一种利用注疏的《左传》分词新方法*

徐润华，陈小荷
南京师范大学 语言信息科技研究中心，南京 210097
E-mail: runhuaxu@163.com

摘　要：先秦文献的注疏文献中包含有大量词汇语义知识，是先秦文献自动分词的重要依据。本文以篇幅最大的先秦文献《左传》为研究对象，在对《左传》及其注疏文献进行自动对齐的基础上，提出了一种利用注疏的《左传》分词新方法。分词实验的F值达到89.0%，较之baseline有明显提升。该方法无需训练语料，利用注疏文献辅助分词的思想也适合推广到其他先秦文献的自动分词任务中去。

关键词：先秦文献；注疏文献；自动对齐；自动分词

A New Method of Segmentation on "Zuo Zhuan" by Using Commentaries

Xu Run-hua[1], Chen Xiao-he[2]
Research Center of Language and Informatics, Nanjing Normal University, Nanjing 210097
E-mail: runhuaxu@163.com

Abstract: Commentaries of Pre-Qin documents contains a large vocabulary semantic knowledge which can be an important basis for segmentation. This paper uses "Zuo Zhuan" as the research object, proposes a new segmentation method based on commentaries which has been already aligned to "Zuo Zhuan". Segmentation experiments F-score reached 89.0%, much higher than the baseline. This method needs no training, and the idea of commentaries assisted segmentation is also available to the segmentation of other pre-Qin documents.

Keywords: Pre-Qin documents; commentaries documents; automatic alignment; automatic segmentation

1　引言

先秦文献专指秦朝统一之前、诞生于春秋战国时代的一大批优秀文学作品，它们形式上丰富多彩、内容上斑驳灿烂，奠定了我国两千多年文学发展历史的坚实基础。目前已知的先秦传世文献不过数十种，而其中又以《左传》的篇幅为最大。随着时代的发展，古籍文献的数字化、语料化的需求越来越大，索引化、结构化的应用越来越广，这也使得对先秦文献进行信息处理方面的研究具有了更加积极的意义。

然而现阶段，针对先秦文献在信息处理方面所做的研究还比较匮乏，并且多停留在使用现代汉语信息处理方法来处理古汉语的模式上，缺乏对先秦文献体裁、古汉语语言风格等特殊之处的有针对性处理。正是在这种背景下，本文以先秦传世文献中篇幅最大的《左传》为研究对象，讨论了一种从先秦文献本身特点出发、充分考虑古汉语信息处理特殊性的《左传》分词新方法，并希望籍此能给整个先秦文献的信息处理研究带来有益的启示和帮助。

2　《左传》自动分词的特点

现代汉语自动分词的一般模式是：人工标注—模型训练—机器标注。人工标注的工作量通常很大，模型训练需要大规模语料支持，机器标注的最终效果取决于标注语料与训练语料的相似程

*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"；江苏高校哲学社会科学重点研究基地重大项目"先秦文献词汇知识挖掘"（2010JDXM023）。

度。先秦文献却并不适用于现代汉语的这套分词模式：先秦传世文献只有数十种，每种文献的篇幅都不大，篇幅最大的《左传》也只有28万字，多数文献只是几万字的篇幅，有的甚至只有几千字。现有的常见统计模型所需要的参数规模往往都很大，即使把某篇先秦文献全部都用作于训练，训练语料的规模也显得不足。

此外，从语料处理角度看，先秦文献的语料大多是封闭的，各篇文献之间的差异比较大，语料的同质性很低。文献间的差异主要体现在：时代差异、学派差异、题材差异和体裁差异等。这些差异意味着，从《左传》中训练得到的分词统计数据，并不能适用于其他文献的自动分词任务，反之亦然。

石民（2009）曾经提出了一种基于CRF模型的《左传》分词方法，该方法的分词精度可以达到93%左右。但这是在训练、测试语料按10:1比例分配的基础上得到的结果：一方面，人工必须事先标注好大部分的《左传》语料，这就降低了自动分词效果的实用性；另一方面，训练得到的分词统计数据并不能很好的适用于其他先秦文献，这就使得自动分词方法的通用性也大打折扣。

《左传》语料中单音节词居多，不适宜使用现代汉语通用的分词词表。多字词大多是专名，专名的结构和语境与现代汉语差异很大，加上词类活用、繁简字、通假字、异体字、文献传抄讹误等因素，都给《左传》的自动分词增加了困难。由于单音节词居多，即使全部切为单字词，《左传》的分词精度也能达到80%左右。《左传》自动分词的基线较高，分词精度的提升空间小。

3 利用注疏文献辅助分词的思想

先秦文献由于年代久远，语言生涩，故后人为其注释，谓之“注”；由于“注”仍然存在语言难懂、解释不全的问题，后人为“注”进行注释，谓之“疏”。注疏文献的信息十分丰富，而且往往有不同时期的叠加：一部《论语》有《论语注疏》、《论语笔解》、《论语集注》、《论语全解》四部不同时期的注疏文献；《左传》的注疏文献更是经历了“经(春秋) ⇒ 传(公羊传) ⇒ 注疏 (公羊传注疏)”这样的演变过程。

先秦文献的注疏中，包含有大量的半结构化词汇、语义知识，可以为先秦文献的自动分词提供重要帮助。例如，根据《春秋左传正义》，可以对下面五个《左传》句子里的相关词语做出正确的切分：

《左传》	《春秋左传正义》
1．秋，大/雨/雹	雨，于付反。
2．/邾子克/也	克，仪父名。
3．/六/人叛楚	六国，今庐江六县。
4．鲁有名而無/情/	有大国名，无情实。
5．冬，来，/反馬/也	反其所留之马。

陈小荷（2009）指出，注疏犹如现今语文教学中的“串讲”，是对先秦文献进行自动分词和标注的重要依据。他认为，语言信息处理需要启动知识。现代汉语信息处理的一般模式是用训练语料作为启动知识（有监督的学习）。先秦文献信息处理则应将相关文献（即注疏文献）作为启动知识，因为处理先秦文献所需的知识已经存在于相关文献之中。而来自于相关文献的证据要比统计模型更可靠和好用。

4 《左传》与其注疏文献的自动对齐

注疏文献中虽然包含了大量的词汇语义知识，但它尚未和原文建立起对应关系；而自动对齐则正是要找到注疏和原文之间的这种关联并将其形式化，将半结构化的注疏文献结构化，为自动分词乃至其他信息处理任务提供更为可靠和有效的帮助。

4.1 《左传》注疏文献的基本格式

注疏文献是一种半结构化的文献，其内部构成方式呈现出明显的规律性，《左传》注疏文献也不例外。本文研究所选用的《左传》注疏文献为《春秋左传正义》，以下是其部分内容示例：

【傳】元年，春，王周正月。言周以别夏殷。○别，彼列反。夏，戶雅反。不書即位，攝也。假攝君政。不脩即位之禮，故史不書於策，傳所以見異於常。

【疏】"不書即位，攝也"。○正義曰：攝訓持也。隱以桓公幼少，且攝持國政，待其年長，所以不行即位之禮。史官不書即位，仲尼因而不改，故發傳以解之。

上例中，"元年，春，王周正月"、" 不書即位，攝也" 都是援引自《左传》原文的引文，引文后面的内容是对该引文所做的注释。《春秋左传正义》基本由"傳"和"疏"构成，"傳"和"疏"均以段落为界，每段文字由引文和注释构成，引文常常间断为若干小句。

4.2 《左传》及其注疏文献的对齐任务

自动对齐的最终目的，是要找到原文在注疏中的引文、注疏对引文所作的解释以及该解释中所出现的原文词汇。因此，《左传》及其注疏文献的对齐任务可以细化为句子对齐、注释对齐、词汇对齐这三个子任务：

原文：孟子卒，繼室以聲子，生隱公。

注疏：繼室以聲子，生隱公。聲，谥也。蓋孟子之侄娣也。諸侯始娶，則同姓之國以侄娣媵。元妃死，則次妃攝治内事，猶不得稱夫人，故謂之繼室。

句子对齐：繼室以聲子，生隱公。 ⟺ 繼室以聲子，生隱公。

注释对齐：繼室以聲子，生隱公。 ⟺ 聲，谥也。蓋孟子之侄娣也。諸侯始娶，則同姓之國以侄娣媵。元妃死，則次妃攝治内事，猶不得稱夫人，故謂之繼室。

词汇对齐：

聲 ⟺ 聲，谥也。

繼室 ⟺ 故謂之繼室。

图1 对齐任务示例

注释对齐是词汇对齐的基础；句子对齐是注释对齐的基础。因此，句子对齐是自动对齐的核心任务。

4.3 《左传》及其注疏文献自动对齐的实验结果

《左传》及其注疏文献的自动对齐采用顺序有关的对齐技术，在对原文及其引文进行相似度比较的同时，利用局部回溯机制增强相似度比较结果的可信度，利用全局回溯机制减少错位匹配发生的机率和错位匹配的长度。通过对《左传》及其注疏文献《春秋左传正义》进行自动对齐实验，得到如下实验结果：《左传》全文共 37588 个小句，其中对齐成功 36917 个，占全部小句数的 98.2%。考虑到少量由于错位匹配而造成的错误对齐结果，最终的自动对齐正确率略低于 98%。

5 利用注疏的《左传》分词新方法

5.1 分词算法的基本框架

利用注疏的《左传》自动分词，就是要根据注疏文献来生成注疏词表，然后以注疏词表为基础来进行自动分词，这是一种规则驱动、词表驱动的分词方法。因此，本文引入最大匹配分词算法作为《左传》自动分词的算法主框架：每次从待分词文本中取长度等于最大词长的字串，该字串逐字递减直至在注疏词表中查找成功为止，否则将该字串最后剩下的单字作为单字词输出。

5.2 分词算法的核心内容

5.2.1 注疏词表的生成

分词算法以词表驱动，因此，如何利用《左传》的注疏文献生成高质量的注疏词表就成为了分词算法的核心问题。基本的思路是：在注疏文献中查找原文所有可能成词的字串，将找到的那些字串添加进注疏词表中。这实质上是一种子串查找和匹配的过程，其目的是找到注或疏中出现过的《左传》原文词语。本文分词算法使用了两种串匹配方法：

宽匹配——如果原文中的字串出现在注或疏中，并且该字串没有包含在其他原文字串内，则匹配成功。

严匹配——如果原文中的字串出现在注或疏中，并且该字串的前后位置有显式分隔标记（标点、提示词等），则匹配成功。

除了串匹配方法，查找方式和查找范围也都是影响注疏词表生成的因素。按照是否利用自动对齐的信息，查找方式分为两种：

利用自动对齐——对原文中的每个字串，都到和当前原文对齐成功的注或疏部分去查找。

不利用自动对齐——对原文中的每个字串，都到全部的注和疏部分去查找。

此外，由于注是对原文直接的注释，给自动分词提供的依据可信度更大；疏是对注的注释，和原文不相关的信息比较多，干扰较大，但仍不能就此完全排除疏对于自动分词的启发作用。本文分词算法的查找范围也分为两种：一、只在注部分进行查找；二、在注和疏部分都进行查找。

综合考虑“串匹配方式”、“是否利用自动对齐”、“查找范围”这三个因素，可以分别生成 8 个注疏词表。注疏词表中的条目由“词语”、“词语所在原文小句”、“频率”这三个数据项构成。

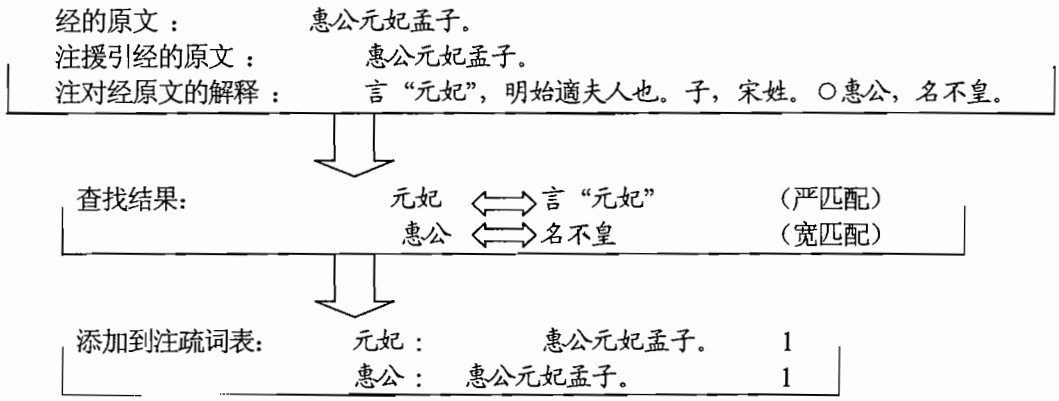

图2 生成注疏词表示例

5.2.2 局部特征和全局特征

在利用注疏文献的自动分词过程中，局部特征指的是原文中的某个字串，恰好可以在注疏对其引文进行注释的内容中找到成词证据。即，设当前待分词的字串为 A，所在小句为 B，若在注疏词表中查到 A，并且其所在原文的小句恰好就是 B，那么字串 A 查找成功。

但是，原文中词语的每一次出现，注疏并不会都给出解释。很多时候，需要把词语出现在其他地方的解释传递到当前的这次出现中。如果能够在注疏对其他引文进行注释的内容中找到成词证据，那么依然可以把当前原文的字串看成是一个词语。这就是全局特征的运用。

5.2.3 分词算法模型

不同的注疏词表生成方式也就形成了不同的查找词表方式，而查找词表的过程是分词算法中最重要的一环，查找是否成功直接决定当前字串是否被切分成词语。如何给出《左传》中每个字串查找词表成功与否的判断，是分词算法的关键。加上局部、全局特征的运用，整个分词算法共

受到四个方面因素的制约："自动对齐"、"查找范围"、"匹配方式"、"特征运用"。根据这四个制约因素，本文构建了一个利用注疏对《左传》进行自动分词的算法模型，如下：

模型公式：$\mathbf{Res = S_i \times R_j \times M_k \times F_n}$

说明：S、R、M、F 都是长度为 2 的数组，i、j、k、n 是数组下标。数组元素的值为 0 或 1，Res 的值为 1 表示查找词表成功，值为 0 表示查找失败。S_i 表示是否"利用自动对齐"，R_j 表示是否"利用注和疏"，M_k 表示是否"采用严匹配"，F_n 表示是否"利用局部特征"。

5.3 模型参数的优选

根据 i，j，k，n 取值的不同，算法模型共有 12 种参数组合方式（四个参数组合应该有 16 种情况，但由于在不利用自动对齐的时候，查找过程无法使用局部特征，故少了四种），每组参数形成一个子算法，共 12 个子算法，如下表：

算法序号	1	2	3	4	5	6	7	8	9	10	11	12
范围	注	注	注	注	注疏	注疏	注疏	注疏	注	注	注疏	注疏
匹配	宽	宽	严	严	宽	宽	严	严	宽	严	宽	严
特征	局部	全局	局部	全局	局部	全局	局部	全局	全局	全局	全局	全局
对齐	是	是	是	是	是	是	是	是	否	否	否	否

本文选取 1/10 的《左传》语料来对这 12 个子算法进行优选，由于《左传》语料以单音节词为主，故以全部切分为单字词的分词精度作为《左传》自动分词的 baseline 并与子算法的分词精度进行比较：

算法序号	1	2	3	4	5	6	7	8	9	10	11	12	Base line
正确率(%)	79.2	**86.1**	78.9	85.8	79.9	86.1	79.4	85.3	83.2	84.2	83.3	85.0	73.0
召回率(%)	88.4	**90.1**	88.2	89.6	87.6	87.9	86.5	85.4	85.8	86.7	81.0	84.2	86.6
F值(%)	83.6	**88.1**	83.3	87.6	83.6	87.0	82.8	85.4	84.5	85.5	82.1	84.6	79.2

分词结果表明，子算法 2 的分词效果最优。12 种子算法的分词精度全部都高于 baseline，这说明引入注疏来辅助分词是可行的，对分词精度的提高确有帮助。同时，利用自动对齐的子算法 2 的分词效果最优，也进一步验证了自动对齐对于分词工作的重要意义。

5.4 《左传》全文的自动分词实验

在模型参数优选的基础上，利用注疏文献的《左传》自动分词实验选取《左传》全文和《春秋左传正义》全文作为实验语料，采用效果最优的子算法 2 作为分词算法的核心成分，并将《左传》全部切分为单字词的分词精度作为实验的 baseline。实验得到的分词结果与人工切分的《左传》分词语料进行比对，最终的实验数据如下表：

	实有词语数	全部分词数	正确分词数	正确率	召回率	F值
子算法2	195139	203675	177175	87.0%	90.8%	**89.0%**
Baseline	195139	229487	172444	75.1%	88.4%	**81.2%**

利用注疏的《左传》分词方法的 F 值可以达到 89%，相对于 baseline 有近 10% 的 F 值提升，这说明利用注疏文献来对《左传》进行自动分词的做法是可行的，并且能够收到比较好的效果；子算法 2 的分词 F 值接近 90%，已经具备一定的实用性；虽然在数据上仍不及基于 CRF 模型的《左

传》分词效果，但较之统计模型对于训练数据的大量需求，注疏分词方法的优势也很明显：无需任何训练语料，整个分词过程中只用到了一本注疏文献而已。

6 结语

6.1 存在的问题

切分不一致问题。只靠查找词表无法解决组合型歧义，而组合型歧义在《左传》语料中并不少见。例："請 其 不足"、"也 不 足以 容 從者"。

低频词问题。有些词语有着明显的形式特征，如"郑公"，"梁伯"等，但由于出现次数过少，未收入词表。

人工标注问题。数词及部分量词究竟切开还是不切开，存在不一致。一些含虚词的结构，到底切不切开，存在不一致。如："在上"、"之後"、"不過"等。

6.2 方法的特色

统计模型的本意，是从较小规模的训练语料中学习模型参数，用学到的模型来自动处理较大规模的、与训练语料相似的其余语料。但这与《左传》语料、先秦文献语料规模较小、同质性低的特点相冲突。而本文所提出的利用注疏的分词方法的最大特色就在于无需人工事先标注，不需要任何训练语料；同时，利用相关文献来处理目标文献的分词思路具有通用性，完全可以移植到其他先秦文献的自动分词乃至各种信息处理的任务中去。

6.3 改进的方向

《左传》自动分词方法中所使用的8个注疏词表以及12个分词子算法目前只是被孤立地看待，可以尝试将其组合起来进行实验；尚未利用到词表中的频率信息，分词算法模型中应当引入更多的统计量来对查找结果进行筛选；查找词表过程中所使用的串匹配方法还很粗糙，需要相关句法知识的引入来形成约束规则，提高匹配结果的精度；《左传》的注疏文献不止一种，同一种注疏文献也存在着不同的版本，下一步的工作应将多本《左传》注疏综合起来以更加充分地挖掘其中的词汇语义知识。

参 考 文 献

[1] 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, 2(24): 39-45.

[2] 肖磊, 陈小荷. 古籍版本异文的自动发现[J]. 中文信息学报, 2010, 5(24): 50-55.

[3] 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.

[4] 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.

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

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

[7] 常娥, 侯汉清, 曹玲. 古籍自动校勘的研究和实现[J]. 中文信息学报, 2007, 21(2): 83-88.