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

Chinese ancient classics have a long tradition 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. By Machine Translation (MT), it is significant to disseminate the information and wisdom contained in Chinese ancient books to the world and promote cultural exchanges China and foreign countries. However, there is very little reliable data available in term of the existing cross-language corpus of Chinese classical texts. In contrast, there is an urgent need for comprehensive studies on cross-language machine translation.

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. The automatic alignment of chapters, paragraphs and sentences for the cross-lingual Chinese classic texts has been achieved based on the corresponding rules, statistics as well as traditional machine learning and deep learning strategies. Furthermore, two pre-trained language models oriented to intelligent processing of ancient texts, SikuBERT and SikuRoBERTa, are developed, based on which the “SIKUBERT intelligent processing platform of classics” is built to provide online services such as automatic processing, retrieval and automatic translation of classics [1].It is hoped that this can make a very positive contribution to the extension of Chinese culture into the whole world.

EvaHan 2023 is the second event dedicated entirely to the evalution of Natural Language Processing(NLP) tools in ancient Chinese. The previous edition is EvaHan 2022[2][[1]](#footnote-1).

The event is held to cope with the follow three challenging problems :

* How to make machine translation techniques deliver excellent performances on ancient Chinese?
* How to boost the development of ancient Chinese resources and related language processing techniques?
* How to foster effective teamworking and researchers’ partnership with each other in the field of ancient Chinese studies?

The task of EvaHan 2023 is machine translation, including the translation from **Ancient Chinese to Modern Chinese** and that from **Ancient Chinese to English**, based on the given cross-lingual parallel corpus of Chinese classic texts. Shared data and scorers will be allotted 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.

Co-organized by EvaHan 2023 and ALT 2023, the Symposium On Ancient Chinese Translation will be held in Macau SAR on Sep 4, 2023. Timed to coincide with the MT-SUMMIT2023 (https://mtsummit2023.scimeeting.cn/en/web/index/), this symposium will be a marvellous opportunity to learn about the chanllenges and latest developments in the field of ancient Chinese machine translation. 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.

Please check the EvaHan 2023 website for any update.

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

# 2. Data

The 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)”. Among them, the 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(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 over a span of 1362 years. The PDF version were converted into word format through OCR recognition before the team members could manually proofread the texts and build a parallel corpus by using the convenient platform for chapter, paragraph and sentence alignment.

The Chinese ancient classic texts in the corpus feature both diachronicity(i.e. spanning thousands of years)(i.e. covering the four traditional types of Chinese canonical texts). The four are *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

As to released data, no sentence of any length is tokenized (including empty sentences). All data is in Unicode (UTF-8) format within the parallel corpus. Table 1 gives an example as fllows:

**Table 1.** Example of the parallel corpus

|  |  |
| --- | --- |
| **Ancient-Chinese** | **Modern-Chinese** |
| 后妃表 | 后妃表 |
| 后妃之制，厥有等威，其來尚矣。 | 后妃的制度，有它的等級威儀，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。 |
| 當時史臣為舅甥之貴，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。 |
| 然其居則曰斡耳朵之分；沒，復有繼承守宮之法。 | 然而這些后族在位就會有資産、私屬人户的分别；死後又有親族繼承守宫的法規。 |
| 位號之淆，名分之瀆，則亦甚矣。 | 位號的混淆，名分的褻瀆，就更加嚴重了。 |
| 累朝嘗詔有司修后妃傳，而未見成書。 | 歷朝都曾詔令有關部門編寫后妃傳，而未見成書。 |
| 內廷事祕，今莫之考，則其氏名之僅見簡牘者，尚可遺而不錄乎？ | 内廷事情隱秘，現在没辦法考證，而在簡牘上僅見氏名的人，還可以省去不收録嗎？ |
| 且一代之制存焉，闕疑而慎言，斯可矣。 | 况且關係一代制度的保存，對有疑問的不記述，有没有疑問的謹慎地記述，就可以了。 |
| 作《后妃表》。 | 作《后妃表》。 |

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

## 2.2 Training Data

The source of the training data includes the Ancient-Chinese-to-Modern-Chinese parallel texts of China Twenty-four Histories, the Ancient-Chinese-to-English parallel texts of Pre-Qin classcis and “Zizhi Tongjian”.

Descriptions about the overall parallel texts for machine translation are presented in Table 2.

**Table 2**. Detail of training data in EvaHan 2023.

|  |  |  |
| --- | --- | --- |
| **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 |

Briefly, the training data employed in this task embodies many new features, such as the large-scale and well-balanced data as well as its comprehensiveness from the diachronic perspective.

## 2.3 Test Data

Test data will be provided in txt format, including Ancient-Chinese characters, Modern-Chinese characters, English words and punctuations. The test data, i.e. the annotation used for the evaluation, will be evaluated before it is supplied to the participants.

Two test data sets are built for Ancient Chinese-Modern Chinese machine translation (testa.txt, TBD) and Ancient Chinese-English machine translation (testb.txt, TBD).

More details 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 serve as an indispensable part of Chinese traditional culture. In the field of ancient literature research, the translation of ancient Chinese texts is of great significance. Ancient Chinese differs greatly from modern Chinese in grammar, syntax, vocabulary, and other aspects. To improving the performance of machine translation from Ancient Chinese to Modern Chinese can better deepen studies on ancient literature. Meanwhile, it can also accelerate the promotion of Chinese traditional culture worldwide.

## 3.1 Task Objectives

The goals of the translation task are:

* 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 problems).
* To built free corpora for machine translation and evaluation of ancient Chinese.
* To afford novice researchers the chance to gain experience in the field of machine translation.
* To further machine translation research for ancient Chinese and the exploration of forefront machine translation technology.

## 3.2 Task Requirements

Parallel corpora of Ancient Chinese-Modern Chinese based on the Twenty-Four Histories and Ancient Chinese-English based on Pre-Qin texts are provided as training and testing data for Ancient Chinese-Modern Chinese and Ancient Chinese-English machine translation. Participants are also supplied with unified models. Specifically, Chinese-RoBERTa-wwm-ext[3] for Modern Chinese , Siku-RoBERTa[1] for ancient Chinese and RoBERTa[4] for English. The participants’ task is to improve the models in terms of machine translation performance.

You can choose either or both of them to participate in, with the same metrics for evaluation is employed. Each task, contains two tracks, i.e., closed track and open track. To ensure the fairness of the competition, only the given training data and model are allowed to use in the closed track. However, when building the translation system in open track other models, resources and self-built modes are permitted. In this case, if any additional data or model is employed, participants should make it clear to us so that we can evaluate the performance of the models on our provided dataset separately from their performance on external data.

The results submitted by each team should include a brief introduction to the translation system (i.e. basic information, such as the models (if any), techniques, methods used, etc.). Besides, teams are required to submit a technical report detailing the improvements to the model and the methods used.

Although the main goal of this evaluation is to identify the best performing machine translation project, we encourage creative projects to enter the competition even if their performance is not optimal. Participants can also use this evaluation to further improve their project.

# 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 3. The evaluation will automatically calculate the scores based on the generated outputs and the corresponding reference translations.

**Table 3.** Example of scorers' output.

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

## 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 Table 4. Other resources are not allowed in the closed modality.

**Table 4.** Pre-trained models for 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. |

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.

**Table 5.** Limitations on the two modalities.

|  |  |  |
| --- | --- | --- |
| **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 |

## 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 for 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)