

文章编号: 1003-0077(2023)03-0046-08

第一届古代汉语分词和词性标注国际评测

李 斌¹, 袁义国¹, 芦靖雅¹, 冯敏萱¹, 许 超¹, 曲维光², 王东波³

(1. 南京师范大学 文学院, 江苏 南京 210097;

2. 南京师范大学 计算机与电子信息学院, 江苏 南京 210023;

3. 南京农业大学 信息管理学院, 江苏 南京 210095)

摘 要: 中文古籍数量庞大, 亟待智能处理方法进行自动处理。古文的自动分词和词性标注, 是古汉语信息处理的基础任务。而大规模词库和标注语料库的缺失, 导致古汉语自动分析技术发展较慢。该文介绍了第一届古代汉语分词和词性标注国际评测的概况, 评测以人工标校的精加工语料库作为统一的训练数据, 以 F_1 值作为评测指标, 比较了古汉语词法分析系统在测试数据(基测集和盲测集)上的优劣。评测还根据是否使用外部资源, 区分出开放和封闭两种测试模式。该评测在第十三届语言资源与评测会议的第二届历史和古代语言技术研讨会上举办, 共有 14 支队伍参赛。在基测集上, 封闭测试模式分词和词性标注的 F_1 值分别达到了 96.16% 和 92.05%, 开放测试模式分词和词性标注的 F_1 值分别达到了 96.34% 和 92.56%。在盲测集上, 封闭测试分词和词性标注的 F_1 值分别达到 93.64% 和 87.77%, 开放测试分词和词性标注 F_1 值则分别达到 95.03% 和 89.47%。未登录词依然是古代汉语词法分析的瓶颈。该评测的最优系统把目前古汉语词法分析提高到新的水平, 深度学习和预训练模型有力地提高了古汉语自动分析的效果。

关键词: 古汉语; 评测; 自动分词; 词性标注; 古文信息处理

中图分类号: TP391

文献标识码: A

Review of the First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff

LI Bin¹, YUAN Yiguo¹, LU Jingya¹, FENG Minxuan¹, XU Chao¹, QU Weiguang², WANG Dongbo³

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;

2. School of Computer and Electronic Information, Nanjing Normal University, Nanjing, Jiangsu 210023, China;

3. College of Information Management, Nanjing Agricultural University, Nanjing, Jiangsu 210095, China)

Abstract: Automatic word segmentation and part-of-speech tagging of ancient texts are the basic tasks of ancient Chinese information processing. The lack of large-scale vocabulary and annotated corpus leads to the slow development of ancient Chinese processing technology. The paper summarizes the First International Ancient Chinese Word Segmentation and POS Tagging Bakeoff, which provides manually annotated corpus as unified training data and basic test set and blind test set. The bakeoff also distinguishes open and close test mode according to whether external resources are used. The bakeoff was held at the Second Workshop on Language Technologies for Historical and Ancient Languages (LT4HALA), which is in the context of the 13th Edition of the Language Resources and Evaluation Conference (LREC). A total of 14 teams participated in the bakeoff. On the basic test set, the F1-scores of word segmentation and POS tagging reaches 96.16% and 92.05%, respectively, in the close test, while 96.34% and 92.56%, respectively, in the open test. On the blind test set, the F1-scores of word segmentation and POS tagging reaches 93.64% and 87.77%, respectively, in the close test, while 95.03% and 89.47%, respectively, in the open test. The out-of-vocabulary words are still the barrier of ancient Chinese lexical analysis, and the deep learning and pre-training model effectively improve the performance of automatic ancient Chinese processing.

Keywords: ancient Chinese; evaluation; word segmentation; POS tagging; ancient language information processing

收稿日期: 2022-12-12 定稿日期: 2023-02-07

基金项目: 国家社会科学基金(21ZD&331); 江苏省社会科学基金(20JYB004); 国家语委项目(YB145—41); 古籍工作重点课题(22GJK006)

0 引言

我国古籍数量庞大,蕴含着丰富的人文知识,亟待用智能处理方法进行自动处理。汉语文本的词与词之间没有边界标记,词语也没有形态变化,所以从文本中识别出词和词性就成为中文信息处理的基础任务^[1-2],即自动分词和词性标注^[3]。为了推动现代汉语文本分词和词性标注的技术进步,已经举办了诸多评测竞赛,特别是2003年到2014年间,国际学术组织 SIGHAN(Special Interest Group of Han)举办了七届中文分词评测^[4-10],推动了现代汉语自动分词技术的发展,SIGHAN2014 最优分词系统在现代汉语多领域数据集上的分词 F_1 值达到了97.3%^[10]。

古籍文本也亟待分词与词性标注的技术突破,以进行内容挖掘和利用。古籍文本,主要由文言文写成。文言文包括两大类:一类是先秦和西汉时期文献的语言,一类是东汉以后的书面语,它们在词汇语法系统上具有一致性^[11],具有悠久的历史。根据《中国古籍总目》^[12],现存的古代文献总量超20万种,数量庞大,亟需快速高效的自动处理工具对文本进行分词和词性标注。然而,古汉语的用字、词汇、语法和现代汉语差异显著,无法直接使用现代汉语词法分析工具处理古汉语文本。同时,古汉语缺少大规模的词库和标注语料库,没有足够的训练语料提高精度,也缺乏有效的手段评估软件系统的优劣。作为基础的古汉语自动分词任务,在深度学习模型应用方面仍存在不足^[13]。因此,古汉语词法分析需要建立一个标准的评测任务,以规范的数据集和评测指标推进古汉语词法分析技术的发展。

在此背景下,我们组织了第一届古代汉语分词和词性标注国际评测(EvaHan2022)^①。该评测在第二届历史和古代语言技术研讨会(Workshop on Language Technologies for Historical and Ancient Languages, LT4HALA)上举办,该研讨会是在法国马赛举办的第十三届语言资源与评测会议(13th Conference on Language Resources and Evaluation, LREC2022)^②的子会议。本次评测由南京师范大学文学院计算语言学与数字人文研究组和南京农业大学信息管理学院组织,由北京大学数字人文研究中心、中国人工智能学会语言智能专委会、中国中文信息学会青年工作委员会、江苏省人工智能学会自然语言处理专委会、江苏省语言学会协办。评测的详

细信息已经在 GitHub^③ 上公布。

1 评测任务

EvaHan2022 的词法分析评测任务是一个综合任务,包括分词和词性标注。该任务需要在古汉语文本的词与词之间添加边界标记(空格符),并标注出每个词的词性,参见表1。评测将会分别对分词和词性标注给出评分,但不接受仅分词后的结果。

表1 EvaHan2022 分词与词性标注示例

原始文本	亟請於武公,公弗許。
分词后的文本	亟 請 於 武公, 公 弗 許。
词性标注后的文本	亟/d 請/v 於/p 武公/nr ,/w 公/n 弗/d 許/v 。/w

2 数据集

EvaHan2022 的数据集分为训练集和两个测试集,使用了《左传》《史记》和《资治通鉴》三本古籍的繁体语料,都经过标点、分词和词性标注处理,并经过双重人工校对。

2.1 标记集与数据格式

EvaHan2022 的数据集均按照南京师范大学先秦汉语分词规范和词类标记集^[14]进行标注,使用21个词性标记(表2),数据存储格式为纯文本 UTF-8,原始文本包括字符和标点。人工标注的结果则以“词/词类(空格)”的方式排列,如表1所示,每一个词都用空格分开,标点也以相同的方式处理。

表2 词类标注基本集

序号	标记	名称	例词
1	a	形容词	大
2	c	连词	则
3	d	副词	不
4	f	方位词	前
5	j	连词	焉
6	m	数词	一

① <https://circse.github.io/LT4HALA/2022/EvaHan>

② <https://lrec2022.lrec-conf.org/en/>

③ https://github.com/CIRCSE/LT4HALA/tree/master/2022/data_and_doc

续表

序号	标记	名称	例词
7	n	普通名词	人
8	nr	人名	孔子
9	ns	地名	齊
10	p	介词	於
11	q	量词	匹
12	r	代词	吾
13	s	拟声词	噫噫
14	t	时间词	五月
15	u	助词	之
16	v	动词	如
17	vs	使动用法	驚
18	vy	意动用法	寶
19	vw	为动用法	泣
20	y	语气词	乎
21	w	标点	,

2.2 训练集和测试集

训练集为《左传》的前十卷,约占《左传》全文的 5/6。测试集分为基测集 TestA 和盲测集 TestB。基测集 TestA 为《左传》后两卷,约占全文的 1/6,用来评测词法分析系统在同质语料上的表现。训练集与 TestA 已经在语言数据联盟(Linguistic Data Consortium, LDC)^①上发布。评测不允许参赛团队将 TestA 用作训练数据。已有一些相关研究在《左传》数据集上进行了实验与评测^[15-19],可以作为各参赛队的参考。盲测集 TestB 则为新标注数据,从未公开过,其语料选自《史记》的《周本纪》《秦本纪》和《资治通鉴》的《周纪》,规模略大于 TestA。与 TestA 相似的是,TestB 同样选自记载了春秋战国历史的史书典籍,目的是尽可能地减少未登录词的影响,更有效地评估各参赛系统。训练集与测试集的规模如表 3 所示。

表 3 EvaHan2022 的数据集规模

数据集	字例数	词例数
训练集	194 995	166 142
测试集	基测集 TestA	33 297
	盲测集 TestB	62 969

本次评测仅将人工标校过的训练集与原始的标点测试集文本提供给参赛队,人工标校过的测试集则用于评估各参赛队词法分析系统的表现。评测结束后,人工标校过的测试集也提供给了各参赛队,以进行错误分析与技术报告的撰写。

3 评测

3.1 评测指标

评测指标是衡量一个参赛队表现的关键考量。我们主要参考了现代汉语分词与词性标注评测 SIGHAN2008^[7],方法是将参赛队提交的自动分析结果与人工标校过的黄金标准数据(Gold Data)进行对比计算。

评测指标为准确率 P (Precision)、召回率 R (Recall) 和 F_1 值(F_1 -score),计算方法如式(1)~式(3)所示。

$$P = \frac{\text{参赛系统标注正确的词例数量}}{\text{参赛系统标注出的词例数量}} \times 100\% \quad (1)$$

$$R = \frac{\text{参赛系统标注正确的词例数量}}{\text{黄金标准标注词例数量}} \times 100\% \quad (2)$$

$$F_1 = \frac{2 \times P \times R}{P + R} \times 100\% \quad (3)$$

分词和词性标注的结果分别计算,且词性标注的结果是在分词结果的基础上进行评测的,如果参赛系统词性标注正确但分词错误,我们同样认为其词性标注是错误的。最终基于 F_1 调和平均值对参赛队结果进行排名。

3.2 两种测试模式

参照国际评测的通用方式,EvaHan2022 设置了封闭测试和开放测试两种模式。两种测试模式可以使用的资源情况见表 4。在封闭测试模式下,每

表 4 EvaHan2022 的两种评测模式的资源使用要求

资源	测试模式	
	封闭模式	开放模式
预训练模型	仅 SikuRoBERTa	无限制
训练数据	仅《左传》训练集	无限制
外部特征	禁止	无限制
人工校改	禁止	禁止

① <https://catalog.ldc.upenn.edu/LDC2017T14>

个团队可以使用的训练资源是给定的,每个团队只能使用指定的《左传》训练集和预训练模型 SikuRoBERTa 进行训练。SikuRoBERTa 是在五亿字的繁体《四库全书》(部分)上,基于 RoBERTa 预训练模型^[20]训练得到的古汉语预训练模型。该模式不允许使用任何其他资源,包括选用外部特征,更不允许人工校改数据。在开放测试模式下,资源、数据和模型均没有限制。例如,可以融合汉字的部首、音韵特征,使用更多的训练语料,使用其他的预训练模型等,但同样不允许人工校改数据。每个参赛队必须在最终报告中说明其系统使用的所有资源、数据和模型。

3.3 评测流程

评测的训练数据自 2021 年 12 月 20 日发布以供下载,测试数据于 2022 年 3 月 31 日发布,评测结

果于 2022 年 4 月 7 日发布。各参赛队于 5 月 27 日提交报告论文最终版。6 月 25 日在法国马赛召开的 LREC2022 的 LT4HALA 会议上,举行了 EvaHan2022 的研讨。

4 参赛队以及评测结果

4.1 参赛队

共有 14 支团队参加了 EvaHan2022,提交了 55 份运行结果。表 5 列出了参赛队的基本信息。所有的参赛团队都在封闭测试模式中提交了运行结果文件,少数参赛队参加了开放测试模式。在参赛队提交的结果文件里,有四份文件存在格式错误(表 5 中标为“+”),我们对其进行了格式更正后予以评测;有两份文件逾期提交(表 5 中标为“*”)。

表 5 EvaHan2022 参赛队的基本信息

ID	队伍名称	单位	TestA		TestB	
			封闭	开放	封闭	开放
1	BIT1	北京理工大学	1	0	1	0
2	BIT2	北京理工大学	1	0	1	0
3	BLCU	北京语言大学	2	2	2	2
4	BUPT	北京邮电大学	1	0	1	0
5	FDU	复旦大学	2	2	2	2
6	GDUFS	广东外语外贸大学	2	0	2	0
7	HIT	哈尔滨工业大学	2	2	2	2
8	IMUT	内蒙古工业大学	1	0	1	0
9	NJU	南京大学	2	0	2	0
10	NJUPT	南京邮电大学	1 *	1 ⁺	1 *	0
11	NUAA	南京航空航天大学	1	0	0	2
12	THU	清华大学	1	0	1	0
13	ZNNU	南京师范大学中北学院	1 ⁺	1 ⁺	0	1 ⁺
14	ZYB	作业帮教育科技(北京)有限公司	2	0	2	0
合计		55	20	8	18	9

4.2 评测结果

55 份结果文件的评测结果整体呈现出较好的分析效果。为了梳理在两个测试集和两种模式下的评测结果,我们分作四张表(表 6~表 9)进行呈现。其中,分词的准确率、召回率和 F_1 值缩写为 WP、

WR 和 WF;词性标记的准确率、召回率和 F_1 值的缩写为 PP、PR 和 PF。各参赛队的成绩按照 PF(词性标记 F_1 值)来降序排序。

在 TestA 中,封闭测试模式分词和词性标注的最高 F_1 值分别达到 96.16% 与 92.05%,开放测试模式分词和词性标注的最高 F_1 值为 96.34% 和

表 6 封闭测试模式下各参赛队在 TestA 上的表现 (单位: %)

参赛队	WP	WR	WF	PP	PR	PF
FDU	95.39	96.68	96.03	91.43	92.67	92.05
	95.57	96.67	96.12	91.50	92.55	92.02
BIT1	95.18	96.49	95.83	90.96	92.22	91.59
BUPT	95.81	96.52	96.16	90.90	91.57	91.24
NUAA	95.63	96.33	95.98	90.88	91.54	91.21
GDUFS	94.85	96.52	95.68	90.34	91.93	91.13
THU	94.70	95.72	95.20	89.59	90.55	90.07
NJU	94.15	95.46	94.80	89.29	90.53	89.90
	94.18	95.44	94.81	89.28	90.47	89.87
GDUFS	92.27	95.46	93.84	88.14	91.18	89.63
BIT2	94.48	94.99	94.74	88.95	89.43	89.19
IMUT	94.67	93.10	93.88	89.73	88.24	88.98
ZYB	94.90	95.07	94.99	88.30	88.46	88.38
ZNNU	92.76	91.45	92.10	88.80	87.54	88.16
ZYB	94.86	94.95	94.90	87.49	87.58	87.53
HIT	90.78	93.03	91.89	84.70	86.80	85.74
	90.81	92.99	91.89	84.72	86.77	85.73
BLCU	91.39	93.22	92.29	84.39	86.09	85.23
	91.39	93.27	92.32	84.20	85.93	85.05
NJUPT	78.13	86.32	82.03	58.48	64.61	61.39

表 7 开放测试模式下各参赛队在 TestA 上的表现 (单位: %)

参赛队	WP	WR	WF	PP	PR	PF
FDU	95.81	96.88	96.34	92.05	93.07	92.56
	95.73	96.84	96.28	91.88	92.94	92.41
ZNNU	92.78	90.18	91.46	88.97	86.48	87.71
HIT	91.20	93.49	92.33	85.41	87.56	86.47
	91.09	93.41	92.24	85.27	87.45	86.35
BLCU	90.91	92.40	91.65	83.55	84.92	84.23
	90.56	92.29	91.41	83.13	84.72	83.92
NJUPT	78.14	86.31	82.02	57.35	63.35	60.20

表 8 封闭测试模式下各参赛队在 TestB 上的表现 (单位: %)

参赛队	WP	WR	WF	PP	PR	PF
FDU	94.72	91.99	93.34	89.07	86.50	87.77
	94.65	91.68	93.14	88.98	86.19	87.57
BIT1	94.48	91.70	93.07	88.40	85.80	87.08

续表

参赛队	WP	WR	WF	PP	PR	PF
GDUFS	94.59	92.70	93.64	87.87	86.12	86.99
	92.81	93.20	93.01	86.58	86.94	86.76
BUPT	94.04	90.59	92.28	86.86	83.67	85.24
THU	93.51	90.35	91.90	86.38	83.32	84.82
IMUT	93.65	86.43	89.89	87.05	80.33	83.56
BIT2	93.07	88.90	90.94	85.45	81.61	83.49
ZYB	93.59	89.89	91.70	84.69	81.34	82.98
	93.61	89.97	91.75	84.00	80.74	82.33
NJU	90.00	87.94	88.96	80.89	79.03	79.95
	89.56	87.31	88.42	80.56	78.53	79.53
BLCU	87.72	84.50	86.08	75.32	72.55	73.91
	87.65	84.61	86.10	75.21	72.60	73.88
HIT	82.79	78.82	80.75	71.37	67.95	69.62
	82.19	77.82	79.94	70.21	66.45	68.27
NJUPT	81.24	85.13	83.14	58.25	61.04	59.62

表 9 开放测试模式下各参赛队在 TestB 上的表现 (单位: %)

参赛队	WP	WR	WF	PP	PR	PF
ZNNU	95.26	94.79	95.03	89.70	89.25	89.47
FDU	94.97	92.26	93.60	89.16	86.62	87.87
	94.81	91.94	93.35	88.85	86.16	87.48
NUAA	94.50	91.69	93.07	87.79	85.18	86.47
	94.49	91.69	93.07	87.79	85.18	86.46
BLCU	87.09	83.43	85.22	73.99	70.88	72.40
	87.03	83.38	85.16	73.48	70.40	71.91
HIT	83.27	79.30	81.24	71.81	68.38	70.05
	82.23	78.31	80.22	70.77	67.40	69.04

92.56%。在 TestB 中,封闭测试模式分词和词性标注的最高 F_1 值分别达到 93.64% 与 87.77%,开放测试模式分词和词性标注的最高 F_1 值为 95.03% 和 89.47%。根据各参赛系统在封闭测试模式和开放测试模式中对两个测试集的标注效果,FDU 系统综合效果最优。

总体来看,由于词性标注是在分词结果上进行的,所有参赛队的分词任务得分均高于词性标注任务。对比各参赛队封闭测试模式与开放测试模式的结果,外部资源可以小幅提升各参赛系统的词法分

析表现。对比不同测试集的评测结果,各参赛队在更接近训练集的 TestA 上表现更好。

4.3 评测结果的比较与分析

为了更直观地比较与说明最优参赛系统标注效果的优劣,EvaHan2022 参考了 SIGHAN2008 的评测方法^[7],为分词与词性标注任务分别设置了 Baseline(基线)与 Topline(顶线)作为性能上下界进行参考。

4.3.1 分词

分词 Baseline 和 Topline 都是基于正向最大匹配算法完成的,区别在于使用的词表不同。Baseline 的词表源自训练集,Topline 模型的词表源自测试集。表 10 和表 11 给出了分词 Baseline 模型和 Topline 模型在两个测试集上的效果。

表 10 分词 Baseline 模型的效果 (单位:%)

测试集	WP	WR	WF
TestA	84.98	89.20	87.04
TestB	80.43	85.28	82.78

表 11 分词 Topline 模型的效果 (单位:%)

测试集	WP	WR	WF
TestA	99.04	98.20	98.62
TestB	98.48	97.11	97.79

我们对比了参赛系统中分词最高 F_1 值对 Baseline 模型的提高率以及同 Topline 模型的差距,如表 12、表 13 所示。大多数参赛系统的分词 F_1 值超过了 Baseline 模型,但仍与 Topline 模型有一定差距。

表 12 最优参赛系统对分词 Baseline 模型的提高率 (单位:%)

测试集	WP	WR	WF
TestA	+10.83	+7.68	+9.30
TestB	+14.83	+9.51	+12.25

表 13 最优参赛系统与分词 Topline 模型的差距 (单位:%)

测试集	WP	WR	WF
TestA	-3.23	-1.32	-2.28
TestB	-3.22	-2.32	-2.76

4.3.2 词性标注

词性标注 Baseline 模型和 Topline 模型采用最

大概率法完成标注,具体标注方法为:①从标注语料中统计得到词与其词性标记的频率;②直接将仅有一种词性标记的词标注为该词性标记;③对于有多种词性标记的词,将其标注为标注语料中该词最高频的词性标记;④对于未在标注语料中出现的词,将其标注为标注语料中最高频的词性标记。两者的区别在于 Baseline 模型使用的标注语料为训练集,Topline 模型的为测试集。

词性标注 Baseline 模型在分词 Baseline 模型的分词结果上进行标注,其效果如表 14 所示。词性标注 Topline 模型在分词 Topline 模型的分词结果上进行标注,其效果如表 15 所示。

表 14 词性标注 Baseline 模型的效果 (单位:%)

测试集	PP	PR	PF
TestA	75.93	79.70	77.77
TestB	66.83	70.87	68.79

表 15 词性标注 Topline 模型的效果 (单位:%)

测试集	PP	PR	PF
TestA	91.76	90.99	91.37
TestB	89.77	88.51	89.14

我们同样对比了参赛系统中词性标注最高 F_1 值对 Baseline 模型和 Topline 模型的提高率,如表 16 与表 17 所示。

表 16 最优参赛系统对词性标注 Baseline 模型的提高率 (单位:%)

测试集	PP	PR	PF
TestA	+16.12	+13.37	+14.79
TestB	+22.87	+18.37	+20.68

表 17 最优参赛系统对词性标注 Topline 模型的提高率 (单位:%)

测试集	PP	PR	PF
TestA	+0.29	+2.08	+1.19
TestB	-0.07	+0.74	+0.33

绝大多数参赛系统的词性标注效果要优于 Baseline 模型,部分参赛系统的得分甚至超过了 Topline 模型,可以看出深度学习模型的高性能。

4.4 模型与方法

根据赛后论文和会议研讨的内容,我们进一步

分析了 14 支参赛队在模型与方法上的得失,总结出两点结论。

(1) 预训练和深度学习模型有效提高了分析效果。大部分参赛队的得分超过了本次评测设置的分词与词性标注 Baseline 模型,小部分参赛队的得分超过了 Topline 模型,这与大部分参赛队采用的深度学习方法有关。与 Baseline 模型和 Topline 模型采用的方法相比,深度学习可以自动地对输入的低阶特征进行组合、变换,得到高阶特征,大大减少了特征工程的工作量,而且能够从训练数据中自动学习数十个甚至上百个连续的表示层。加上基于《四库全书》的预训练模型,通过运用神经网络模型,大部分参赛队在古汉语分词和词性标注上取得了不错的分数。

(2) 融合外部知识与多种模型,解决数据稀疏问题。本次评测中,不少参赛队将几种模型与方法创新地引入古汉语自动处理领域,非常具有启发性意义。例如,在本次竞赛中获得最好成绩的 FDU 队,针对 BERT 模型局部语义特征建模能力不足以及古汉语标注语料稀少的现状,引入了局部语义增强(Local Semantic Enhancement)和知识融合(Knowledge Fusion)两种策略。局部语义增强策略引入的是二元语法(Bigram)特征,具体来说,句子中每个字的向量化表示由三个向量拼接而成,包括 SikuRoBERTa 模型提供的字向量,以及由两个线性二元语法层(Linear Bigram Layer)分别得到的当前字与前字的二元语法向量化表示、当前字与后字的二元语法向量化表示。外部知识融合策略则是先使用 Monte-Carlo Dropout^[21]获得待标注句中的不确定成分,再通过几部先秦古籍中含有该成分且语义相似度最接近的几个句子作为外部知识辅助该句的标注,较好地实现了语言知识与深度学习的融合。

此外,获得第二名的 BIT1 团队,引入了集成学习(Ensemble Learning)和对抗训练(Adversarial Training),丰富了应用于古汉语自动分析的方法。获得第三名的 BUPT 团队,将常用于序列标注的 BERT-BiLSTM-CRF 中的条件随机场模型(Conditional Random Fields, CRF)替换为多层感知机

(Multi-layer Perceptron, MLP),取得了不错的成绩。

同时,针对古汉语标注数据集规模不大的现状,BIT2、GDUFs、NJU 三个参赛队均采用了数据增强(Data Augmentation)的方法,通过近义词、同词性词互相替换或动态掩码(Mask)方法,实现了数据增强,为今后古汉语文本自动处理研究提供了新思路。

4.5 未登录词问题

深度学习方法本质上仍是在学习数据基础上进行特征提取,即对训练集进行建模,如果测试集与训练集差异较大,神经网络模型的表现往往不佳。在词法分析上,这种差异主要表现为未登录词问题。对古汉语来说,人名和地名是两类主要的未登录词。对于处在一定句式中的人名和地名,即使这些词是未登录词,训练好的神经网络模型往往也能标注正确,例如,TestB 的“絳疵謂智伯曰”一句中,“絳疵”和“智伯”都是训练集中未出现的词语,但由于处在“某人謂某人曰”的句式中,经过训练的神经网络模型很容易将其标注正确,但像 TestB 的“唯輔果在”一句中,“輔果”这一未登录词的数据比较稀少,绝大部分参赛系统未能标注正确。

我们以训练集的词表为分词词表,统计了 TestA 和 TestB 中未登录词的词型数占比和词例数占比,如表 18 所示。可以看到,与 TestA 相比,TestB 中未登录词的占比要高很多,这是各参赛队对 TestA 的标注效果往往要比 TestB 更好的原因之一。未登录词仍是古汉语分词与词性标注中待解决的难题。

表 18 TestA 与 TestB 中未登录词的占比 (单位:%)

测试集	未登录词 词型数占比	未登录词 词例数占比
TestA	38.72	6.80
TestB	51.13	12.57

以本次评测中成绩最好的 FDU 参赛系统的最优分词和词性标注结果为例,我们计算了其模型对各测试集未登录词的分词和词性标注效果,如表 19

表 19 FDU 参赛系统对未登录词的分词和词性标注效果

(单位:%)

测试集	WP _{OOV}	WR _{OOV}	WF _{OOV}	PP _{OOV}	PR _{OOV}	PF _{OOV}
TestA	89.89	83.17	86.40	83.16	76.95	79.93
TestB	82.04	91.50	86.51	76.22	85.01	80.37

所示。FDU 参赛系统对未登录词的分词和词性标注的 F_1 值分别在 86% 和 80% 左右,远低于其对整个测试集分词和词性标注的 F_1 值。而在 TestA 和 TestB 中,未登录词的占比较高。本次评测的参赛队仍未较好地解决古汉语未登录词问题。

整体来看,古汉语文本中未登录词的词例数占比较高,影响了未登录词的分词和词性标注效果,如何解决好未登录词的识别和标注问题应是今后古代汉语自动分词和词性标注的研究重点。

5 总结与展望

本文介绍了第一届古汉语自动分词与词性标注国际评测的概况。EvaHan2022 作为国内外首次举办的古代汉语分词和词性标注评测,吸引了 14 支队伍参赛。不少参赛队将多种机器学习模型与方法创新地引入古汉语自动处理领域,特别是预训练和深度学习方法有效提升了古文自动分析的效果。其中,复旦大学的团队在本次评测中取得了最优成绩,其引入的局部语义增强方法,有效增强了模型对局部语义特征的建模效果,其不确定度采样和外部知识融合策略对古汉语这种资源稀缺语言有较高的应用价值。与基测集相比,各参赛队在盲测数据集上的表现均有所下降,说明未登录词问题依然是古汉语词法分析的瓶颈难题。

未来的工作中,我们将继续扩大古汉语高质量标注语料库的规模,以更好地评估古汉语词法分析系统的性能。其次,我们将在不同文体和跨时代的古汉语文本上进行古汉语词法分析评测,以评估各参赛系统的泛化性和鲁棒性。最后,逐步将古汉语的断句、标点、专名识别、句法分析等自然语言处理任务也纳入评测,让古汉语自动分析技术逐步走向成熟与实用。

参考文献

- [1] 孙茂松,邹嘉彦.汉语自动分词研究评述[J].当代语言学,2001(01): 22-32.
- [2] XUE N.Chinese word segmentation as character tagging[C]//Proceedings of the International Journal of Computational Linguistics & Chinese Language Processing, Special Issue on Word Formation and Chinese Language Processing, 2003: 29-48.
- [3] 张开旭,孙茂松.执行中文分词和词性标注的统一框架[J].中国通信,2012,9(03): 1-9.
- [4] SPROAT R,EMERSON T.The first international Chinese word segmentation bakeoff[C]//Proceedings of the 2nd SIGHAN Workshop on Chinese Language Processing, 2003: 133-143.
- [5] EMERSON T.The second international Chinese word segmentation bakeoff [C]//Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing, 2005: 123-133.
- [6] LEVOW G A.The third international Chinese language processing bakeoff: Word segmentation and named entity recognition[C]//Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing, 2006: 108-117.
- [7] JIN G,CHEN X.The 4th international Chinese language processing bakeoff: Chinese word segmentation, named entity recognition and Chinese POS tagging [C]//Proceedings of the 6th SIGHAN Workshop on Chinese Language Processing, 2008: 69-81.
- [8] ZHAO H,LIU Q.The CIPS-SIGHAN CLP Chinese word segmentation backoff [C]//Proceedings of the CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2010.
- [9] DUAN H,SUI Z,TIAN Y,et al.The CIPS-SIGHAN CLP Chinese word segmentation on microblog corpora bakeoff [C]//Proceedings of the 2nd CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2012: 35-40.
- [10] DUAN H, SUI Z, GE T. The CIPS-SIGHAN CLP Chinese word segmentation bakeoff [C]//Proceedings of the 3rd CIPS-SIGHAN Joint Conference on Chinese Language Processing, 2014: 90-95.
- [11] 蒋绍愚.也谈文言和白话[J].清华大学学报(哲学社会科学版),2019,34(02): 1-13.
- [12] 中国古籍总目编纂委员会.中国古籍总目[M].北京: 中华书局,2009-2013.
- [13] 邓三鸿,胡昊天,王昊,等.古文自动处理研究现状与新时代发展趋势展望[J].科技情报研究,2021,3(01): 1-20.
- [14] 陈小荷,冯敏萱,徐润华.先秦文献的信息处理[M].北京: 世界图书出版公司,2013.
- [15] 石民,李斌,陈小荷.基于 CRF 的先秦汉语分词标注一体化研究[J].中文信息学报,2010,24(02): 39-45.
- [16] 程宁,李斌,葛四嘉,等.基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究[J].中文信息学报,2020,34(04): 1-9.
- [17] 俞敬松,魏一,张永伟,等.基于非参数贝叶斯模型和深度学习的古文分词研究[J].中文信息学报,2020,34(06): 1-8.
- [18] 王东波,刘畅,朱子赫,等.SikuBERT 与 SikuRoBERTa: 面向数字人文的《四库全书》预训练模型构建及应用研究[J].图书馆论坛,2022,42(06): 31-43.

(下转第 64 页)

搜索空间的前提下有一定提升,验证了路径语义集成和空洞循环卷积的作用。由于卷积网络参数共享,以及采用了分阶段并行训练方法,嵌入维度为

500 时 $APAC_q$ 得分超越 QuatE 最优表现(维度 1 000)只需要 1.5 小时,仅为后者训练时间的约 1/5。



陈新元(1984—),硕士,副教授,主要研究领域为自然语言处理、知识表示与推理。

E-mail: 3122955356@qq.com



周忠眉(1965—),通信作者,博士,教授,主要研究领域为数据挖掘。

E-mail: 20163063@qq.com



陈庆强(1965—),硕士,教授,主要研究领域为图像处理、统计学习。

E-mail: 3204193260@qq.com



(上接第 53 页)

- [19] 刘畅,王东波,胡昊天,等.面向数字人文的融合外部特征的典籍自动分词研究:以 SikuBERT 预训练模型为例[J].图书馆论坛,2022,42(06): 44-54.
- [20] LIU Y,OTT M,GOYAL N,et al.Roberta: A robustly optimized bert pretraining approach[J].arXiv preprint arXiv: 1907.11692,2019.

- [21] GAL Y,GAHRAMANI Z.Dropout as a bayesian approximation: Representing model uncertainty in deep learning[C]//Proceedings of the International Conference on Machine Learning. PMLR,2016: 1050-1059.



李斌(1981—),博士,副教授,主要研究领域为计算语言学。

E-mail: libin.njnu@gmail.com



袁义国(1996—),通信作者,硕士,主要研究领域为计算语言学。

E-mail: lexcliff1023@gmail.com



芦靖雅(1999—),硕士研究生,主要研究领域为计算语言学。

E-mail: dorismurraylu@163.com