Ancient Chinese Corpus (ACC) V1.0

Author(s): Xiaohe Chen, Bin Li, Minxuan Feng, Chao Xu, Runhua Xu, Min Shi, Lili Yu, Lei Xiao, Qingqing Wang

Introduction

The Ancient Chinese Corpus (ACC) V1.0, contains the word segmented, POS-tagged data of *Zuozhuan* (an ancient Chinese history classical book). It has 180,000 Chinese characters, 195,000 segment units (including words and punctuations). It is separated to 2 parts, training data (166,138 words) and test data (28,131 words). The POS tagging set has 17 tags. The details of the tagging set are shown in table 1.

The Ancient Chinese Corpus project began at the Nanjing Normal University in 2009. The project goal is to provide a large, part-of-speech tagged Ancient Chinese corpus. In this first delivery, ACC 1.0, contained only one book *Zuozhuan*. We will continue to release much more data.

Data

There are two text files in this release, containing 268 paragraphs, 10,560 lines. Each line is one sentence or a statement of a person. Each paragraph is separated by one empty line. Each word is tagged its part-of-speech and separated by a space.

Example: 夏/n 四月/t ，/w 費伯/nr 帥/v 師/n 城/v 郎/ns 。/w

We designed the POS tagging set, which has 17 tags shown in table 1. The users could refer the following paper or Chinese book for further information.

Table 1. The 17 part-of -speech tags

ID	Tag	Part-of-speech	Example(Chinese_English Trans)
1	a	adjective	大_big
2	c	conjunction	则_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	y	modal	乎_interrogative

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

- Xiaohe Chen, Minxuan Feng, Runhua Xu, et al. Information Processing of Pre-Qin Chinese. World Publishing Corporation, Beijing, 2013. (陈小荷,冯敏萱,徐润华,等.先秦文献信息处理, 世界图书出版公司, 2013)
- Bin Li, Minxuan Feng, Xiaohe Chen. Corpus Based Lexical Statistics of Pre-Qin Chinese. Lecture Notes in Computer Science Volume 7717, 2013, pp 145-153.

Samples

Please view the following sample file:

example.txt

Acknowledgement

This work was supported in part by the Ministry of Education of China (16YJC740034) National Social Science Foundation of China (15ZDB127).

Updates

We will continue to release more annotated data of Ancient Chinese.

Copyright