EvaHan 2022

Guidelines

Version v1.0

Oct 18, 2021

Bin Li[1] Yiguo Yuan[1] Minxuan Feng[1] Chao Xu[1] Dongbo Wang[2]

1. School of Chinese Language and Literature, Nanjing Normal University
2. College of Information Management, Nanjing Agricultural University
E-mail: libin.njnu.at.gmail.com

Contents

Chapter 1

Introduction

EvaHan 2022 is the first campaign totally devoted to the evaluation of Natural Language Processing (NLP) tools for the Ancient Chinese language. The Ancient Chinese language dates back around 1000BC-221BC. The campaign is designed following a long tradition in NLP, see for example other campaigns such as MUC, SemEval, CoNLL, EVALITA and SIGHAN[1], with the aim of answering two main questions:

- How can we promote the development of resources and language technologies for the Ancient Chinese language?
- How can we foster collaboration among scholars working on Ancient Chinese and attract researchers from different disciplines?

EvaHan first edition has one task (i.e. a joint task of **Word Segmentation** and **POS Tagging**). Shared data and a scorer 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 is organized within the "Workshop of Language Technologies for Historical and Ancient Languages", co-located at LREC 2022[1]. EvaHan is organized by the Computational Linguistics and Digital Humanities (CLDH) Group at Nanjing Normal University in Nanjing, China.
For any update, please check the LT4HALA website:
https://circse.github.io/LT4HALA/2022/organization .

[1] https://lrec2022.lrec-conf.org/en/

Chapter 2

Data

The dataset of EvaHan 2022 is made of texts from the Classic Texts like *Zuozhuan* and other historical texts[2]. The training and gold texts have been automatically punctuated, word segmented and POS tagged, and then manually corrected by Ancient Chinese language experts.

2.1 Data Format

Training data is distributed following the word segmentation and POS tagging guidelines for Ancient Chinese by Nanjing Normal University[3]. According to such format, annotations are encoded in UTF-8 plain text files. There are no word boundaries in Chinese texts. Thus, the raw texts contain characters and punctuation. After manual annotation, word boundaries and POS tags are added to the text. As shown in Table 1, each word is labelled with a POS tag, in the form of **Word/POS**. And each word is separated by a space. Punctuations are treated as words too.

Type	Example
Raw Text with Punctuations	亟請於武公，公弗許。
Annotated Text with word boundaries and POS tags	亟/d 請/v 於/p 武公/nr ，/w 公/n 弗/d 許/v 。/w

Table 1: Examples of the data format.

2.2 Training Data

The training data contains punctuated, word-segmented and part-of-speech tagged text from *Zuozhuan (左传)*, an ancient Chinese work believed to date from the Warring States Period (475-221 BC). *Zuozhuan* is a commentary on the *Chunqui (春秋)*, a history of the Chinese Spring and Autumn period (770-476 BC).

The files are presented in UTF-8 plain text files using traditional Chinese script. It is released via Linguistic Data Consortium (LDC)[1].

[1] https://catalog.ldc.upenn.edu/LDC2017T14

Data Sets	Data name	Sources	Word Tokens	Char Tokens
Train	*Zuozhuan_Train*	*Zuozhuan*	166,142	194,995
Test A	*Zuozhuan_Test*	*Zuozhuan*	28,131	33,298
Blind Test B	*Blind_Test*	Other similar ancient Chinese Book	Around 40,000	Around 50,000

Table 2: Texts distributed as training/test data in EvaHan 2022.

2.3 Test Data

Test data will be provided in raw format, only Chinese 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. Test A is designed to see how a system perform on the data from the same book. *Zuozhuan_Test* is extracted from *Zuozhuan*, not overlapping with *Zuozhuan_Train*. *Zuozhuan_Test* has been released by LDC. But the teams are not allowed to use it as training data. There have been several papers reporting their performance on this data[4-7].

Blind Test B is designed to see how a system performs on similar data (texts of similar content but from different books). *Blind_Test* has not been released publicly. Its size is similar to that of *Zuozhuan_Test.*

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

Chapter 3

Task

3.1 Task

This Section provides details on the task included in EvaHan2022.

3.1.1 Word Segmentation and POS tagging

Word segmentation is the process of transforming Chinese character sequence to word sequence. Each word is separated by one space. And POS tagging is the process of labelling the Part-of-Speech sequence to word sequence. The example of shown in Table 3 below. Word segmentation and POS tagging have been treated as a joint task in many Chinese language processing systems. Therefore, in this shared task, a sentence should be automatically parsed from raw text to POS tagged text. And the evaluation toolkit will give the scores on both word segmentation and POS tagging.

Raw Text with Punctuations	亟請於武公，公弗許。
Annotated Text with word boundaries	亟 請 於 武公 ， 公 弗 許 。
Annotated Text with word boundaries and POS tags	亟/d 請/v 於/p 武公/nr ，/w 公/n 弗/d 許/v 。/w

Table 3: Examples of Word Segmentation and POS Tagging.

Please note that EvaHan2022 does not accept running results with word segmentation only.

3.1.2 POS Tagging Set

In the Part-of-Speech (POS) Tagging task, systems are required to assign a lexical category (POS tag) to each token as shown in Table 1.

The POS tagging set, which has 22 tags, is shown in Table 4. The users could read ref [3] for further information.

ID	Tag	Part-of-speech	Example(Chinese_English Trans)
1	a	adjective	大_big
2	c	conjunction	則_then
3	d	adverb	不_not
4	f	locative	前_front
5	j	combined	焉_at there
6	m	number	一_one
7	n	noun	人_human
8	nr	person	孔子_ Confucius
9	ns	location	齊_Qi(state name)
10	p	prepositional	於_at
11	q	classifier	匹_classifier for horse and wolf
12	r	pronoun	吾_me
13	s	onomatopoeia	譆譆_LOL
14	t	time	五月_the fifth month
15	u	aux	之_of
16	v	verb	如_go
17	vs	causative usage of predicate	驚_make someone **shock**ed
18	vw	benefitive usage of predicate	泣_**cry** for
19	vy	conative usage of predicate	寶_take something as **treasure**
20	y	modal	乎_interrogative
21	w	punctuation	。_ full stop
22	x	others	abc

Table 4. The 22 part-of -speech tags

•a: adjective. They modify nouns and specify their properties or attributes. They usually describe a person or thing. Examples, 大(*big),* 高(*tall).*

•c: conjunction. Conjunction is a word that joins words, phrases or sentences. Examples*,* 與(*and),* 若(if*).*

•d: adverb. Adverbs modify other words in the sentence, especially verbs, providing information about manner, degree, cause, place, or time. Examples, 弗(*no),* 初(*at the beginning)*.

•f: locative. the form of a noun, pronoun or adjective when it expresses the idea of place. Examples, 西(*east),*下(*down).*

•j: compatible. This tag refers to a monosyllabic word combined by two monosyllabic words, thus having two meanings like a phrase. Examples, 諸(*it at),* 焉*(at there).*

•m: number. It is a sign or symbol that represents a number. Examples, 一(*one),* 十(*ten)*，.

•n: noun. This tag is used for common nouns typically denoting a person, place, thing, animal or idea. Examples, 人*(human)*, 天*(heaven)*.

•nr: person. This tag refers to the name of the person in the text. Examples, 孔子(*Confucius),* 楚王*(Lord of Chu State)*.

•ns: location. This tag refers to the place name in the text. Examples, 齊(*Qi State),* 秦*(Qin State)*.

•p: preposition. Preposition is a word which usually has a noun group as its object. Examples, 為*(for),* 以*(by).*

•q: classifier. Classifier is a word or morpheme used in some languages in certain contexts (such as measure) to indicate the semantic class to which the counted item belongs. Example, "匹"*classifier for horse and wolf*.

•r: pronoun. This tag refers to a word that is used instead of a noun or noun phrase. Examples, 之*(it)*, 是*(this thing/situation).*

•s: onomatopoeia. Onomatopoeia refers to the use of words which describe the sounds. Example, 譆譆*(sounds like sisi)*.

•t: time. This tag refers to a word describing a day, a month or a year. Examples, 正月(*the first month*）*,* 今(*today),* 明年(*next year)*.

•u: auxiliary. In grammar, an auxiliary or auxiliary verb is a verb which is used with a main verb. Examples, 之(*of)*.

- v: verb. Verbs convey actions, occurrences, or states of being. Examples, 如(*go),* 伐*(invade)*.

- vs, vw, vy. The three special usages of predicates (verbs and adjectives). Examples, 泣*(cry)* 之*(it)* means “cry for it”.

- y: model. Model refers to modal particles that express emotions such as exclamation and surprise. Examples, 乎(*interrogative).*

- w: punctuation. Examples, 。(full stop), ？ (question mark).

- x: others. Not appeared in the train/test texts.

The data is provided in the UTF-8 encoding. All files were automatically verified and manually checked.

Chapter 4

Evaluation

4.1 Metrics

Each participating team will initially have access only to the training data. Later, the unlabeled test data will also be released. After the assessment, the labels for the test data will also be released. The scorer employed for EvaHan is a modified version of the one developed for the ref [5]. An example of the output of the scorer is given in Table 5. The evaluation will align the system-produced words to the gold standard ones. Then, Word Segmentation (WS) and Part-of-Speech (POS) are evaluated: precision, recall and F1 score are calculated. The final ranking will be based on F1 score.

Metric	Precision	Recall	F1 Score
WS	95.00	92.00	93.48
POS	90.00	91.00	90.50

Table 5: Example of scorer 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 *Zuozhuan_Train*, and the pretrained model *SIKU-Roberta*[7][1]. It is word embeddings pretrained on a very large corpus of traditional Chinese collection, *Siku Quanshu* (四库全书)[2]. Other resources are not allowed in the closed modality.

In the open modality, there is no limit on the resources, data and models. Annotated external data, such as the components or Pinyin of the Chinese characters, word embeddings can be employed. But each team has to state all the resources, data and models they use in each system in the final report.

[1] https://huggingface.co/SIKU-BERT/sikuroberta

[2] https://en.wikipedia.org/wiki/Siku_Quanshu

Limits	Closed Modality	Open Modality
Machine learning algorithm	No limit	No limit
Pretrained model	Only *SIKU_Roberta*	No limit
Training data	Only *Zuozhuan_Train*	No limit
Features used	Only from *Zuozhuan_Train*	No limit
Manual correction	Not allowed	Not allowed

Table 6: Limitations on the two modalities.

4.3 Baselines

As a baseline, we provide the scores obtained on *Zuozhuan_test* using CRFs (Conditional Random Fields) Training on *Zuozuan_train* without additional resources [4].

Data Set	Data Name	Word Segmentation			POS Tagging		
		Precision	Recall	F1 Score	Precision	Recall	F1 Score
Close test A	*Zuozhuan_test*	90.64	92.08	91.35	89.06	89.54	89.30

Table 7: Baseline for the WS and POS tagging task on Close test A in the closed modality.

Chapter 5

How to Participate

Participants will be required to submit their runs and to provide a technical report for the task they participated in.

5.1 Submitting Runs

Each participant can submit runs for each subtask within each task. A run should be produced according to the 'closed modality'. The second run will be produced according to the 'open modality'. The closed run is compulsory, while the open run is optional.
Once the system has produced the results for the task over the test set, participants have to follow these instructions for completing your submission:

- Name the runs with the following filename format:
  testID_teamName_systemID_modality.txt

  For example: *testa_unicatt_1_closed.txt* would be the first run of a team called *unicatt* using the closed modality for the task using *testa.txt* document.

  testb_unicatt_2_open.txt would be the second run of a team called *unicatt* using the open modality for the task using the blind *testb.txt* document.

- Send the file to the following email address: libin.njnu[AT]gmail.com, using the subject "EvaHan Submission: task - teamName", where the "task" is either *testa* or *testb*.

- Each team could only submit up to 2 running files for each test file in each modality. Thus, each team could submit up to 8 running files in all.

5.2 Writing the Technical Report

Technical reports will be included in the proceedings of the Workshop on Language Technologies for Historical and Ancient Languages 2022 (LT4HALA 2022) as short papers and will be published along the LREC 2022 proceedings. Reports must be submitted through the START platform (URL available soon). All the reports must meet the following requirements:

- they must be written in English;

- they must be formatted according to the LREC 2022 conference style[1];
- the maximum length is 4 pages (excluding references);
- they should contain (at least) the following sections: description of the system, results, discussion, references.

Reports will receive a light review: we will check for the correctness of the format, the exactness of results and ranking, and overall exposition. If needed we will contact the authors asking for corrections.

[1] https://lrec2022.lrec-conf.org/en/submission2022/authors-kit/

Appendix A

POS Tags Modified in PreQin Corpus

A.1 Modify the POS tags

- sv→vs
- yv→vy
- wv→vw

Appendix B

Selection of Resources for Ancient Chinese

- SikuRoBERTa: https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing; https://huggingface.co/SIKU-BERT/sikuroberta
- Ancient Chinese GPT-2: https://huggingface.co/uer/gpt2-chinese-ancient; https://github.com/Morizeyao/GPT2-Chinese
- 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] Sproat, Richard, and Thomas Emerson. The first international Chinese word segmentation bakeoff. *Proceedings of the second SIGHAN workshop on Chinese language processing.* 2003, pages 133-143. (Note: This shared task is only for Modern Chinese).

[2] Bin Li, Minxuan Feng, Xiaohe Chen. Corpus Based Lexical Statistics of Pre-Qin Chinese. *Lecture Notes in Computer Science*, Volume 7717, pages 145-153, 2013.

[3] Xiaohe Chen, Minxuan Feng, Runhua Xu and et al. *Pre-Qin literature information processing*. Beijing: World Publishing Corporation, 2013.

[4] Min Shi, Bin Li and Xiaohe Chen. CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese. *Journal of Chinese Information Processing*, 34(04):1-9, 2010.

[5] CHENG Ning, LI Bin, XIAO Liming, XU Changwei, GE Sijia, HAO Xingyue, FENG Minxuan. Integration of Automatic Sentence Segmentation and Lexical Analysis of Ancient Chinese based on BiLSTM-CRF Mode. *1st Workshop on Language Technologies for Historical and Ancient Languages, (LT4HALA 2020)*, pp 52-58. Marseille, 11–16 May 2020.

[6] Jingsong Yu, Yi Wei, Yongwei Zhang, Hao Yang. Word Segmentation for Ancient Chinese Texts Based on Nonparametric Bayesian Models and Deep Learning. *Journal of Chinese Information Processing*, 34(06):1-8, 2020.

[7] Dongbo Wang, Chang Liu, Zihe Zhu, Jiangfeng Liu, Haotian Hu, Si Shen, Bin Li. Construction and Application of Pre-training Model of “Siku Quanshu” Oriented to Digital Humanities, *Library Tribune*, 1-14[2021-08-20], http://kns.cnki.net/kcms/detail/44.1306.G2.20210819.2052.008.html.

[8] Renfen Hu, Shen Li, Yuchen Zhu. Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on Deep Language Models. *Journal of Chinese Information Processing*, 35(04):8-15, 2021.