基于 CRF 的古汉语分词标注一体化研究*

石民　陈小荷　于丽丽　李斌

南京师范大学语言科学及技术系，南京，210097

E-mail:sdfcsm@yahoo.com.cn

摘要： 本文在计算机自然语言处理和古代汉语、特别是先秦文献的交叉领域进行了新的探索。首先对《左传》文本进行了词汇处理（分词和词性标注）和分析，然后采用条件随机场模型（CRF），基于两个模板进行自动分词、词性标注、分词标注一体化的对比实验。研究表明，一体化分词方法比单独分词的准确率和召回率均有明显提高，开放测试的最高 F 值达到了 90.89%，满足古代汉语词汇研究和语料库建设的需求，而且较好地弥补了人工标注的不足。

关键词： 古汉语，分词，词性标注，左传，条件随机场模型(CRF)

The Integrated Research about Ancient Chinese Word Segmentation and POS Tagging Based on CRF

SHI Min , CHEN Xiaohe , YU Lili，LI Bin

Department of Linguistic Science and Technology, Nanjing Normal University, Nanjing 210097

E-mail:sdfcsm@yahoo.com.cn

Abstract: In this paper,we made a new exploration in the cross field between NLP and Ancient Chinese,in particular of the pre-Qin literature. We processed and analyzed the vocabularies of "Zuo Zhuan" text,then used the conditional random model (CRF) to do automatic WS,POS tagging,joint WS and POS tagging experiments based on two different templates. This research has shown that the P and R value of the integrated approach have markedly improved than the independent WS,and the highest F value has reached to 90.89%. This method meets the requirements of the study of Ancient Chinese vocabulary and corpus construction,and it can make up for the shortcomings of manual tagging .

Keywords: Ancient Chinese, Word Segmentation, POS Tagging, Zuo Zhuan, Conditional Random Model(CRF).

1　前言

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进一步探索。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。加工手段上，依然只能采用传统的逐字逐句的人工标注方法，耗时耗力，且一致性较差。

我们正在做一个项目“先秦汉语词汇统计与知识检索”，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。而要实现这一目标，古汉语的切分标注是

基金项目： 南京师大 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“先秦文献词汇统计研究”。

作者简介： 石民（1984—），男，硕士生，研究方向为计算语言学；陈小荷（1952—），男，教授，博士生导师，研究方向为计算语言学；于丽丽（1983—），女，硕士生，研究方向为计算语言学。

古汉语语料库建设必需的基础性工作。虽然先秦汉语以单字词为主，但多字词也是不可忽视的，且随着历史的演进，多字词逐渐增多，汉语在历史的不同阶段呈现出不同的面貌。李斌（2007）曾提出面向中文陌生文本的人机交互式分词方法，在没有分词底表和训练语料的条件下，由系统自动地发现未登录词，提交给用户进行增删，不断重复此过程以得到领域词表，最后进行最大匹配法分词，这一思路在面向不同朝代古汉语的分词研究中具有重要的参考价值。针对古汉语的自动分词，邱冰（2008）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，通过这种措施进行古汉语分词。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。

对于汉语的词性标注及句法分析，通常的做法是把自动分词作为语料加工的第一道工序，然后进行词性标注等后续加工。这种两步走的方法，难以避免分词错误导致的词性标注正确率下降。白拴虎（1995）提出分词跟词性标注结合起来的一体化方法；Yue Zhang and Stephen Clark 在 ACL-08：HLT 联合学术会议上提出使用单一感知器的联合分词和词性标注方法，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。

本文着力研究面向古汉语文献的分词和词性标注，参照前人成果，设计了基于两个不同特征模板的自动分词、词性标注、分词标注一体化的对比实验，以此检验一体化方法的性能。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放出来，而仅需要对机器处理结果进行人工校对，一致性较高。正如古汉语计算语言学家尉迟治平(2000)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。”

2 古汉语分词标注规范

在具体研究分词之前，首先碰到的问题就是怎样确定古汉语的分词单位及词类。陈克炯在《春秋左传详解词典》中把古汉语词类划分为 14 类，我们制定面向计算机信息处理的古汉语分词规范，必须把握好词类区分的颗粒度。参考北京大学《古代汉语知识教程》，同时采用词汇意义和语法功能兼顾的标准，我们确定出适合古汉语的分词单位及词类，可以查阅南京师范大学 CIPP 中文信息处理平台网站[①]。我们目前还只是在划分“主要的”词类方面做了一些工作，划分词的子类则必须考虑词汇意义的标准，例如把副词分为范围副词、时间副词、情态副词等小类，它们的语法功能几乎没有什么差别，为了便于前期语料标注，对其内部所属子类不做区分。后期可以从意义的角度再细分成几个小类，进行词汇统计的细化研究。

根据《春秋左传》语料考察结果及其词类分布状况，本文确定了古汉语主要词类的标注集：

表 1　　古汉语词类标注集

序号	名称		标记	解释
1	名词	普通名词	n	noun 首字母
		人名	nr	noun 首字母+人（ren）首字母
		地名	ns	noun 首字母+space 首字母
		方位词	f	“方”的声母

[①] 《古代汉语分词标注规范》http://www.cipp.cn/news_view.asp?id=76。

	时间词	t	time 首字母
2	动词	v	verb 首字母
	使动用法	sv	“使”的声母+verb 首字母
	意动用法	yv	“意”的声母+verb 首字母
	为动用法	wv	“为”的声母+verb 首字母
3	形容词	a	adjective 首字母
4	数词	m	number 第 2 个字母
5	量词	q	quatity 首字母
6	代词	r	pronoun 的第 2 个字母
7	介词	p	prepositional 的首字母
8	连词	c	conjunction 的首字母
9	助词	u	auxiliary 的第 2 个字母
10	副词	d	adverb 的第 2 个字母
11	语气词	y	“语”的声母
12	拟声词	s	sound 的第 1 个字母
13	兼词	j	“兼”的声母
14	标点	w	参考北大现代汉语标记集

3 算法描述

3.1 模型简介

条件随机场（CRF）是一个无向图的判别模型，它能够被用来定义在给定一组需要标记的观察序列的条件下，一个标签序列的联合概率分布。对于一组长设为 n 的观察序列 $X=X_1X_2X_3...X_n$(要标记的汉字序列或词序列)，输出为 $Y=Y_1Y_2Y_3...Y_n$(相应的标注序列)。这样就把分词问题（或词性标注）转化为相应的序列标注问题。同时，CRF 模型允许增加复杂特征，可以有效地处理标记偏置问题。本文相关实验采用 Taku Kudo 开发的“CRF++0.51”工具包进行训练和测试，下载地址为：http://crfpp.sourceforge.net/。

3.2 实验数据

本文使用香港中文大学建立的汉达文库《春秋左传》传文作为实验语料，约 20 万字。“先秦汉语词汇统计”课题组将该语料切分标注，数据存储格式为纯文本, 采用 UNICODE 编码。由于分工标注，尽管严格按照分词规范，难免存在一些意见分歧或标注失误，在一定程度上可能会影响实验结果，语料还要进一步完善。实验过程，以前十卷作为训练语料，后两卷（定公卷、哀公卷）作为测试语料，测试语料与训练语料比例为 16%。

表 2　　训练语料样例

自动分词			词性标注			分词标注一体化	
第一列	第二列		第一列	第二列		第一列	第二列
晉	S		正月	t		晉	S-ns

魏	B		辛巳	t		魏	B-nr
舒	O		，	w		舒	O-nr
合	S		晋	ns		合	S-v
諸	B		魏舒	nr		諸	B-n
侯	O		合	v		侯	O-n
之	S		諸侯	n		之	S-u
大	B		之	u		大	B-n
夫	O		大夫	n		夫	O-n
于	S		于	p		于	S-p
狄	B-		狄泉	ns		狄	B-ns
泉	O		。	w		泉	O-ns
。	w		...	...		。	S-w

3.3 特征模板

特征是基于 CRF 自动识别的核心，特征选择的好坏将影响 CRF 模型识别的性能。但特征选择越多，模型的搜索数据量越大，对机器性能是个严峻的考验。本文结合古代汉语单音节词占优势的特点，采用如下两个特征模板进行了各项实验的模板横向效果对比，以及一体化分词与单独分词、一体化标注与单独标注的纵向效果对比。

表3　　特征模板

	特征
模板一(Template1)	$W_{i-2},W_{i-1},W_i,W_{i+1},W_{i+2}$
模板二(Template2)	W_{i-1},W_i,W_{i+1}, W_{i-1}/W_i， W_i/W_{i+1}

注意：CRF分词时采用由字构词原理，这里的W为单字，词性标注时为词；下标表示所考察的特征位置，例如i表示当前位置、i-1表示左边第一个位置、i+1表示右边第一个位置；模板一考察一元特征，窗口为5；模板二增加了二元共现特征，CRF搜索特征量也将随之激增。

4 实验结果及分析

4.1 基于 CRF 的分词实验

本实验采用由字构词原理，参考现代汉语六词位分词相关文献，设计了适合古代汉语的四词位分词方法，词位标记{B、I、O、S}，因为古代汉语单音节词占优势，标记过多会造成词位冗余，利用率低。这里 S 代表单字词或标点，B 代表词首第一个字，O 代表词尾最末字，I 代表一个词中间的所有字，语料样例见表 2。评测结果如下表所示：

表 4　　CRF 分词评测结果

分词	模板	测试总词数	CRF 切分	切分正确	P(%)	R(%)	F(%)
Close	Template1	166746	167674	158627	94.60	95.13	94.86
	Template2	166746	167008	165073	98.83	98.98	98.90
Open	Template1	27429	28060	24394	86.94	88.94	87.93

	Template2	27429	28127	25049	89.06	91.32	90.18

从表 4 可以看出，CRF 的封闭实验效果非常好，开放测试的 F 值略低。且增加了二字共现特征的模板二的两次测试的结果均高于模板一：封闭测试 F 值提高了 4.04%，开放测试 F 值提高了 2.25%。可见增加一定量的复杂特征能显著提高 CRF 的性能。开放测试 F 值达到 90%以上，基本满足语料库自动分词的需求。

4.2 基于 CRF 的词性标注实验

词性标注是 CRF 模型的典型应用，这里不再详述。评测结果如下表所示：

表 5　　CRF 词性标注评测结果

词性标注	模板	测试总数	标注总数	标注正确	P(%)=R(%)
Close	Template1	166749	166749	160039	95.98
	Template2	166749	166749	162543	97.48
Open	Template1	27429	27429	24838	90.55
	Template2	27429	27429	24877	90.70

词性标注的结果与分词结果的规律相吻合：模板二的测试结果优于模板一，封闭测试 F 值提高了 1.5%，开放测试提高了 0.15%。

4.3 基于 CRF 的分词标注一体化实验

4.3.1 语料预处理

CRF 分词我们考虑的是由字构词的汉字词位信息，在此基础上，我们把词的词性标记也赋予该汉字，这样汉字就承载了双重信息，即该字所属词的词性以及该字在词中的词位信息(B,I,O,S)。例如：諸侯/n,“諸”为词首 B，“侯”为词尾 O，则标注格式为“諸/B-n　侯/O-n”，预处理过程可用程序自动完成。 CRF 模型进行序列化标注时，规定格式为每行一个词例，句间用空行隔开，因此语料第一列为单字，第二列为复合标记信息，语料样例见表 2。

4.3.2 实验结果

为了便于与前两个实验对比，本文对一体化标注结果分别进行了分词和标注的评测。分词评测参照单独分词的评测方法，而一体化标注评测需要在正确的分词结果上考虑标注是否正确，本文给出了详细数据。评测结果如表 6、表 7 所示：

表 6　　一体化分词评测结果

一体化分词	模板	测试总词数	CRF 切分	切分正确	P(%)	R(%)	F(%)
Close	Template1	166746	166634	162143	97.30	97.24	97.27
	Template2	166746	166723	164874	98.89	98.88	98.88
Open	Template1	27429	27747	24963	89.97	91.01	90.49
	Template2	27429	27759	25080	90.35	91.44	90.89

表 7　　一体化词性标注评测结果

一体化标注	模板	测试总数	标注总数	标注正确	P(%)	R(%)	F(%)
Close	Template1	166746	166634	161742	97.06	97.00	97.03
	Template2	166746	166723	163958	98.34	98.33	98.33

Open	Template1	27429	27747	24901	89.74	90.78	90.26
	Template2	27429	27759	25053	90.25	91.34	90.50

4.3.3 实验分析

从三个实验的内部横向对比来看，增加了二字（词）共现特征的模板二的两次测试的结果均高于模板一。可见，增加适当的语境信息能显著提高 CRF 的识别性能，这也与古代汉语单音节词占优势的特点有关。

分词标注一体化实验的分词结果优于单独分词实验，但训练数据量过大，效果提高不是太大，最高 F 值仅比单独分词最好结果提高 0.71%。且对机器性能要求较高，下一步应着力进行模型优化，解决机器的学习速度问题。

一体化实验的词性标注结果封闭测试 F 值高于单独标注结果，开放测试 F 值反而比单独标注低。开放测试的极值（即假设一体化实验中正确切分的单位，其相应标注也全部正确），模板一 P 为 89.97%，R 为 91.01%，F 为 90.49%，比单独标注的 F 值低；模板二 P 为 90.36%，R 为 91.44%，F 为 90.90%，比单独标注略高（0.20%）。

分词标注一体化方法，将汉字的词位信息和所属词的词性信息结合起来，能更有效的提高分词精度；但词性标注效果不太明显，分词的精度也限制了标注精度的提高。

5 结论及未来工作

本文在自然语言处理和古代汉语、特别是先秦文献的交叉领域进行了新的探索，研究表明，计算机分词标注一体化方法可以用于古代汉语语料库建设，与单独分词相比，准确率和召回率均有明显提高。分词开放测试的最高 F 值达到了 90.89%，而且较好地弥补了人工分词方法的缺陷和不足，保证了语料加工质量，具有良好的经济效益和科研效益。

下一步工作是继续探索改善 CRF 性能的特征模板和方法，以便将 CRF 模型集成到面向不同朝代的古汉语词汇处理专门系统中去，应用此方法，完成先秦 25 种传世文献的切分标注和后期校对，初步建立起先秦文献切分标注语料库，来服务于语言学工作者的研究工作。

参 考 文 献

[1] 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.

[2] 陈克炯. 春秋左传详解词典(第 1 版)[M]. 河南:中州古籍出版社, 2004.

[3] 陈小荷. 现代汉语自动分析——Visual C++实现[M]. 北京：北京语言文化大学出版社, 2000.

[4] 国家技术监督局、中华人民共和国国家标准厅. 信息处理用现代汉语分词规范[M]. 北京：中国标准出版社, 1993.

[5] 李斌. 面向中文陌生文本的人机交互式分词方法[J]. 中文信息学报，2007 年第 3 期.

[6] 刘开瑛. 中文文本自动分词和标注[M]. 北京：商务印书馆, 2000.

[7] 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 年第 1 期.

[8] 尉迟治平. 计算机技术和汉语史研究[J]. 古汉语研究，2000 年第 3 期.

[9] Yue Zhang and Stephen Clark. Joint Word Segmentation and POS Tagging using a Single Perceptron[C]// Proceedings of ACL-08：HLT. 2008 ： 888-896.

[10] 张双棣等. 古代汉语知识教程[M]. 北京：北京大学出版社, 2002.