MTS Machine Translation Summit 2023

September 4-8, 2023 Macau SAR, China

Proceedings of ALT2023:
First Workshop on Ancient Language Translation

September 5, 2023

Editors: Bin Li, Shai Gordin

Preface

The proceedings include the papers accepted for presentation at the First Workshop on Ancient Language Translation (Machine Translation from Ancient Languages to Modern Languages, ALT2023 for short)[1]. The workshop was held on September 5th in Macau SAR, China, co-located with the 19th Machine Translation Summit (MT Summit 2023)[2].

The workshop seeks to provide an opportunity to learn about the challenges and latest developments in the field of machine translation for ancient languages. Participants engaged in discussions and hands-on activities to develop a deeper understanding of the field and the techniques used to address the unique challenges posed by translating texts written in ancient languages. The workshop concluded with a discussion of the results of the hands-on activities and a summary of the key takeaways from the workshop. Participants left the workshop with a deeper understanding of the field of ancient language machine translation and the tools and techniques used to address its unique challenges. In this year's workshop, we proposed shared tasks on Machine Translation for Ancient Chinese and Cuneiform languages (Akkadian and Sumerian), respectively, to provide an opportunity to address the unique challenges faced by ancient language machine translation. The topics of the workshop were closely related to the special features of translation in ancient languages that distinguish them from modern languages and have a significant impact on machine translation.

ALT 2023 is the venue for the second edition of EvaHan, an event dedicated to the evaluation of NLP tools for Ancient Chinese. EvaHan [3] is a series of international evaluations focusing on the information processing of Ancient Chinese. In 2022, together with EvaLatin for automatic analysis and evaluation of Ancient Latin, EvaHan 2022 focused on the task of Part-of-Speech tagging. More than ten teams participated in the evaluation, and Evahan2022 achieved the best results ever in the field.

EvaHan2023 focused on Machine Translation from Ancient Chinese to Modern Chinese/English. EvaHan2023 was organized by the Center of Language Big Data and Computational Humanities at Nanjing Normal University, College of Information Management at Nanjing Agricultural University, School of Economics & Management at Nanjing University of Science and Technology.

Training data for evaluation was 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 test data was only provided in Ancient Chinese, which was derived from the ancient Chinese books Jinlouzi (金楼子) and Houshan Tanshong (后山谈丛). The test dataset consisted of about 2,000 sentences. Each participant could submit runs following two modalities. In the closed modality, the resources each team could use are limited. Each team could only use the training data, and the pre-trained models supplied by the organizers. In the open modality, however, there was no limit on the resources, data and models. Participants were required to submit a technical report for the task in which they participated. EvaHan received a total of eight technical reports, all of which were briefly reviewed by the organizers to check for correct formatting, accuracy of reported results and rankings, and overall presentation. There is also an overview paper in the proceedings detailing some specific aspects of the second EvaHan, such as the datasets, metrics, and results of the shared task.

Besides EvaHan, ALT 2023 hosted also the first edition of EvaCun[4], an evaluation series of NLP tools for the Ancient languages written in the Cuneiform script (3,400 BCE-75CE), organized by Adam Anderson (Data Science Discovery Partner, UC Berkeley, California), Shai Gordin (Digital Pasts Lab,

[1] https://github.com/GoThereGit/ALT
[2] https://mtsummit2023.scimeeting.cn/en/web/index/
[3] https://github.com/GoThereGit/EvaHan
[4] https://digitalpasts.github.io/EvaCUN/

Ariel University, Israel), and their research students. Cuneiform is one of the earliest writing systems in recorded human history (ca. 3,400 BCE-75 CE). Hundreds of thousands of such texts were found over the last two centuries in the Middle East. Most of these texts are found on a clay or stone medium, and are written in Sumerian and Akkadian, beside relatively smaller corpora (still in the tens of thousands) in Elamite, Eblaite, Hittite, Hurrian, Urartian, Hattian, and Luwian, as well as languages which use alphabetic Cuneiform like Ugaritic and Old Persian. EvaCun 2023 consists of three machine translation tasks – Akkadian (in Cuneiform) to English, Akkadian (transcription) to English and Sumerian (transcription) to English, based on the corpora of royal, administrative, and financial texts we provide. For the Akkadian part we used the corpora from the Open Richly Annotated Cuneiform Corpus (ORACC)[5]. Chronologically, the great majority of the texts are Neo-Assyrian (NA) and the best attested genres are the royal inscriptions (2,997) and administrative letters (2,003). Nevertheless, the chosen corpus represents a variety of genres. For the transcription to English we used 56,160 sentences, where we treat each sentence as an independent example for training. We call them in these guidelines "sentences", even if they are made up of a single word, a group of words, a phrase or a group of phrases. This is mostly because Cuneiform does not have punctuations that separate sentences like modern languages do. For the Sumerian part we used a corpus from the Cuneiform Digital Library Initiative (CDLI) and of a neural network-based encode-decoder architecture for English-Sumerian and Sumerian-English. The Sumerian data is only available in transliterated form. The project carries out English to Sumerian and Sumerian to English Translation using a parallel corpus of about 20K sentences for both languages as the parallel corpora. We evaluated the performance of the cuneiform/transcription/Sumerian-to-English machine translation model based on BLEU. EvaCun received one technical report overdue. The task will move to the next year.

[5] http://oracc.museum.upenn.edu/

Organizers:

Shai Gordin (shaigo@ariel.ac.il), Ariel University, Israel
Bin Li (lib@njnu.edu.cn), Nanjing Normal University, China

Program Committee:

Adam Anderson, US Berkeley (USA)
Congjun Long, Chinese Academy of Social Sciences (China)
Dongbo Wang, Nanjing Agricultural University (China)
Ethan Fetaya, Bar-Ilan University (Israel)
Gabriel Stanovsky, Hebrew University of Jerusalem (Israel)
Konstantin Margulyan, Ariel University (Israel)
Liu Liu, Nanjing Agricultural University (China)
Luis Sáenz, Ariel University/Heidelberg University (Israel/Germany)
Minxuan Feng, Nanjing Normal University, (China)
Morris Alper, Tel Aviv University (Israel)
Renfen Hu, Beijing Normal University (China)
Sanhong Deng, Nanjing University, (China)
Si Shen, Nanjing University of Science and Technology (China)
Stav Klein, Ariel University (Israel)
Xiaodong Shi, Xiamen University (China)
Yudong Liu, Western Washington University (USA)

EvaCUN 2023 Organizers:

Adam Anderson, University of California, Berkeley (USA)
Shai Gordin, Ariel University (Israel)
Stav Klein, Ariel University (Israel)
Konstantin Margulyan, Ariel University (Israel)

EvaHan 2023 Organizers:

Dongbo Wang, Nanjing Agricultural University (China)
Si Shen, Nanjing University of Science and Technology (China)
Minxuan Feng, Nanjing Normal University (China)
Chao Xu, Nanjing Normal University (China)
Lianzhen Zhao, China Pharmaceutical University (China)
Wenlong Sun, Nanjing Tech University (China)
Kai Meng, Nanjing Agricultural University (China)
Liu Liu, Nanjing Agricultural University (China)
Wenhao Ye, Nanjing Agricultural University (China)
Weiguang Qu, Nanjing Normal University (China)
Bin Li, Nanjing Normal University (China)

Sponsors:

Phoenix Media

Jiangsu Wenku

Table of Contents

Conference Program

Monday, September 5, 2023

14:00–14:10 Opening Remarks

Invited Talks

14:10–14:30 Prof. Zhiwei Feng, Xinjiang University (China)

14:30–15:00 Prof. Jinxing Yu, Peking University (China)

Oral Reports

15:00–15:15 *EvaCun: The first shared task on Cuneiform Machine Translation*
Shai Gordin

15:15–15:30 *EvaHan2023: Overview of the First International Ancient Chinese Translation Bakeoff*
Dongbo Wang, Litao Lin, Zhixiao Zhao, Wenhao Ye, Kai Meng, Wenlong Sun, Lianzhen Zhao, Xue Zhao, Si Shen, Wei Zhang and Bin Li

15:30–16:00 ***Coffee Break***

16:00–16:15 *The Ups and Downs of Training RoBERTa-based models on Smaller Datasets for Translation Tasks from Classical Chinese into Modern Standard Mandarin and Modern English*
Stuart Michael McManus, Roslin Liu, Yuji Li, Leo Tam, Stephanie Qiu and Letian Yu

16:15–16:30 *Pre-trained Model In Ancient-Chinese-to-Modern-Chinese Machine Translation*
Jiahui Wang, Xuqin Zhang, Jiahuan Li and Shujian Huang

16:30–16:45 *Some Trials on Ancient Modern Chinese Translation*
Li Lin and Xinyu Hu

16:45–17:00 *Istic Neural Machine Translation System for EvaHan 2023*
Ningyuan Deng, Shuao Guo and Yanqing He

Monday, September 5, 2023 (continued)

17:00–17:15 *BIT-ACT: An Ancient Chinese Translation System Using Data Augmentation*
Li Zeng, Yanzhi Tian, Yingyu Shan and Yuhang Guo

17:15–17:30 *Technical Report on Ancient Chinese Machine Translation Based on mRASP Model*
Wenjing Liu and Jing Xie

17:30–17:45 *AnchiLm: An Effective Classical-to-Modern Chinese Translation Model Leveraging bpe-drop and SikuRoBERTa*
Jiahui Zhu and Sizhou Chen

17:45–18:00 *Translating Ancient Chinese to Modern Chinese at Scale: A Large Language Model-based Approach*
Jiahuan Cao, Dezhi Peng, Yongxin Shi, Zongyuan Jiang and Lianwen Jin

18:00–18:10 Closing Remarks