文章编号: 1003-0077(2010)02-0039-07

基于 CRF 的先秦汉语分词标注一体化研究

石 民, 李 斌, 陈小荷

(南京师范大学 文学院, 江苏 南京, 210097)

摘 要: 该文探索了古代汉语, 特别是先秦文献的词切分及词性标注。首先对《左传》文本进行了词汇处理(分词和词性标注)和考察分析, 然后采用条件随机场模型(CRF), 进行自动分词、词性标注、分词标注一体化的对比实验。结果表明, 一体化分词比单独分词的准确率和召回率均有明显提高, 开放测试的 F 值达到了 94.60%; 一体化词性标注的 F 值达到了 89.65%, 比传统的先分词后标注的"两步走"方法有明显提高。该项研究可以服务于古代汉语词汇研究和语料库建设, 以弥补人工标注的不足。

关键词: 计算机应用; 中文信息处理; 先秦汉语; 分词; 词性标注; 左传; 条件随机场模型

中图分类号: TP391 文献标识码: A

CRF Based Research on a Unified Approach to Word Segmentation and POS Tagging for Pre-Qin Chinese

SHI Min, LI Bin, CHEN Xiaohe

(School of Chinese language and literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China)

Abstract: This paper explores the cross field between NLP and ancient Chinese, particularly the pre-Qin documents. The text of "Zuo Zhuan" is firstly analyzed after manual segmentationand POS tagging. Then the Conditional Rarndom Fields model (CRF) is adopted for the word segmentation (WS), POS tagging (PT) and a unified process of WS and PT, respectively. The precision and recall of the unified approach are much higher than the independent WS and PT in the open test, with a F-score of 94.60% in WS and 89.65% in PT. This method is suitable for the study of ancient Chinese vocabulary and corpus construction, and can be applied to compensatethe manual tagging.

Key words: computer application; Chinese information processing; Pre-Qin Chinese; word segmentation; POS tagging; Zuo Zhuan; conditional random fields model

1 引言

中文信息处理研究在现代汉语领域已经取得了比较丰硕的成果, 但古代汉语信息处理还有待探索。目前, 先秦文献的信息处理大体还处于字处理阶段, 以解决古文字的输入输出、文献逐字索引等问题为主要内容, 实用成果仅限于古籍文献的专题索引和查询。

我们正在实施的项目是"先秦汉语词汇统计与知识检索系统", 准备对 25 种最重要的先秦传世文献进行词语切分、词性标注、个别常用词(包括古今字和通假字)的词义标注, 建立先秦文献的词汇知识库以及历史知识库, 并研制相应的检索系统。要实现这一目标, 古文献的切分标注是古汉语语料库建设的一项基础性工作。先秦汉语以单字词为主, 也存在着一定量的多字词, 在缺少分词词典和训练语料的条件下, 分词标注仍有难度。正如古汉语计算

收稿日期: 2009-08-26 定稿日期: 2009-11-26

基金项目: 国家"211 工程"三期重点学科建设项目"语言科技创新及工作平台建设"子课题"先秦文献词汇统计与知识检索系统"

作者简介: 石民(1984—), 男, 硕士生, 研究方向为计算语言学; 李斌(1981—), 男, 博士, 讲师, 研究方向为计算语言学; 陈小荷(1952—), 男, 教授, 博导, 研究方向为计算语言学。

语言学家尉迟治平的呼吁:“我们期望能有可以用于汉语史电子文献自动分词、自动断句、自动标注的软件早日问世,专家只需对结果刊谬补缺,这将大大减轻属性式标注的劳动强度,加快工作进度[1]。”

针对古汉语的自动分词,已经有了一些研究成果。台北中研院的“汉籍电子文献”对以《十三经》为主的先秦文献进行了分词和词性标注,可以通过“瀚典全文检索系统”对文献进行检索、统计、搭配[2]。但文献数量还较少,分词标注方法也以较为传统的最大概率和隐马尔科夫模型为主。邱冰则提出一种启发式的混合分词方法,以反向最大匹配分词为主,同时统计已出现词语的频率和汉字间的互信息,一方面对高频词进行直接的提取,另一方面调整词表增加新的词语[3]。由于采用《汉语大词典》作为通用分词词典,存在一定的局限性。

汉语的分词和词性标注工作,通常是在自动分词的基础上,再进行词性标注。这种“两步走”的方法,存在错误扩散问题,会影响到最后的标注精度。白栓虎给出了汉语词切分和词性标注一体化的隐马尔科夫模型,并进行了小规模试验[4]。Hwee Tou Ng 和 Jin Kiat Low 则深入比较了两步走和一体化的优劣,提出基于字标注的一体化方法是最佳的方案,其分词系统获得 Sighan2003 四个测试语料中的三项封闭测试第一,同时又肯定了两步方案在训练和测试时间上的优势[5]。Yue Zhang 和 Stephen Clark 提出使用单一感知器模型的分词和标注一体化方法,由于充分利用了词性信息,分词准确率和召回率均有大幅提高[6]。这些研究表明,在现代汉语语料上,分词标注一体化方法效果较好,只是训练时间开销较大。

本文着力研究面向先秦文本的分词和词性标注,以人工标校的《左传》作为实验对象。首先进行了语料分析,然后分别设计了基于条件随机场模型(CRF)的自动分词、词性标注、分词标注一体化实验,以寻找适合古汉语分词标注的最佳方案。研究成果可以服务于古籍文献的语料库建设,将研究人员从繁重的语料标注工作中解脱出来,仅需校对机器自动处理的结果,也可以有效缓解人工标注一致性较差的问题。

2 先秦汉语分词标注规范及语料考察

2.1 语料来源

《左传》是先秦文献的经典之作,内容是传《春秋》的,即春秋时期各国的历史。篇幅约 23 万字,是先秦传世文献中单本字数最多的文献,非常适合用来作为机器学习的对象,服务于先秦其他文献的自动标注。

本文使用的语料底本,是由香港中文大学中国古籍研究中心建设的汉达文库的《左传》的“传”文。该文库收录的文献版本,均为旧刻善本,后由研究人员重新标点、校勘。为了保证语料质量,我们参照了古文献界较为公认的杨伯峻的《春秋左传注》[7],以解决异文(添字、缺字、异体字等)问题,不一致处按中华书局版校正。语料采用 Unicode 编码存储。

2.2 先秦汉语分词标注规范

确定古汉语的分词标准及词类体系,是分词标注的基本前提。我们参照了台北中研院的《资讯处理用分词规范》,采用词汇意义和语法功能兼顾的标准,确定出适合古汉语的分词单位及词类体系①。和中研院的主要差别是,将数词进行了捆绑处理,区分了三种常见的词类活用方式,共设立了 21 个词类标记(见表 1)。

表 1 先秦汉语词类标注基本集及词类统计信息

序号	名称		标记	词型/词例
1	名词	普通名词	n	4 794/28 722
		人名	nr	3 817/12 925
		地名	ns	1 150/6 866
		方位词	f	38/454
		时间词	t	179/2 348
2	动词		v	3 405/44 948
	使动用法		sv	20/26
	意动用法		yv	3/4
	为动用法		wv	3/5
3	形容词		a	449/1 732
4	数词		m	92/1 358
5	量词		q	20/65
6	代词		r	126/11 796
7	介词		p	34/5 992
8	连词		c	73/7 699
9	助词		u	13/3 533

① 可以查阅南京师范大学 CIPP 中文信息处理平台网站《先秦汉语分词标注规范》,http://www.cipp.cn/news_view.asp?id=76。

续表

序号	名称	标记	词型/ 词例
10	副词	d	329/ 10 072
11	语气词	y	50/ 6 032
12	拟声词	s	2/ 3
13	兼词	j	3/ 536
14	标点	w	15/ 49 678
合计			14 615/ 194 794

2.3　语料的人工标注和考察

四位语言学专业的研究生, 参照杨伯峻的注释和《春秋左传详解词典》[8], 对语料进行了人工切分标注和校对。本文所用语料版本为 V2. 0①。《左传》的传文部分, 共 179 792 个汉字, 除去标点, 共 3 308 个字型、14 600 个词型(区分词性)。其中, 多字词有 9 973 个词型, 占全部词型的 68. 31%, 但只占词例数的 21. 02% (见表 2), 平均词长为 1. 81 字。由此可见, 先秦汉语的基本特点是以单字词为主的, 同时, 多字词也是不可忽视的。如果整个语料按照单字来切分, 正确率大约只有 79%。因此, 如何处理多字词应成为分词的重点研究对象。

表 2　左传词型、词例统计(除去标点, 区分词性)

词长(字)	1	2	3	4	5	合计
词型数	4 627	8 348	1 363	256	6	14 600
词例数	114 616	26 814	3 202	477	7	145 116

3　实验及分析

3.1　实验语料

《左传》按照鲁国 12 个国君的谥号, 共分 12 卷。在实验中, 本文将前十卷作为训练语料, 后两卷作为测试语料, 训练测试比约为 6∶ 1(见表 3)。

表 3　训练测试语料情况

语料	汉字频次	切分单位频次
训练语料	153 648	166 536
测试语料	26 144	28 258
合计	179 792	194 794

3.2　基于 CRF 的分词实验

本文采用由字构词原理进行汉语自动分词, 将分词问题转化为词位信息的序列标注问题。CRF 是一个应用广泛的序列标注模型, 该模型允许增加复杂特征, 可以有效地处理标记偏置问题。实验采用 Taku Kudo 开发的"CRF++ 0. 53"工具包进行训练和测试。② 由于《左传》的平均词长为 1. 81 字, 且存在三字以上的词, 因此使用四词位标注集, 即 T = {B, M, E, S}, 其中 B 代表词首第一个字, E 代表词尾最末字, M 代表一个词中间的任意字, S 代表单字词和标点。语料样例见表 5 的"字符"列和"分词格式"列。

仿照 SIGHAN 竞赛, 我们给分词精度设定了 Baseline 和 Topline。分别为采用训练和测试语料的词表, 对测试语料进行正向最大匹配法分词, F 值分别为 83. 39% 和 96. 46%。

实验一　采用字面信息作为特征, 比较了上下文窗口为左右 1~ 3 个字, 以及二字、三字同现情况下的分词结果。

从表 4 可以看出, 任何一个分词结果都超过了 Baseline。增加二元、三元同现特征, 比单字上下文特征效果要好。在窗口为 ±1 个字、二元字同现 (1W+ 2) 的情况下, 精度最高, 达到了 93. 75%。

表 4　基于字面特征的分词评测结果

模板	1W	2W	3W	1W+ 2	1W+ 3	2W+ 2	2W+ 3	3W+ 2	3W+ 3
P/ %	90. 64	90. 50	90. 28	92. 95	91. 84	92. 94	92. 01	92. 85	91. 89
R/ %	92. 08	91. 75	91. 58	94. 57	93. 52	94. 37	93. 44	94. 37	93. 30
F/ %	91. 35	91. 12	90. 93	93. 75	92. 67	93. 65	92. 72	93. 60	92. 59

注: W 表示字符, nW 表示窗口[- n, n] 的字符, 2W 就表示上下文窗口[- 2, 2] 的字符;"+ 数字"表示字符的同现, + 2 就表示 2 个字符同现。以字符 W_i(i 为当前位置) 为例, 2W+ 2 就表示 W_{i-2}, W_{i-1}, W_i, W_{i+1}, W_{i+2}, W_{i-1}/W_i, W_i/W_{i+1}, W_{i-1}/W_{i+1}。

① CNCCL2009 会议论文所用语料为 V1. 0, 详见《中国计算语言学研究前沿进展》, P46- P51, 清华大学出版社, 2009 年 7 月出版。会后对语料进行了一次校对工作, 形成现在的版本 V2. 0。

② 下载地址为: http: / / crfpp. sourceforge. net/。

实验二 为了获得更佳的分词效果,以分词效果较好的“1W+ 2”、“2W+ 2”、“3W+ 2”3 个模板为基础,增加了一些语言学特征进行实验。这些特征包括字符分类、声、韵、调、部首。我们将字符分为“汉字(HZ)、普通标点(Punc)、句末标点(SenPunc)、西文数字(Num)、汉字数字(CNum)、干支(CCNum)”等类别。由于先秦汉语的声、韵、调皆为拟音推测,也没有比较公认的数据库,因此选取了描写中古汉语的《广韵》作为基本数据库来近似,为了保证字符的覆盖率,部首信息取自《康熙字典》。语料样例见表 5(“分词标注一体化格式”列除外)。

表 5 增加语言学特征的分词/一体化训练和测试语料样例

字符	col1 字符分类	col2 声	col3 韵	col4 调	col5 部首	分词格式		分词标注一体化格式	
						标准答案	标注结果	标准答案	标注结果
惠	HZ	匣	齊	去	心	B	B	B-nr	B-nr
公	HZ	見	東	平	八	E	E	E-nr	E-nr
元	HZ	疑	元	平	儿	B	B	B-n	B-n
妃	HZ	滂	微	平	女	E	E	E-n	E-n
孟	HZ	明	庚	去	子	B	B	B-nr	B-nr
子	D1	精	之	上	子	E	E	E-nr	E-nr
。	SenPunc	*	*	*	*	S	S	S-w	S-w

根据是否采用字符分类特征以及不同的特征组合、上下文窗口,分别进行了四组实验(见表 6)。第一组与第二、三组的区别为是否增加字符分类,二至四组主要是上下文窗口长度不同。

从实验结果来看:

(1) 增加字符分类特征有助于提高分词精度。使用字符分类特征的结果普遍好于不使用的结果。在“2W+ 2+ C1”和“2W+ 2+ C123”下,精度最高,F 值达到了 93.79%,且以“2W+ 2+ C1”为基础的模板性能最为稳定,实验效果普遍较好。因此,我们进一步增加了字符分类的二元同现特征,F 值提高到 93.92%(见表 7 前三列)。

(2) “2W+ 2+ C1”效果好也说明,字符二元同现是有效的特征。而宋彦在现代汉语分词实验中,六词位标记集在字符三元同现条件下效果最好[9]。这可能正是先秦汉语的特点造成的。现代汉语以多字词为主,三元同现可以提供充足的构词信息,而在古汉语中单字词居多,三元同现可能是冗余信息。

(3) 在字符分类基础上再增加声韵、声韵调、声韵调及部首,实验效果差别不大,特别是增加部首后,甚至出现了下降。究其原因,声韵调这三个特征本身也需要消除歧义。每个字的声韵调,在不同的词性或义项下往往是不同的,还需要仔细分析。而汉字的部首是不需要消歧的,分词精度的下降,说明

表 6 增加语言学特征模板的分词评测结果 1

模板	1W+ 2	1W+ 2+ C23	2W+ 2+ C23	1W+ 2+ C2345	2W+ 2+ C5
P/%	92.95	92.92	93.07	93.02	93.06
R/%	94.57	94.48	94.47	94.38	94.41
F/%	93.75	93.69	93.76	93.70	93.73
模板	1W+ 2+ C1	1W+ 2+ C123	1W+ 2+ C1234	1W+ 2+ C12345	1W+ 2+ C15
P/%	93.01	92.99	92.94	92.95	92.94
R/%	94.56	94.50	94.45	94.36	94.44
F/%	93.78	93.74	93.69	93.65	93.68
模板	2W+ 2+ C1	2W+ 2+ C123	2W+ 2+ C1234	2W+ 2+ C12345	2W+ 2+ C15
P/%	93.10	93.11	93.05	93.05	93.00
R/%	94.49	94.49	94.43	94.38	94.35
F/%	93.79	93.79	93.73	93.71	93.67
模板	3W+ 2+ C1	3W+ 2+ C123	3W+ 2+ C1234	3W+ 2+ C12345	3W+ 2+ C15
P/%	92.91	93.02	92.90	92.92	92.93
R/%	94.37	94.44	94.28	94.27	94.33
F/%	93.63	93.72	93.58	93.59	93.62

部首特征对于分类并无贡献。

实验三　先秦汉语的声韵系统本身就比较复杂, 我们使用的《广韵》是中古音系, 有 206 韵, 对于先秦汉语的声韵来说可能不太准确, 但调类只有“平、上、去、入”四类, 消歧也许相对容易, 为此本文在声韵调内部又做了对比实验。在模板选择上, 以“2W+ 2+ C1′”为基础模板, 然后分别增加声、韵、调特征(见表 7)。

表 **7**　增加语言学特征模板的分词评测结果 **2**

模板	1W+ 2+ C1′	2W+ 2+ C1′	3W+ 2+ C1′	2W+ 2+ C1′2	2W+ 2+ C1′3	2W+ 2+ C1′4	2W+ 2+ C1′24
P/%	93.09	93.20	93.03	93.22	93.11	93.22	93.25
R/%	94.65	94.65	94.52	94.64	94.53	94.65	94.65
F/%	93.86	93.92	93.77	93.92	93.81	93.93	93.94

注: C1′ 表示字符分类的二元同现。

通过表 7 与表 6 的对比, 我们发现字符分类二元同现特征能够提高分词精度, F 值最多提高了 0.15 个百分点。增加声、调特征后也有不同程度提高, 而加韵后明显降低, “2W+ 2+ C1′24”模板实验效果最佳, F 值达到了 93.94%。可见声、调对于汉字也是有效的特征, 但作用并不显著, 还需要进一步探讨。可以得出的初步结论是: 基于上下文两个汉字、二字同现、字符分类二元同现的模板“2W+ 2+ C1′”, 最适合《左传》的自动分词。

3.3　基于 **CRF** 的词性标注实验

词性标注是 CRF 模型的典型应用, 可以将词性标注问题视为词语的词类属性的序列化标注问题, 这里不再详述。特征选择上, 仅使用词形信息, 分别在上下文词语观察窗口为[-1, 1]、[-2, 2]、[-3, 3]的基础上增加词语二元同现。为了验证“两步走”方案在先秦语料上是否存在弊端, 在词性标注时, 分别对标准分词文本(Right, 即人工校对过的标准答案)和实验得到的最佳分词文本(BestSeg, 由 3.2 节复杂特征模板“2W+ 2+ C1′24”得到)进行了评测。

表 **8**　**CRF** 词性标注评测结果

模板	1W+ 2		2W+ 2		3W+ 2	
	BestSeg	Right	BestSeg	Right	BestSeg	Right
P/%	86.18	91.95	86.10	91.94	85.81	91.77
R/%	87.46		87.38		87.09	
F/%	86.82		86.74		86.45	

与单纯使用字面信息的分词实验一样, 表 8 中“1W+ 2”特征模板下的词性标注效果最好。在 BestSeg 和 Right 分词文本基础上, F 值分别达到了 86.82% 和91.95%。如果把 BestSeg 文本的分词精度 93.94% 和 Right 文本的词性标注精度 91.95% 相乘, 则可得到 BestSeg 文本词性标注的预测值 86.38%, 和实际测得的 86.82% 是相近的。实际测得的精度略高, 是由于标点部分的词性标注都是正确的, 不会受到分词错误的影响。

3.4　基于 **CRF** 的分词标注一体化实验

我们将“由字构词”的方案应用到词性标注问题上, 让汉字承载分词和词性的双重信息, 即该字所属词的词性标记(n、v 等)以及该字在词中的词位信息(B、M、E、S)。例如: “范献子/nr”, “范”为词首 B, “子”为词尾 E, “献”为词内字 M。则词性标注格式为“范 B-nr, 献 M-nr, 子 E-nr”。语料样例见表 5(“分词格式”列除外)。

在 3.2 节的分词实验中, 使用语言学特征时, 我们得出模板“2W+ 2+ C1′”最适合《左传》的自动分词, 分别增加声、调特征也都有不同程度提高, 在模板“2W+ 2+ C1′24”上效果最佳, 因此在基于字的一体化标注时, 我们设计了“2W+ 2”、“2W+ 2+ C1”、“2W+ 2+ C1′”、“2W+ 2+ C1′2”、“2W+ 2+ C1′4”、“2W+ 2+ C1′24”六个模板进行对比实验。为了和上文的实验结果对比, 对一体化标注分别给出了分词和词性标注的评测结果。

从实验结果来看:

(1) 分词精度有较大提升。表 9 与表 7 相比, 一体化实验效果均优于单独分词, F 值最多提高了 0.66 个百分点, 说明一体化方法能将汉字的词位信息和所属词的词性信息结合起来, 有效提高分词效果。

(2) 词性标注精度明显提升。表 9 与表 8 中基于 BestSeg 文本的词性标注最好结果相比, F 值提高了2.83 个百分点, 说明一体化方法能有效减少

表 **9** 一体化分词标注评测结果

模板	分词精度			词性标注精度		
	P/%	R/%	F/%	P/%	R/%	F/%
2W+ 2	94.14	94.64	94.39	89.06	89.54	89.30
2W+ 2+ C1	94.24	94.90	94.57	89.26	89.89	89.57
2W+ 2+ C1′	94.28	94.92	94.60	89.28	89.90	89.59
2W+ 2+ C1′2	94.26	94.95	94.60	89.18	89.83	89.50
2W+ 2+ C1′4	94.26	94.89	94.57	89.35	89.95	89.65
2W+ 2+ C1′24	94.23	94.91	94.57	89.19	89.83	89.51

“两步走”方法分词错误导致的扩散。

(3) 字符分类依然是有效特征,增加声、调特征性能并不稳定。由于测试语料的标准切分单位总数是固定的,从召回率上考虑,分词最佳模板为“2W+ 2+ C1′2”,词性标注最佳模板为“2W+ 2+ C1′4”;从综合性能上考虑,“2W+ 2+ C1′”是比较稳定的方式,研究者可以根据侧重点的不同进行取舍。当然,更好的特征模板仍然是我们进一步寻找的目标。

从平均时间消耗(T_{ave})上来看,一体化方法在时间开销上,确实比较大。本文实验采用的硬件配置为 Intel 四核处理器,4G 内存。3.2 节分词实验 T_{ave} 为 326 秒;3.3 节词性标注实验 T_{ave} 为6 732 秒,约 1.87 小时;3.4 节一体化方法 T_{ave} 为98 945 秒,约 27.48 小时。虽然分词标注一体化方法性能优于两步方法,但由于分类的类别数量大,时间消耗也大了很多。

总的来说,一体化方法不仅提高了分词精度,词性标注效果也有了明显提升。由于先秦语料库的建设,往往是人工标校出一部分语料作为训练数据,使用一体化方法来标注,可以满足实际需要。而在训练时间的开销方面,问题并不是很大,因为 20 多种先秦文本的规模总共只有 200 多万字,训练语料的数量更是有限的。

3.5 分词和标注错误分析

本节对一体化最佳标注结果的分词和词性标注错误类型做了分类统计(见表 10)。在分词错误中,未登录词和分词标准问题导致的错误占到77.97%。测试语料中未出现于训练语料的未登录词(OOV)共1 817 个,OOV 率为 8.75%,切分个数为1 693 个,正确个数为 1 214 个,准确率为 71.70%,召回率为 66.81%,F 值为 69.17%。在错误的 603 个未登录词中,多字词占 97.18%。同时,多字词的错误率占全部错误总数的 70.87%。可见,多字词是古汉语信息处理的难点。分词标准问题是指,机器自动切分的结果,分与合在意义上是两可的,只是与人工标注不同。切分歧义中,交集型歧义很少,组合型歧义居多。我们采用全切分算法统计了测试语料中的交集型歧义字段,总计只有 84 个段型和 128 个段例,其中错误的仅为 9 例。组合型歧义错误,则多是将二字词误切为两个单字,这主要是这些字在训练语料中多为单字词。人工标注错误而机器标注正确的词也有部分存在,这也可以看到自动标注具有一定

表 **10** 分词标注错误统计

错误类型		错误比例/%	示 例
分词	未登录词	53.55	“國人 施 公孫有山氏”(错标为“公孫 有 山氏”)
	分词标准问题	24.42	“陵虐 小 国”(陵虐,欺压凌辱,错标为“陵 虐”)
	切分歧义	5.86	“有 不 腆 先人 之 產馬”(產馬,本地所产之马,错标为“產 馬”)
	人工标注错误	4.26	“公斂處父”(人工错标为“公 斂 處父”)
	其他	11.91	“縱 子 忘 之”(错标为“縱子 忘 之”)
词性标注	名词动词混淆	22.91	v→n;n→v
	人名错识	20.28	nr→n;nr→v;nr→ns
	地名错识	14.41	ns→n;ns→nr;ns→v
	错识为人、地名	10.89	n→nr;v→nr;m→nr;n→ns
	错识为动词	10.56	d→v;p→v;c→v;a→v;m→v;t→v;f→v
	错识为名词	4.27	r→n;a→n;d→n;f→n
	其他	16.68	u→r;d→c;j→y;c→p;p→c;r→u;v→c;v→d;n→r;c→d;r→c;v→a;v→p;d→a

的自动纠错能力。

在词性标注错误中, n、ns、nr 三个词类之间混淆错标的占全部标注错误的 37. 04%, 这源于《左传》中的姓氏多取自爵位、职官、封邑等, 造成识别困难。由 n、v 混淆错标以及错标为 n 或 v 的共占 43. 89%, 其中"v → n" 占 13. 43%, "n → v" 占 9. 48%。这是由于古汉语词的兼类和活用现象比较频繁, 造成词类消歧困难。

4　结论及未来工作

本文在古代汉语自然语言处理领域进行了新的探索。在《左传》传文上的一系列实验表明, 基于 CRF 的分词标注一体化方法可以用于古代汉语语料库建设。与两步方法相比, 分词、词性标注性能均有明显提高, 开放测试的 F 值分别达到了 94. 60% 和 89. 65%。该方法可以应用于先秦其他语料的自动标注工作, 有效降低人工标注的工作量, 加快语料库的建设。从《左传》得到的训练模型, 可以用于先秦语料中内容相近的语料的自动标注, 如《公羊传》、《谷梁传》和《吕氏春秋》等, 给我们的项目进展带来了巨大的效益。

我们下一步的工作主要是: (1) 考虑先秦语料中诗词、语录体、典章制度等与《左传》差异较大的文本的自动标注。采取"人工标注训练语料→机器学习自动标注→人工校对" 的方式, 完成先秦 25 种传世文献的切分标注和后期校对, 建立起先秦文献切分标注语料库。(2) 继续探索改善 CRF 标注性能的特征模板和方法, 如采用多分类器集成技术和迁移学习技术。(3) 进一步细化词类体系。本文分词标注遵循的是《先秦汉语分词标注规范基本集》, 仅仅给出了 21 个词类标记, 对各词类的内部子类没有细分, 今后要尝试对词类进一步扩展, 制定出《扩展集》, 将先秦汉语的语料库加工技术研究深入下去, 在此基础上进行词汇统计和知识检索的工作。

致谢　感谢硕士生于丽丽、汪青青、肖磊同学在语料标注校对方面所做的大量工作。

参考文献

[1] 尉迟治平. 计算机技术和汉语史研究[J]. 古汉语研究, 2000, 3: 56-60.

[2] 魏培泉, 黄居仁, 等. 建构一个以共时与历时语言研究为导向的历史语料库[J]. 中文计算语言学期刊, 1997, 2(1): 131-145.

[3] 邱冰. 基于中文信息处理的古代汉语分词研究[J]. 微计算机信息, 2008, 1: 100-102.

[4] 白拴虎. 汉语词切分及词性标注一体化方法[C]//计算语言学进展与应用. 北京: 清华大学出版社, 1995: 56-61.

[5] Hwee Tou Ng and Jin Kiat Low. Chinese Part-of-Speech Tagging: One-at-a-Time or All-at-Once? Word-Based or Character-Based? [C]// Proceedings of ACL-04: 277-284.

[6] Yue Zhang and Stephen Clark. Joint Word Segmentation and POS Tagging using a Single Perceptron[C]// Proceedings of ACL-08: 888-896.

[7] 杨伯峻. 春秋左传注(修订版)[M]. 北京: 中华书局, 1990.

[8] 陈克炯. 春秋左传详解词典[M]. 河南: 中州古籍出版社, 2004.

[9] 宋彦, 等. 一种基于字词联合解码的中文分词方法[J]. 软件学报, 2009, 9: 2366-2375.