EvaHan 2023

Guidelines

Version v1.0

Mar 1, 2023

School of Literature, Nanjing Normal University

School of Economics & Management, Nanjing University of Science and Technology

College of Information Management, Nanjing Agricultural University

Contents

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. There are a large number of ancient books written in Chinese characters, and it is time-consuming and laborious to read and translate only by human beings. 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 ancient Chinese texts. In contrast, there is an urgent need for comprehensive studies on cross-language machine translation.

EvaHan is a series of international evaluation focusing on the information processing of ancient Chinese. In 2022, together with EvaLatin for automatic analysis and evaluation of Ancient Latin, Evahan 2022 was held in Marseille, France. More than ten teams participated in the evaluation, and refresh the best results in the industry. In order to promote the progress of ancient book processing technology, the second EvaHan Evaluation will be held in Macao, China in 2023.

EvaHan 2023 is the second International Evaluation of Ancient Chinese Information Processing. This year's task is machine translation of ancient Chinese, which includes two sub-tasks:

- **Ancient Chinese to Modern Chinese**
- **Ancient Chinese to English**

The task was carried out based on the cross-language parallel corpus of ancient texts provided by the sponsor. Training data, test data, and evaluation scripts will be provided.

The evaluation 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?

Co-organized by EvaHan 2023 and ALT 2023, the Symposium On Ancient Chinese Translation will be held in Macau SAR on Sep 4, 2023 at MT-SUMMIT2023 (https://mtsummit2023.scimeeting.cn/en/web/index/), one of the top international conferences on machine translation.

EvaHan 2023 is organized by Computational Linguistics and Digital Humanities (CLDH) Group at School of Literature, Nanjing Normal University, College of Information Management, Nanjing Agricultural University, School of Economics & Management, Nanjing University of Science and Technology.

Please check the EvaHan 2023 website for any update.
https://github.com/GoThereGit/EvaHan

2. Data

Training data for evaluation 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)”,which was finished by the research group of 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) . Among them, the Twenty-Four Histories is the general name of the twenty-four official histories written by 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 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), *shi* (史), *zi* (子) and *ji* (集).

2.1 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 1 and Table 2.

Table 1. Example of the Ancient Chinese to Modern Chinese 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》。

Table 1 shows an example of the Ancient Chinese to Modern Chinese parallel corpus. On the left side is the Ancient Chinese text, and on the right side is the Modern Chinese (traditional Chinese) text corresponding to the sentence unit.

Table 2. Example of the Ancient Chinese to English corpus

Ancient-Chinese	English
杜密素與李膺名行相次，	Du Mi had shared in reputation with Li Ying,
起，對之揖，勸令從學。	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.
獵者來還，莫不潤涕。	When the men came back from the hunt, every one of them wept for sorrow.
榮恐不免，詣闕自論。	Kou Rong was afraid he would not escape [this combination of hostility] and he [sought to] go back to the palace to plead his case.
悲夫，久生亦復何聊！	Alas! One may live so long, but what is there to hope for?
三公統外，御史察內。	The Three Excellencies deal with matters outside, the imperial clerks investigate matters within.
帝不得已，竟免覽官。	And the emperor in the end had no recourse but to dismiss Hou Lan.
三月，辛巳，赦天下。	In the third month on the day xinsi [14 Apr] there was an amnesty for the empire.

Table 2 shows an example of the Ancient Chinese to English parallel corpus. On the left side is the Ancient Chinese text, and on the right side is the English text corresponding to the sentence unit.

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 3.

Table 3. 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

The test data only provides Ancient Chinese texts(the source language),and one test data set is provided for Ancient Chinese to Modern Chinese machine translation and Ancient Chinese to English machine translation , about 2000 sentences each.

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**.

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).
- Provide a platform for the enthusiasts of machine translation in ancient Chinese
- To further machine translation research for ancient Chinese and the exploration of forefront machine translation technology.

3.2 Task Requirements

The difference between Ancient Chinese-Modern Chinese translation task and Ancient Chinese-English translation task is that the source corpus is different and the model specified in the closed mode is different.

You can choose either or both of them to participate in, with the same metrics for evaluation is employed. Although the Chinese data in the training corpus is all traditional, you can also choose to submit simplified translation results when submitting the Ancient Chinese-Modern Chinese translation results, and the test data in the evaluation includes both versions.

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[1], chrF[2] and COMET-QE[3-5].

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[1], 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 2. Example of scorers' output.

Metric	Score
BLEU	45
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.

[1] https://github.com/mjpost/sacreBLEU

Table 3. Pre-trained models for closed modality.

Model name	Language	Description
Siku-RoBERTa[6][2]	Ancient Chinese	Ancient Chinese RoBERTa pre-trained on high-quality "Siku Quanshu (四库全书)" full-text corpus.
Chinese-RoBERTa-wwm-ext[7][3]	Modern Chinese	Modern Chinese pre-trained RoBERTa with Whole Word Masking strategy.
RoBERTa[8][4]	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 4. 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.

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

5. How to Participate

5.1 Registration

If you would like to participate in this shared task, please fill out the registration form[5] (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

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. 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:

  tasknumber_teamName_systemnumber_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>]

[5] https://forms.office.com/Pages/ResponsePage.aspx?id=DQSIkWdsW0yxEjajBLZtrQAAAAAAAAAAAAMAAExHmLlUMURNSUNHQTQ5SUhQMzFIR05GSEo2QUFONi4u&lang=en

Each field should be delimited by a single tab character.

Where:

<id> is the number 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.

A sample of Ancient Chinese-Modern Chinese translation result submission is given in Figure 1.

id	source	translation
1	孝元皇后，王莽之姑也。	孝元皇后，是王莽的姑媽。
2	後八十年，當有貴女興天下云。	八十年後，王家會有貴女興於天下。
3	母，適妻，魏郡李氏女也。	他們的母親是正妻，魏郡李氏的女兒。
4	後以妒去，更嫁爲河内苟賓妻。	後來因爲妒嫉被休，又改嫁爲河内苟賓的妻子。

Figure 1 A sample of Ancient Chinese-Modern Chinese translation result submission

A sample of Ancient Chinese-English translation result submission is given in Figure 2.

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?

Figure 2 A sample of Ancient Chinese-English translation result submission

- 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 num ber 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 an d 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] 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.

[2] 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.

[3] 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.

[4] 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.

[5] 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.

[6] 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)

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

[8] 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.