1	0	1	0	1
7	PKU	Peking University	2	2	0	0	0	0
8	SCUT	South China University of Technology	0	0	0	2	0	0
9	USST	University of Shanghai for Science and Technology	1	0	0	0	0	0
Total Files		18	7	4	0	3	2	2

Table 6 : Result submission status of participating teams in closed (C) and open (O) modalities

5.2. Results

EvaHan2023 uses the BLEU score as the ranking basis, and lists the CHRF score as a reference for readers. Both BLEU and CHRF are calculated in units of corpus, instead of calculating the scores of each single sentence and then averaging them. Among the 18 submitted results, some teams achieved excellent translation results. This paper presents the results of the contest according to different tracks and modalities.

Table 7 and Table 8 show the results of each team in the translation task from Ancient Chinese to Modern Chinese (traditional form). In the closed mode, CHUK scored the best, with a BLEU score of 26.76. In the open mode, ISTIC has the best score, with a BLEU score of 24.34, which is only about 0.17 points ahead of PKU.

Team	BLEU	CHRF
CUHK_1	26.7634	24.5946
PKU_2	24.1719	22.0529
PKU_1	24.1629	22.0451
NJUNLP_1	22.0524	20.5356
BIT_1	21.9485	20.5911
USST_1	21.7537	20.1962
Lemontree_1	20.7738	19.6607

Table 7: The performance of each team in the translation from Ancient Chinese to Modern Chinese (traditional Chinese format) in the closed modality

Team	BLEU	CHRF
ISTIC_2	24.3419	21.4651

PKU_2	24.1719	22.0529
PKU_1	24.1629	22.0451
NJUCM_2	7.3135	10.0544

Table 8 : The performance of each team in the translation from Ancient Chinese to Modern Chinese (traditional Chinese format) in the open modality

In the open modality, two teams submitted three translation results of Modern Chinese in simplified format. We evaluated the translation results of the three submitted Simplified Modern Chinese translations using the Simplified Modern Chinese format test dataset reference translations. At the same time, we also converted other Modern Chinese translation results submitted in the form of traditional Chinese in the open modality into simplified format for re-evaluation, and jointly presented them in Table 9. Their simplified Modern Chinese translations converted from the results submitted in traditional Chinese.

It can be seen from Table 9 that SCUT has achieved the best results, and the BLUE score is 29.68. The results of the two simplified Modern Chinese forms submitted by SCUT are better than the simplified results obtained by converting the traditional Chinese results of PKU. In addition, as far as PKU submitted results in traditional Chinese, after converting to simplified Chinese, the evaluation score of the translation results improved.

Team	BLEU	CHRF
SCUT_2	29.6832	26.1363
SCUT_1	29.5355	26.0515
PKU_2*	26.6438	24.0231
PKU_1*	26.5925	23.9902
ISTIC_2*	24.9170	21.9074
NJUCM_1	9.3807	11.1137

Table 9 : The performance of each participating team in the translation from Ancient Chinese to Modern Chinese (simplified format) in the open modality(The translation result of the team marked with * is converted from their submitted traditional format results)

In this competition, no team submitted results of Modern Chinese translation in simplified format in closed modality. In this regard, we converted all the submitted results in the traditional format in the closed mode to the simplified format, and conducted evaluation with reference to the translation in the simplified format, and the results are shown in Table 10. Compared with the evaluation results of Modern Chinese in traditional form, the top three and their rankings have not changed. CUHK still ranks the first, and the rankings of BIT and L&C have risen.

Team	BLEU	CHRF
CUHK_1*	27.3315	25.0665
PKU_2*	26.6438	24.0231
PKU_1	26.5925	23.9902
BIT_1*	24.3132	22.4501
NJUNLP_1*	24.0682	22.1297
Lemontree_1*	22.5412	21.0501
USST_1*	22.2126	20.5707

Table 10 : The performance of each team in the translation from Ancient Chinese to Modern Chinese (simplified Chinese form) in the closed modality (The translation result of the team marked with * is converted from their submitted traditional format results)

Table 11 and Table 12 show the socres of each team in the Ancient Chinese to English translation task. Under the closed modality, a total of two teams submitted two translation results. The BLEU values are 1.11 and 1.08 respectively, and CHUK's score is slightly better. Under the open modality, two teams also submitted two translation results. The BLEU values are 6.55 and 3.00 respectively, and ISTIC has achieved a clear advantage. The results of the teams on the open modality are significantly better than those on the closed modality.

Team	BLEU	CHRF
CUHK	1.1102	24.2297
BIT	1.1084	23.0841

Table 11 : The performance of the participating teams in the translation from Ancient Chinese to English in the closed modality

Team	BLEU	CHRF
ISTIC	6.5493	26.4452
NJUCM	3.0024	22.8333

Table 12 : The performance of each team in the translation from Ancient Chinese to English in the open modality

5.3. Comparison with Baselines and Toplines

In order to more intuitively present the pros and cons of the translation models and translation results, EvaHan set a baseline and a topline as references. EvaHan2023 selects the single-layer Transformer (Vaswani et al., 2017) model as the baseline model. The parameters of the Transformer model are set as follows: The embedding size or dimensionality of the input tokens is 512; The number of attention heads in the multi-head attention mechanism of the Transformer model is 8; The hidden dimension size of the feed-forward neural network (FFN) within the Transformer model is 512; The number of layers in both the encoder and decoder stack of the Transformer model are 3. EvaHan2023 uses the same training corpus specified under the closed modality to train the Transformer translation model for different tasks, and uses its translation results as the baseline reference translation.

As for the topline, EvaHan2023 selects Baidu's classical Chinese translation API as the topline model, and uses BLEU and CHRF to evaluate the traditional, simplified and English translations. For the English translation results, EvaHan2023 also adds the English translation results obtained by Google's general translation function as a topline reference.

The test data of the baseline model and the topline model are the same test data provided to the participating teams.

5.3.1 Translation from Ancient Chinese to Modern Chinese

Table 13 shows the comparison of the best model with the baseline and topline under the open modality and closed modality. In the task of translating Ancient Chinese to traditional form of Modern Chinese, the Transformer-based translation model was trained for 3 rounds using the parallel corpus of *Twenty-Four Histories*, which consists of 300,000 Ancient Chinese and traditional Chinese translation sentence pairs. Comparing with Table 7 and Table 8, it can be found that all submitted traditional Chinese translation results outperformed the baseline model's translation results. In the open modality, only one team scores below baseline, and no participating team exceeds the topline. Under the closed modality, only CUHK scores more than the topline.

Model Type	Model	BLEU	CHRF
Baseline Model	Transformer	8.9554	10.3511
Topline Model	Baidu Classical Chinese Translation	24.9731	23.0771
Best model in open modality	ISTIC_2	24.3419	21.4651
Best model in closed modality	CUHK_1	26.7634	24.5946

Table 13 : Baseline and topline effects from Ancient Chinese to Modern Chinese (traditional form)

Table 14 shows the baseline and topline in the task of translating Ancient Chinese to simplified format of Modern Chinese, as well as the best results under the open and closed modalities. In the task of translating Ancient Chinese to simplified Modern Chinese, the Transformer translation model was trained for 10 rounds based on 5,899 Ancient Chinese original text from pre-Qin classics and *Zizhi Tongjian* and their corresponding simplified Modern Chinese translation sentence. For the Transformer translation model, 5,899 sentence pairs are not enough to achieve sufficient training, so no good translation results have been achieved. Considering the scores of the translated results in simplified format, there are two teams with a total of 4 submissions exceeding the topline under the open modality; and two teams with a total of 3 submissions exceeding the topline under the closed modality.

Model Type	Model	BLEU	CHRF
Baseline Model	Transformer	9.0368	11.2385
Topline Model	Baidu Classical Chinese Translation	25.5667	23.5617
Best model in open modality	SCUT_2	29.6832	26.1363
Best model in closed modality	CUHK_1*	27.3315	25.0665

Table 14 : Baseline and topline effects from Ancient Chinese to Modern Chinese (simplified format) (The translation result of the team marked with * is converted from their submitted traditional format results)

5.3.2. Translation from Ancient Chinese to English

Table 15 shows the best results for the tasks translated from Ancient Chinese to English under open and closed modalities, as well as the results for the baseline model and the topline model. In this task, the training corpus of baseline model Transformer is 5,899 ancient English parallel sentence pairs of pre-Qin classics and *Zizhi Tongjian*. According to the data in Table 11 and Table 12, the best results on either closed or open modalities did not exceed topline, all teams performed better than the baseline under the closed modality, but neither the closed modality nor the open modality had a better performance than topline. On the whole, the effect of Ancient Chinese to English translation is not as good as that of Ancient Chinese translation to Modern Chinese translation.

Model Type	Model	BLEU	CHRF
Baseline Model	Transformer	0.8901	18.2355
Topline Model①	Baidu Classical Chinese Translation	12.3526	34.5937
Topline Model②	Google Translation	10.7757	31.6446
Best model in open modality	ISTIC_2	6.5493	26.4452

Best model in closed modality	CUHK_1	1.1102	24.2297

Table 15 : Baseline and topline effects from Ancient Chinese to English translation

5.4. Models and Methods

ISTIC: ISTIC uses Transformer as the basic structure of the translation model. ISTIC uses a variety of data preprocessing methods to optimize the quality of the training data set, including removing repetitive sentences, converting traditional characters to simplified ones, unifying punctuation marks, filtering sentence length ratios, and encoding Chinese characters in pinyin. In terms of data enhancement, the team first built an initial model using the training corpus provided by EvaHan, and then used the above model to translate Ancient Chinese data collected from the Internet to form a new parallel corpus. The experimental results show that the new parallel corpus obtained by the above method has a positive effect on improving the performance of the translation model.

BIT: BIT uses Transformer as the basic structure of the translation model. BIT performs word segmentation processing on Ancient Chinese and Modern Chinese in the training corpus, thus constructing a machine translation model encoded in word units. In the data preprocessing part, the team discarded sentences that were too long, taking into account the structural character ISTICs of the model used. In terms of data enhancement, the team first trained a translation model from Modern Chinese to Ancient Chinese based on the training data from Ancient Chinese to Modern Chinese provided by EvaHan, and then used the above-mentioned model to translate Modern Chinese to Ancient Chinese in the training set, thus constructing a new The parallel corpus, the experimental results show that the new data set constructed by this method is beneficial to improve the performance of the model. Based on the expanded new data, the team performed a second round of data augmentation, but experiments found that the second round of data augmentation weakened the performance of the model.

CHUK: CHUK uses RoBERTa and SikuRoBERTa as encoders from Ancient Chinese to English and Ancient Chinese to Modern Chinese respectively, and uses Beam Search to decode the encoded results to obtain translation results.

NJUCM: Based on the Mrasp model, this study fine-tunes the parallel corpus of the Twenty-Four Histories and the ancient English parallel corpus of *Zizhi Tongjian*, so as to evaluate the task. As the application of BERT model in the field of machine translation, the mRASP model uses bilingual parallel corpus in multiple languages for combined training, so that the model fully learns the knowledge of single language and translation between languages. The design idea of the model is similar to the model training fine-tuning paradigm in the current era of large models, and its pre-training task is similar to the downstream task, which can give full play to the performance of the model. However, since there is no Ancient Chinese to English parallel corpus in the pre-training corpus of the mRASP model, its effect is poor in the ancient Chines English translation task. Further increasing the scale of the training data may improve this.

PKU: In this study, the data augmentation method of merging adjacent sentences from the same chapter is adopted, so that the model learns richer contextual information during the training process. At the same time, the first six layer parameters of SikuBERT are used as infrastructure for model building. During the training process, the model training is detected in real time through BLEU and ChrF scores, and the model training is stopped in time, which improves the efficiency of model training to a certain extent. Through data enhancement and model fine-tune, this study has achieved better performance in downstream tasks, and at the same time, it also proves that the data augmentation method adopted in this study has certain feasibility and effectiveness.

SCUT: The study applies large-scale language models to ancient Chinese translation tasks, and performs word list expansion, incremental training and large-scale fine-tuning on the basis of the LLaMA model, and uses a larger training data. Finally, the fine-tuned output of the Ziya-13B(Zhang et al., 2023), a variant of the LLaMA (Touvron et al., 2023), was used as the competition result, and the superior effect was achieved.
NJU: This study used all Twenty-Four Histories corpus for training, and the whole workflow was a standard machine translation process, which employed a standard Transformer-based natural machine translation architecture with the pre-trained model, Chinese-RoBERTa-wwm-ext as the encoder and a randomly initialized decoder. In this study, a relatively large corpus and a join dictionary were used, and only one embedding vector was needed for the same word, which achieved good results while improving training efficiency.
USST: The team built a translation model based on the Transformer architecture, used Siku-Roberta to encode the ancient text, and introduced the method of alternate initialization from Deltalm to initialize the decoder parameters. The process also used BPE-drop to enhance the parallel corpus.

Taken together, pre-trained language models such as SikuRoBERTa and RoBERTa are widely used as encoders. There are also teams that do not use the pre-trained language model in the encoding stage, and achieve good results by refining the training data and adding external knowledge such as pinyin, using the improved Transformer model. There is also a team that applies large-scale language models to Ancient Chinese machine translation. By optimizing the vocabulary, improving the random initialization scheme of unregistered words, and then performing domain-based fine-tuning training on large-scale language models through parallel corpora, they successfully constructed a translation model for Ancient Chinese, and achieved excellent results.

In any case, the scale of high-quality parallel corpora is always a key factor affecting the performance of translation models. In this competition, almost all teams enhanced the training data, and because of the lack of parallel corpus from Ancient Chinese to English, most teams did not achieved the desired effect in the translation task from Ancient Chinese to English.

6. The problems of the text era characteristics

In order to explore the influence of text era characteristics on the performance of machine translation, we split the test data into two parts: *Houshan Tancong* and *Jin Lou Zi*, and re-evaluated the traditional Chinese translation results submitted by CHUK. Table 16 shows evaluation results. From the data in Table 16, we can see that CUHK's translation results of *Houshan Tancong* are significantly better than *Jin Lou Zi*'s. *Houshan Tancong* was written in the Song Dynasty, and *Jin Lou Zi* was written in the Southern and Northern Dynasties. The differences of translation effects are likely to be caused by literary styles in different dynasties.

Test data	BLEU	CHRF
Jin Lou Zi	19.9600	19.2349
Houshan Tancong	36.1354	32.1755

Table 16 : CHUK's performance in the Ancient Chinese to Modern Chinese (traditional Chinese format) translation tasks in *Jin Lou Zi* and *Houshan Tancong*

7. Conclusion

EvaHan2023 is the first bakeoff for Ancient Chinese Mechine Translation. The competition provided a large-scale multilingual parallel corpus of Ancient Chinese. The corpus of this

competition with great diachronicity covers pre-Qin classics, *Zhizhi Tongjian*, *Twenty-four Histories*, which were written in different periods and recorded the contents of different periods, providing better support for training high-quality Ancient Chinese translation models.

The participating teams have shown their unique advantages. On the task of translating Ancient Chinese to traditional Modern Chinese, CUHK achieved the best results under the closed modality, and ISTIC achieved the best results under the open modality; on the task of translating Ancient Chinese to simplified Modern Chinese, SCUT achieved the best results under the open modality, and CUHK achieved the best results under the closed modality. On the task of translating Ancient Chinese to English, CHUK and ISTIC achieved the best results under the closed modality and the open modality respectively.

In this competition, the deep trainning language model has been widely used, and the machine translation model from Ancient Chinese to English did not achieve desired results. In future work, we will consider constructing a parallel corpus of Ancient Chinese that includes reference translations in more languages, so as to promote the progress of Ancient Chinese machine translation technology and the spread of excellent traditional Chinese culture to the world.

8. Acknowledgements

This research is supported by the National Social Science Foundation of China major project "Research on the Construction and Application of Cross-language Knowledge Base of Ancient Chinese Classics" (project No. 21&ZD331), Key Project of Ancient Books Work (22GJK006) and National Language Commission Project (YB145-41).

9. Bibliographical References

Adelani, D. I., Alam, M. I., Anastasopoulos, A., Bhagia, A., Costajussà, M. R., Dodge, J., Faisal, F., Fedorova, N., Federmann, C., Guzmán, F., Koshelev, S., Maillard, J., Marivate, V., Mbuya, J., Mourachko, A., Saleem, S., Schwenk, H., & Wenzek, G. (2022, December 7-8). Findings of the WMT'22 shared task on large-scale machine translation evaluation for African languages. In Proceedings of the Seventh Conference on Machine Translation (WMT), pp. 773-800.

Li B., Yuan Y., Lu J., Feng M., Xu C., Qu W., Wang D. (2022). The First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff: Overview of the EvaHan 2022 Evaluation Campaign. In Proceedings of the Second Workshop on Language Technologies for Historical and Ancient Languages, pages 135–140, Marseille, France. European Language Resources Association.

Liu, C., Wang, D. B., Zhao, Z. X., Hu, D., Wu M. C., Lin L. T., Shen S., Li B., Liu J. F., Zhang, H., & Zhao, L. Z. (2023). SikuGPT: A generative pre-trained model for intelligent information processing of ancient texts from the perspective of digital humanities (arXiv:2304.07778). https://doi.org/10.48550/arXiv.2304.07778

ChrF - a Hugging Face Space by evaluate-metric. (n.d.). Retrieved June 29, 2023, from https://huggingface.co/spaces/evaluate-metric/chrf

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding (arXiv:1810.04805). https://doi.org/10.48550/arXiv.1810.04805

Han, F., Yang T. X., & Song J. H.;S (2015). Ancient Chinese MT Based on Sentence-focused Syntax. Journal of Chinese Information Processing, 29 (2), 103-110,117.

Freitag, M., Rei, R., Mathur, N., Lo, C., Stewart, C., Avramidis, E., Kocmi, T., Foster, G., Lavie, A., & Martins, A. F. T. (2022, December 7–8). Results of WMT22 metrics shared task: Stop using BLEU – Neural Metrics Are better and more robust. In Proceedings of the Seventh Conference on Machine Translation (WMT)(pp.46-68).

Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., Gowda, T., Graham, Y., Grundkiewicz, R., Haddow, B., Knowles, R., Koehn, P., Monz, C., Morishita, M., Nagata, M., Nakazawa, T., Novák, M., Popel, M., Popovic, M., & Shmatova, M. (2022, December 7-8). Findings of the 2022 Conference on Machine Translation (WMT22). In Proceedings of the Seventh Conference on Machine Translation (WMT)(pp. 1–45).

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H., & Neubig, G. (2021). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing (arXiv:2107.13586). https://doi.org/10.48550/arXiv.2107.13586

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: A Method for Automatic Evaluation of Machine Translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (pp. 311–318).

Popović, M. (2015, September). chrF: Character n-gram F-score for automatic MT evaluation. In Proceedings of the Tenth Workshop on Statistical Machine Translation(pp.392–395).

Post, M. (2018). A Call for Clarity in Reporting BLEU Scores. In Proceedings of the Third Conference on Machine Translation: Research Papers, 186–191.

Post, M. (2023). SacreBLEU [Python]. https://github.com/mjpost/sacrebleu (Original work published 2017)

Srivastava, V., & Singh, M. (2022, December 7-8). Overview and results of MixMT shared-task at WMT 2022. In Proceedings of the Seventh Conference on Machine Translation (WMT)(pp.806–811).

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., & Lample, G. (2023). *LLaMA: Open and Efficient Foundation Language Models* (arXiv:2302.13971). https://doi.org/10.48550/arXiv.2302.13971

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosu-khin, I. (2017). Attention is All you Need. Advances in Neural Information Processing Systems 30. https://papers.nips.cc/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html

Zhang, J. X., Gan, R. Y., Wang, J. J., Zhang, Y. X., Zhang, L., Yang, P., Gao, X. Y., Wu, Z. W., Dong, X. Q., He, J. Q., Zhuo, J. H., Yang, Q., Huang, Y. F., Li, X. Y., Wu, Y. H., Lu, J. Y., Zhu, X. Y., Chen, W. F., Han T., Pan, K. H., et al. (2022). Fengshenbang 1.0: Being the foundation of Chinese cognitive intelligence. (arXiv:2209.02970). https://doi.org/10.48550/arXiv.2209.02970