EvaHan 2023



Guidelines

Version v1.0

Mar 1, 2023

Organizers list

**Contents**

[1. Introduction 1](#_Toc128604975)

[2. Data 3](#_Toc128604976)

[2.1 Data Format 3](#_Toc128604977)

[2.2 Training Data 4](#_Toc128604978)

[2.3 Test Data 4](#_Toc128604979)

[3. Task 5](#_Toc128604980)

[3.1 Task Objective 5](#_Toc128604981)

[3.2 Task Requirements 5](#_Toc128604982)

[4. Evaluation 7](#_Toc128604983)

[4.1 Metrics 7](#_Toc128604984)

[4.2 Two Modalities 7](#_Toc128604985)

[4.3 Baselines 8](#_Toc128604986)

[5. How to Participate 9](#_Toc128604987)

[5.1 Registration 9](#_Toc128604988)

[5.2 Submitting Runs 9](#_Toc128604989)

[5.3 Writing the Technical Report 10](#_Toc128604990)

[Appendix: Selection of Resources 11](#_Toc128604991)

[Bibliography 12](#_Toc128604992)

# 

# 1. Introduction

The ancient Chinese classics has a long history and culture, serving as an integral part of the world civilization and a treasure that should be shared by people all over the world. The Machine Translation (MT) task of ancient Chinese books is crucial to disseminate the knowledge and wisdom contained in ancient Chinese books to the world and promote the good exchange of Chinese and foreign cultures. However, the existing cross-language corpus of Chinese classical texts has a relatively small amount of data, and the corresponding cross-language machine translation research is relatively insufficient.

Against such backdrop, we utilized a platform which can facilitate the alignment of chapters, paragraphs and sentences to obtain the distribution characteristics of the cross-lingual Chinese classic texts based on parallel chapters, paragraphs and sentences. We achieved the automatic alignment of chapters, paragraphs and sentences for the cross-lingual Chinese classic texts based on the corresponding rules, statistics as well as traditional machine learning and deep learning strategies. Furthermore, the SikuBERT and SikuRoBERTa pre-trained language model oriented to intelligent processing of ancient texts, and the “SIKUBERT intelligent processing platform of classics”, which can provide online services such as automatic processing, retrieval and automatic translation of classics, are developed[1].

We hope to make contributions to the spread of Chinese culture represented by ancient Chinese books.

EvaHan 2023 is the second edition of the campaign totally devoted to the evaluation of Natural Language Processing (NLP) tools for the Ancient Chinese language. The previous edition is EvaHan 2022[2][[1]](#footnote-1).

The campaign is designed with the aim of answering three main questions:

* How can we make machine translation techniques show the best performance on Ancient Chinese?
* How can we promote the development of resources and language technologies for the Ancient Chinese?
* How can we foster collaboration among scholars working on Ancient Chinese and attract researchers from different disciplines?

The task of EvaHan 2023 is machine translation, including **Ancient Chinese to Modern Chinese machine translation** and **Ancient Chinese to English machine translation**, based on the cross-lingual parallel corpus of Chinese classic texts we provide. Shared data and several scorers are provided to the participants. The organizers rely on the honesty of all participants who might have some prior knowledge of part of the data that will be used for evaluation. Unfairly use of such knowledge is not permitted in the shared task.

EvaHan 2023 is co-organized with ALT 2023: “Ancient Language Translation Workshop”, Macau SAR, China on Sep 4, 2023. As a co-located event with MT-SUMMIT2023 (<https://mtsummit2023.scimeeting.cn/en/web/index/>), this workshop will provide an opportunity to learn about the challenges and latest developments in the field of machine translation for ancient languages. EvaHan 2023 is organized by the Computational Linguistics and Digital Humanities (CLDH) Group at Nanjing Normal University, School of Economics & Management, Nanjing University of Science and Technology, School of Information Management, Nanjing Agricultural University in Nanjing, China.

For any update, please check the EvaHan 2023 website:

<https://github.com/GoThereGit/EvaHan>

# 2. Data

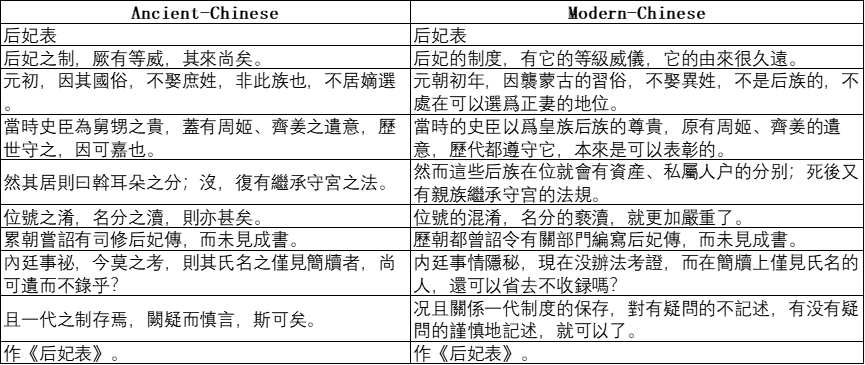
The data comes from China Twenty-four Histories , Pre-Qin canonical texts and “ZiZhi TongJian (资治通鉴, Comprehensive Mirror in Aid of Governance)”. The PDF text is converted into word format through OCR recognition, and the team members manually proofread the corpus by using the convenient platform for chapter, paragraph, and sentence alignment, and finally get the parallel corpus. Among them, China Twenty-four Histories is the general name of the twenty-four official histories written by various dynasties in ancient China, all of which are compiled in the biography style; the Pre-Qin canonical texts are the historical materials of the pre-Qin period, 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 1362 years of history of sixteen dynasties.

The ancient Chinese classic texts in the corpus feature both diachronicity, i.e. spanning thousands of years, as well as diversity, i.e. covering the four traditional types of Chinese canonical texts: *Jing* (经, Confucian classics), *shi* (史, historical works), *zi* (子, philosophical works belonging to schools of thought other than the Confucian but also including works on agriculture, medicine, mathematics, astronomy, divination, art criticism, and other miscellaneous writings) and *ji* (集, collection of literary works).

Both English and modern Chinese translations are selected for these texts in the parallel corpus. The specific parallel texts provided for this test are as follows.

## 2.1 Data Format

The released data is not tokenized and includes sentences of any length (including empty sentences). All data is in Unicode (UTF-8) format. The Figure1. below gives an example of the parallel corpus data format:



**Figure 1.** Data format of the aligned sentences

On the left side is the ancient Chinese text, and on the right side is the modern Chinese text corresponding to the sentence-based unit. For the ancient Chinese-English parallel texts, the same format is followed.

## 2.2 Training Data

The source of the training data is the parallel corpus of Ancient-Chinese-to-Modern-Chinese parallel texts of China Twenty-four Histories and Ancient-Chinese-to-English parallel texts of Pre-Qin canonical texts and “Zizhi Tongjian (资治通鉴, Comprehensive Mirror in Aid of Governance)”.

The overall parallel texts for machine translation are presented as follows.

|  |  |  |
| --- | --- | --- |
| **Data** | **Source language** | **Target language** |
| Ancient-Chinese-to-Modern-Chinese parallel texts of China 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 1.** Detail of training data in EvaHan 2023.

In this task, the cross-lingual parallel corpus of Chinese classic texts is large-scale, diachronic, and well-balanced.

## 2.3 Test Data

Test data will be provided in txt format, including Ancient-Chinese characters, Modern-Chinese characters, English characters and punctuations. The gold standard test data, that is the annotation used for the evaluation, will be provided to the participants after the evaluation.

There are two test data sets, they are designed for Ancient Chinese-Modern Chinese machine translation (testa.txt, TBD) and Ancient Chinese-English machine translation (testb.txt, TBD).

The details of the test data will be provided to the participants after the evaluation.

# 3. Task

The cross-lingual machine translation of Chinese classic texts consists of two parts: **the Ancient-Chinese-to-Modern-Chinese machine translation** and **the Ancient-Chinese-to-English machine translation**. Chinese ancient classics are the important part of traditional Chinese culture. In the field of ancient literature research, the translation of ancient Chinese texts plays a very important role. Ancient Chinese differs greatly from modern Chinese in grammar, syntax, vocabulary, and other aspects. Improving the machine translation performance from Ancient Chinese to Modern Chinese can better promote the study of ancient literature. Improving the machine translation technology from Ancient Chinese to English can also accelerate the promotion of Chinese traditional culture worldwide.

## 3.1 Task Objective

The goals of the translation task are:

* To investigate the applicability of current MT techniques when translating ancient Chinese into English or modern Chinese
* To examine special challenges in translating between ancient Chinese and English or modern Chinese, including word order and syntax
* To create publicly available corpora for machine translation and evaluation of ancient Chinese
* To provide practical experience of the most advanced machine translation methods for beginners in the field of machine translation
* To prompt the development of machine translation research for ancient Chinese and advance the forefront of machine translation technology exploration

## 3.2 Task Requirements

We will provide parallel corpora of Ancient Chinese-Modern Chinese based on the Twenty-Four Histories and Ancient Chinese-English based on pre-Qin texts, respectively, as training and testing data for Ancient Chinese-Modern Chinese and Ancient Chinese-English machine translation. We will also provide several unified models, using Chinese-RoBERTa-wwm-ext[3] for Modern Chinese , Siku-RoBERTa[1] for ancient Chinese and RoBERTa[4] for English. The goal is to improve the model and enhance machine translation performance.

You can choose to participate in one or both of the tasks, and we will use the same metrics for evaluation. For each task, we provide subtasks of two tracks, i.e., closed track and open track. To ensure the fairness of the competition, in the closed track, please use the data we provide as the training data only. However, you can use other models and resources to build the translation system in open track, or just build your own model. If additional data is used, participants should clearly indicate which data is from the provided dataset and which is from external sources. This will allow us to evaluate the performance of the models on our provided dataset separately from their performance on external data.

Each participant should include a brief introduction of their translation system when submitting, including basic information such as the models (if any), techniques, methods used, etc. Each participant should submit a technical reports emphasizing improvements made to the model, techniques used, and methods applied.

Although the primary objective of this evaluation is to identify the most exceptional machine translation system, we encourage participation even if your approach does not achieve the highest performance. If you have developed an interesting approach, this evaluation provides an opportunity to further refine and enhance your system.

# 4. Evaluation

## 4.1 Metrics

We will evaluate the performance of the Ancient-Chinese-to-English machine translation model and Ancient-Chinese-to-Modern-Chinese machine translation model provided by the participants. The scorers employed for EvaHan 2023 are based on BLEU[5], chrF[6] and COMET-QE[7-9].

Each participating team will initially have access only to the training data. Later, test data containing only ancient Chinese texts will also be released. After the assessment, the modern Chinese or English texts corresponding to the ancient Chinese in the test data will also be released.

The BLEU metrics measures machine translation quality by word-level n-grams. It is a modified version of the sacreBLEU[[2]](#footnote-2), which provides hassle-free computation of shareable, comparable, and reproducible BLEU scores. The ChrF metrics evaluates the character-level translation quality and adds a recall metric, thus improving the correlation with human judgment. The COMET-QE is a state-of-the-art metric based on pre-trained models designed to predict human language experts’ judgments of machine translation quality, often with the highest accuracy.

An example of the output of the scorers is given in Table 2. The evaluation will automatically calculate the scores based on the generated outputs and the corresponding reference translations.

|  |  |
| --- | --- |
| **Metric** | **Score** |
| BLEU | 0.47 |
| chrF | 0.6 |
| COMET-QE | 2.6 |

**Table 2.** Example of scorers’ output.

## 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 (Training data name, TBD), and the following pre-trained models listed in Table3. Other resources are not allowed in the closed modality.

|  |  |  |
| --- | --- | --- |
| **Model name** | **Language** | **Description** |
| Siku-RoBERTa[1][[3]](#footnote-3) | Ancient Chinese | Ancient Chinese RoBERTa pre-trained on high-quality “Siku Quanshu (四库全书)” full-text corpus. |
| Chinese-RoBERTa-wwm-ext[3][[4]](#footnote-4) | Modern Chinese | Modern Chinese pre-trained RoBERTa with Whole Word Masking strategy. |
| RoBERTa[5][[5]](#footnote-5) | English | Pre-trained model on English with MLM objective. |

**Table 3.** 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, KGs, 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.

|  |  |  |
| --- | --- | --- |
| **Limits** | **Closed Modality** | **Open Modality** |
| Machine learning algorithm | No limit | No limit |
| Pre-trained model | Only models mentioned in Table 3. | No limit |
| Training data | Only (Training data name, TBD) | No limit |
| Features used | Only from (Training data name, TBD) | No limit |
| Manual correction | Not allowed | Not allowed |

**Table 4.** Limitations on the two modalities.

## 4.3 Baselines

We will evaluate the translated outputs of Google Translate on the test data and use the scores as the baseline.

# 5. How to Participate

## 5.1 Registration

If you would like to participate in this shared task, please fill out the registration form[[6]](#footnote-6) (Links can also be found on the EvaHan 2023 website) and ensure that your information is correct and your email is able to receive messages. Once we receive your registration information, we will send the training data to your email address. Please check your email regularly.

If you have any questions about this shared task, please feel free to send an email to our official email address: evahan2023@gmail.com.

If you do not receive a reply for a long time, please check if your email was sent successfully.

## 5.2 Submitting Runs

Once the system has produced the results for the task over the test set, participants have to follow these instructions for completing your submission:

* File naming:

Name the runs with the following filename format:

taskID\_teamName\_systemID\_modality.tsv

For example: testa\_unicatt\_1\_closed.tsv would be the first run of a team called unicatt using the closed modality for the task using testa.txt(TBD) document (the Ancient-Chinese-to-Modern-Chinese machine translation).

testb\_unicatt\_2\_open.tsv would be the second run of a team called unicatt using the open modality for the task using testb.txt(TBD) document (the Ancient-Chinese-to-English machine translation).

* Submission format:

The output files for system-level rankings should be formatted as a tab-separated values (TSV) in the following way:

<id>\t<source>\t<translation>[\t<translation>]

Each field should be delimited by a single tab character.

Where:

**<id>** is the ID of source data (original ancient Chinese text).

**<source>** is the original ancient Chinese text.

**<** **translation >** is the machine translation result of your system, the second machine translation result is optional.

Below is an example:

|  |
| --- |
| id source translation  1 植，琰之兄女婿也。 Cao Zhi had married a daughter of Cui Yan's elder brother.  2 眾嘉嚴畯能以實讓。 All admired the honest way that Yan Jun had refused the appointment.  3 操曰：“凡人也。” A common fellow, replied Cao Cao.  4 然則何為自往？ Then why go yourself? |

* How to submit:

Before you submit, please run your scores files through a validation script, which we will provide later. You can use it along with either BLEU, chrF or COMET-QE sys.

Submissions should be sent to evahan2023@gmail.com with the subject “EvaHan Submission: taskID - teamName”, where the “taskID” is either testa(TBD) or testb(TBD).

You can make **up to 2 submissions** per language pair, per team.

## 5.3 Writing the Technical Report

Papers should not be longer than **4** pages of content (for references, unlimited number of pages is allowed). The papers must follow the MT Summit 2023 style guides (PDF version, LaTeX version, MS Word version, and Overleaf template(<https://www.overleaf.com/latex/templates/mt-summit-2023-template/knrrcnxhkqxd>) and be submitted in PDF format. To allow for blind reviewing, please do not include author names and affiliations within the paper and avoid obvious self-references.

Papers must be submitted to the following website by the conference submission deadline: <https://softconf.com/mtsummit2023/research>.

# Appendix: Selection of Resources

* Ancient Chinese SikuRoBERTa: https://huggingface.co/SIKU-BERT/sikuroberta; <https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing>
* Modern Chinese RoBERTa: https://huggingface.co/hfl/chinese-roberta-wwm-ext; <https://github.com/ymcui/Chinese-BERT-wwm>
* English RoBERTa: https://huggingface.co/roberta-large; <https://github.com/facebookresearch/fairseq/tree/main/examples/roberta>
* Multilingual version of RoBERTa: https://huggingface.co/xlm-roberta-large; <https://github.com/facebookresearch/fairseq/tree/main/examples/xlmr>
* Ancient Chinese GPT-2: https://huggingface.co/uer/gpt2-chinese-ancient; <https://github.com/Morizeyao/GPT2-Chinese>
* Ancient Chinese SikuGPT: https://huggingface.co/JeffreyLau/SikuGPT2; <https://github.com/SIKU-BERT/sikuGPT>
* GuwenBERT: https://huggingface.co/ethanyt/guwenbert-base; <https://github.com/Ethan-yt/guwenbert>
* Ancient Chinese syntactic corpus: <http://kanji.zinbun.kyoto-u.ac.jp/~yasuoka/kyodokenkyu/2019-03-08/>
* Ancient Chinese Sentence Segmentation: https://seg.shenshen.wiki/; <https://wyd.kvlab.org>
* Tagged Corpus of Old Chinese: <http://lingcorpus.iis.sinica.edu.tw/ancient/>
* A very Large Online Ancient Chinese Corpus Retrieval System: <http://dh.ersjk.com/>
* A GPI Ancient Chinese raw corpus: <https://github.com/garychowcmu/daizhigev20>

# Bibliography

[1] Wang, D., Liu, C, Zhu, Z, et al. 2022. Construction and Application of Pre-trained Models of Siku Quanshu in Orientation to Digital Humanities. *Library Tribune*, 42(6), 31-43. (In Chinese)

[2] 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* (pp. 135-140).

[3] Cui, Y., Che, W., Liu, T., et al. Revisiting pre-trained models for Chinese natural language processing[J]. *arXiv preprint arXiv*:2004.13922, 2020.

[4] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., ... & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

[5] Papineni, K., Roukos, S., Ward, T., & Zhu, W. J. (2002). 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), Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

[6] Popović, M. (2015). chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the tenth workshop on statistical machine translation* (pp. 392-395), Lisbon, Portugal: Association for Computational Linguistics.

[7] Rei, R., Stewart, C., Farinha A. C., & Lavie, A. (2020). COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp.2685-2702), Online: Association for Computational Linguistics.

[8] Rei, R., Farinha, A. C., Zerva, C., van Stigt, D., Stewart, C., Ramos, P., ... & Lavie, A. (2021). Are references really needed? unbabel-IST 2021 submission for the metrics shared task. In *Proceedings of the Sixth Conference on Machine Translation* (pp. 1030-1040), Online: Association for Computational Linguistics.

[9] Kocmi, T., Bawden, R., Bojar, O., Dvorkovich, A., Federmann, C., Fishel, M., ... & Popović, M. (2022). Findings of the 2022 conference on machine translation (WMT22). In *Proceedings of the Seventh Conference on Machine Translation (WMT)* (pp. 1-45), Abu Dhabi, United Arab Emirates: Association for Computational Linguistics.

1. <https://circse.github.io/LT4HALA/2022/EvaHan> [↑](#footnote-ref-1)
2. <https://github.com/mjpost/sacreBLEU> [↑](#footnote-ref-2)
3. <https://huggingface.co/SIKU-BERT/sikuroberta> [↑](#footnote-ref-3)
4. <https://huggingface.co/hfl/chinese-roberta-wwm-ext> [↑](#footnote-ref-4)
5. <https://huggingface.co/roberta-large> [↑](#footnote-ref-5)
6. <https://forms.office.com/Pages/ResponsePage.aspx?id=DQSIkWdsW0yxEjajBLZtrQAAAAAAAAAAAAMAAExHmLlUMURNSUNHQTQ5SUhQMzFIR05GSEo2QUFONi4u&lang=en> [↑](#footnote-ref-6)