DOI:10.16515/j.cnki.32-1745/c.2019.02.019

《左传》及其注疏文献的内容自动对齐研究

徐润华[1],梁社会[2]

(1.金陵科技学院人文学院,江苏 南京 210038;2.南京师范大学国际文化教育学院,江苏 南京 210097)

摘 要:自动对齐的目的是将半结构化的注疏文献转化为结构化形式,从而为先秦文献的信息处理提供可靠的帮助。探讨《左传》及其注疏文献的三种自动对齐方式:句子对齐、注释对齐和词汇对齐。在句子及注释对齐试验中,对齐的正确率为99.8%,召回率为98.2%,效果较为理想。提出两种词汇对齐的原则,并通过具体实例展示《左传》及其注疏文献的词汇对齐效果。

关键词:注疏文献;句子对齐;注释对齐;词汇对齐;内容自动对齐;《左传》

中图分类号:H141;TP391 **文献标识码**:A **文章编号**:1673-131X(2019)02-0084-05

Study on the Automatic Alignment of the Contents Between *Zuo Zhuan* and its Annotated Documents

XU Run-Hua[1], LIANG She-Hui[2]

(1. Jinling Institute of Technology, Nanjing 210038, China;
2. Nanjing Normal University, Nanjing 210097, China)

Abstract: The purpose of automatic alignment is to transform semi-structured annotated documents into a structured form, so as to provide reliable help for the information processing task of pre-Qin literature. Three kinds of automatic alignment processes between *Zuo Zhuan* and its annotated documents are discussed: sentence alignment, annotation alignment and lexicon alignment. In the test of sentence and annotation alignment, the alignment accuracy rate was 99.8%, and the recall rate was 98.2%, which works pretty well. Two kinds of lexicon alignment principles are put forward, and demonstrated the lexicon alignment effect of *Zuo Zhuan* and its annotated documents through a specific example.

Key words: annotated literature; sentence alignment; annotation alignment; lexicon alignment; aufomatic content alignment; *Zuo Zhuan*

先秦传世文献篇幅一般不长,即使篇幅最长的《左传》也仅有28万字,其余文献多数为几万字篇幅,有的甚至仅几千字。现代汉语的信息处理方法往往需要较大的参数规模和大量的训练语料,这与先秦文献篇幅短小的特点相冲突。因此对先秦文献进行信息处理需要探索新方法。

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,为此后人对"注"进行注释,谓之"疏"。先秦文献注疏中的信息十分丰富,包含大量的半结构化词汇和语义知识,是先秦文献信息处理的重要依据。注疏文献犹如现今语文教学中的"串

收稿日期:2019-04-26
基金项目:国家社会科学基金项目(15BYY096);江苏高校哲学社会科学研究基金项目(2018SJA0473)
作者简介:徐润华(1982-),男,江苏南京人,讲师,博士,主要从事自然语言处理和信息计量研究。

讲”，是对先秦文献进行自动分词和标注的重要依据[1]。对语言进行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），而先秦文献由于所需的知识已存于相关文献（即注疏文献）中，且这些文献的证据要比统计模型更可靠和好用[2]，因此，对其进行信息处理应将相关文献作为启动知识。例如，《左传》“六人叛楚”一句，根据《春秋左传正义》“六国，今庐江六县”的注疏可知，此句中的“六人”应被理解为六国之人，而不是六个人，藉此可以帮助计算机对该句做出正确的理解和词语切分。

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。而自动对齐正是要找到注疏文献和原文之间的这种关联并将其形式化，进而将半结构化的注疏文献结构化，从而为自动分词乃至其他先秦文献的信息处理提供更为可靠和有效的帮助。本文以先秦文献中篇幅最长的《左传》为研究对象，对《左传》及其注疏文献进行内容自动对齐研究。

一、《左传》及其注疏文献的自动对齐概述

（一）《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明显的规律性，《左传》注疏文献也不例外。本文选用的《左传》注疏文献为《春秋左传正义》，以下为部分内容示例：

【傳】元年，春，王周正月。言周以别夏殷。○别，彼列反。夏，户雅反。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。[3]

例中，“元年，春，王周正月”和“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。从示例中可以看出，注疏文献在行文结构上具有以下特点：基本上是由“对原文的援引”和“对引文的注解”两部分构成；“对引文的注解”分为“注”和“疏”两部分，“注”紧跟引文之后，“疏”则另起一段；一段引文及“注”的内容，加上一段“疏”的内容，构成注疏文献的最基本单位。

（二）《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文献中的引文、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、词汇对齐三个子任务[4]。三个子任务中，句子对齐最为重要。由于注疏本身就是半结构化的文献，因此句子对齐成功后，注释对齐自然也就完成了。而词汇对齐是基于注释对齐的一个子串匹配过程，所以必须以句子对齐和注释对齐的结果作为前提和依据（图 1）。

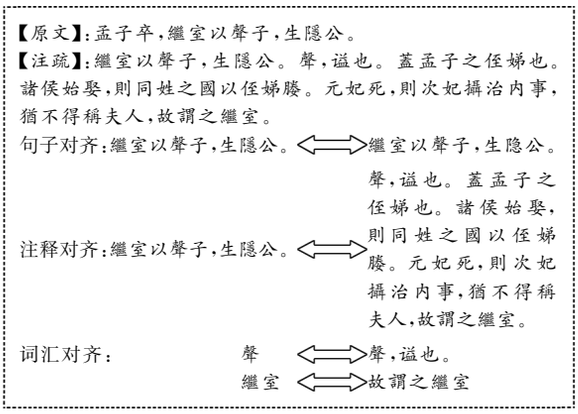

图 1　《左传》自动对齐任务示例

（三）《左传》及其注疏文献自动对齐的特殊性

《左传》及其注疏文献在进行自动对齐时存在以下特殊情况。

一是繁体字存在不同版本，同一个字原文和引文使用的版本不一致[5]，但此时并不存在引用错误的问题。

二是原文和引文的标点位置不一致，并且相同位置的标点也存在不一致的情况，但由于古籍文献的句读本身就是后人加上去的，所以此时也不存在引用错误的问题。

三是有时原文和引文相似程度非常高，但实际上该引文却不是此处原文的对齐结果；有时原文和引文的相似程度并不高甚至只有 50% 左右，但实际上该引文正是此处原文的对齐结果。因此，自动对齐算法的匹配功能，必须要能兼容各种复杂特殊的对齐情况，不能仅由一个相似度计算结果决定。

四是在面对大规模文献的对齐任务时，算法首先要能保证顺利跑完所有的注疏文献而不中途报错，其次要保证不出现因为某一次的匹配失败或者匹配错误而导致接下来所有内容的匹配失败或匹配错误的现象。

二、《左传》及其注疏文献的句子对齐和注释对齐

(一)顺序有关的自动对齐

顺序无关是指,原文的每个句子都到注疏文献库中查找可能的对齐结果。这样的对齐过程,其实相当于一个全文检索过程。每一个句子是否对齐成功,互相之间没有影响。顺序有关是指,按照顺序,原文中的每个句子都到注疏文献的相应部分中查找可能的对齐结果。只有“过去”的原文对齐成功,“现在”和“将来”的原文才可能对齐成功。两者比较,顺序无关算法健壮性更好,因为不能寄希望于注疏中的引文也完全按照原文中的先后顺序出现。但是顺序无关算法的正确性较差,因为和当前原文相似程度高的注疏中的句子可能有很多,但其中只有一句才是真正的引文。此外,由于原文和引文的标点位置存在不一致现象,如果采用次序无关的对齐算法,因为引文被断成了更多的小句,并散落在不同的段落里,许多原文可能因此找不到相关的引文。可见,顺序无关的对齐算法健壮性较好,而顺序有关的对齐算法正确性更优。考虑到《左传》的注疏文献大都是按照原文的先后顺序来援引相关内容的,因此本研究采用了顺序有关的对齐算法来对其进行自动对齐。

(二)局部回溯算法

句子与句子之间的相似度可以用浮点型数值表示,同时设优、良、中、差四个等级。一个好的对齐算法不能仅由一个相似度结果来确定,其必须要有很强的兼容性。在进行局部匹配时,算法不会仅仅只根据当前原文和当前引文的相似度计算结果就给出对齐成功与否的结论,依照当前原文的相似度计算结果的不同,分四种情况。

一是当前原文和引文的相似度等级为优,则对齐成功。回溯上一句原文和引文的相似度等级,若为良或者中,则上一句对齐成功。

二是当前原文和引文的相似度等级为良,则回溯上一句原文和引文的相似度等级:若为优,则当前对齐成功;若为良,则上一句对齐成功并且当前对齐成功;若为其他,则上一句对齐失败,当前句暂不判断。

三是当前原文和引文的相似度等级为中,则回溯上一句原文和引文的相似度等级:若为优,则当前对齐成功;若为其他,则上一句对齐失败,当前句暂不判断。

四是当前原文和引文的相似度等级为差,则对齐失败。回溯上一句原文和引文的相似度等级,若为良或者中,则上一句对齐失败。

(三)全局回溯算法

《春秋左传正义》的字数在100万以上,大约是《左传》字数的5倍。因此,当原文A和注疏B对齐失败后,若一直继续往后匹配直到注疏文献末尾,那么在大多数的情况下,原文A总是能够对齐成功的。但是事实上,A的真正引文,或者就是B,或者就在B的附近,而一直往后匹配所找到的那个匹配成功的C,往往并不是A的真正的引文。这就会造成一种严重的错位现象,并且带来连锁反应,导致A之后的所有原文都将无法正确地进行对齐。

对齐并不意味着每一句原文都必须要找到引文。对齐允许失败,而且有些失败是必要的。只有及时地反馈对齐失败的结果,才能及时地回溯,从而使整个对齐过程继续有效地进行下去。因此,本文在对齐算法中加入了全局回溯机制:当注疏文献中连续10行内容都找不到与当前原文匹配的引文时,则当前原文匹配失败,不再向下匹配,回退到注疏文献内容的10行之前,并开始对原文的下一句进行匹配;若原文中连续10个小句都无法在当前10行的注疏文献内容中匹配成功,则跳转至接下来的10行注疏文献内容中,回退到原文的10个小句之前,继续匹配。这样处理的好处是,无论是由于算法误判而造成的对齐失败,还是由于错误的对齐成功而造成的错位现象,都可以被控制在有限的范围之内,而不影响全局(图2)。

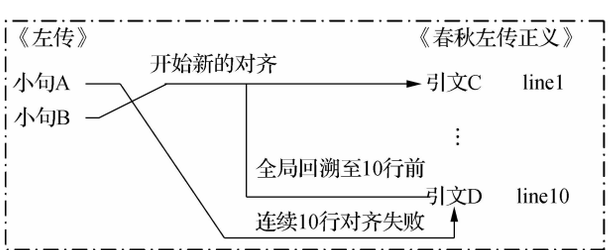

图2 全局回溯示例

(四)《左传》及其注疏文献自动对齐的实验结果

《左传》全文共37 588个小句,在本次实验中,一共对齐了36 917个小句,其中36 835个小句对齐正确,句子对齐的正确率为99.8%,召回率为

98.2%，可见《左传》及其注疏文献的句子对齐效果较为理想。

三、《左传》及其注疏文献的词汇对齐

注疏文献的词汇对齐是对先秦文献进行语义分析及加工的前提。但词汇对齐和句子对齐的过程并不相同，句子对齐的实质是把原文和注疏中的相同文字、相同片段找出来并一一对应，而词汇对齐的任务并不是找出相同的文字或者片段，而是把原文和注疏中的词汇及其相应的释义内容找出来并进行对齐。所以词汇对齐涉及两个具体子任务：一是要找到原文和注疏文献中究竟有哪些词汇；二是要在注疏文献中找到这些词汇的释义内容，并与这些词汇进行对齐。这两个子任务是互相关联的，通过比对原文或引文和释义内容之间的相同文字，可以帮助找出原文或引文中的词汇，通过原文或引文中的词汇，又可以把释义内容中与该词汇相关的所有释义文字准确地定位出来。

通过判断某个字串是否出现在注疏文献中的释义内容部分，可以为判断该字串是否为词汇提供帮助。在确认该字串是词汇的同时，可以把出现该字串的释义内容抽取出来，从而完成词汇对齐。本文设计了两种词汇对齐原则来进行词汇发现和释义内容抽取：一是宽式原则，即只要原文或引文的某部分字串在注疏文献的释义内容部分中出现过，就把该字串看作是一个词汇，并将相关释义内容抽取出来；二是严式原则，即只关注“者、也、称、言、为、曰”等提示词，若原文或引文的某部分字串在注疏文献的释义内容部分和这些提示词紧邻出现过，或者单独出现过（出现在‘’或“”中也算单独出现），则把该字串看作是一个词汇，并将其所位于的释义内容小句（由“。”“？”等标点分隔开的注疏文献部分）抽取出来。由于宽式原则过于宽松，严式原则又有些矫枉过正，两者间需要平衡。因此，本文综合以上两种原则，对《左传》及其注疏文献进行了词汇对齐的实验。以《左传》原文句子“惠公元妃孟子”为例，该句在《春秋左传正义》中相应的注疏内容如下：

傳惠公元妃孟子。言“元妃”，明始適夫人也。子，宋姓。○惠公，名不皇。謚法愛人好與曰惠。其子隱公，讓國之君。元妃，芳非反。傳曰“嘉耦曰妃”。適，本又作嫡，同，丁厝反。

［疏］傳“惠公元配孟子”。正義曰：惠公，名弗皇，孝公之子也。謚法：“愛民好與曰惠。”《釋詁》雲“元，始也。妃，匹也”。始匹者，言以前未曾娶，而此人始爲匹，故注雲：言元妃，明始適夫人也。妃者，名通適妾，故傳雲“陳哀公元妃鄭姬生悼大子偃師，二妃生公子留，下妃生公子勝”。元者，始也，長也。

……

但林父、荀首并得立家，故荀首子孫亦從適長稱伯也。或可春秋之時不能如禮，孟伯之字無適庶之異，蓋從心所欲而自稱之耳。契姓子，宋是殷後，故子爲宋姓。婦人以字配姓，故稱孟子。[3]

综合利用宽式原则和严式原则，对上述原文内容进行自动词汇对齐。

第一步，找出所有可能构成词汇的字串（设最大词长为 3），并把这些字串在《春秋左传正义》中出现过的释义内容部分也找出来并进行对齐，共有 12 个候选词汇（图 3）。

[1]	惠：	謚法愛人好與曰惠、愛民好與曰惠
[2]	惠公：	惠公
[3]	惠公元：	【无】
[4]	元：	元者
[5]	元妃：	言“元妃”、元妃、言元妃
[6]	元妃孟：	【无】
[7]	妃：	傳曰“嘉耦曰妃”、妃、妃者、妃者配匹之言、是大夫之妻亦稱妃也
[8]	妃孟：	【无】
[9]	妃孟子：	【无】
[10]	孟：	《禮緯》雲“庶長稱孟”、則稱爲孟、故或稱孟氏、趙氏恒爲庶而稱孟者也
[11]	孟子：	故稱孟子
[12]	子：	子、孝公之子也、則趙武適妻子也、則荀吳妾子也

图 3　词汇对齐第一步

第二步，去除从未在《春秋左传正义》释义内容部分出现过的候选词，此时还剩下 8 个候选词（图 4）。

[1]	惠：	謚法愛人好與曰惠、愛民好與曰惠
[2]	惠公：	惠公
[3]	元：	元者
[4]	元妃：	言“元妃”、元妃、言元妃
[5]	妃：	傳曰“嘉耦曰妃”、妃、妃者、妃者配匹之言、是大夫之妻亦稱妃也
[6]	孟：	《禮緯》雲“庶長稱孟”、則稱爲孟、故或稱孟氏、趙氏恒爲庶而稱孟者也
[7]	孟子：	故稱孟子
[8]	子：	子、孝公之子也、則趙武適妻子也、則荀吳妾子也

图 4　词汇对齐第二步

第三步，去除包含在更长候选词中的候选词

(分词的长词优先原则),此时还剩下3个候选词,这3个候选词组合在一起正好构成了原文句子"惠公元妃孟子",词汇对齐完成(图5)。

[1] 惠公: 惠公
[2] 元妃: 言"元妃"、元妃、言元妃
[3] 孟子: 故稱孟子

图5 词汇对齐第三步

四、结语

本文探讨了《左传》及其注疏文献的三种自动对齐过程:句子对齐、注释对齐和词汇对齐。在句子及注释对齐试验中,对齐的正确率为99.8%,召回率为98.2%,效果较为理想。相比之下,词汇对齐的任务仍然很艰巨。本文提出了两种词汇对齐的原则,并通过具体的实例展示了《左传》及其注疏文献的词汇对齐效果,但在大规模的文本实验过程中仍然存在两个需要解决的问题。

一是如何有效地从注疏文献的释义内容中抽取信息。"有效"包含两层意思:有用的信息尽量不遗漏,无用干扰的信息尽量筛掉。例如《左传》原文"繼室以聲子",对于"聲子"这个词,注疏文献的释义部分只能找到"聲,谥也"的信息,但是在该句中,"聲子"才是正确的词。又如原文"莊公寤生,驚姜氏",注疏文献的释义部分可以找到"驚姜氏",但通过对注疏的仔细观察可以找到"寤寤而莊公已生,故驚而惡之"这样的句子,"故驚而惡之"可以表明"驚"是个动词,加之"姜氏"可以在上下文全篇中找到,因此可以有把握地推翻"驚姜氏"是一个词的结论。但这样的判断对于人来说很容易,让机器能够自动甄别和正确筛选却很难。

二是如何有效地利用从注疏文献的释义内容中抽取到的信息。对于先秦文献来说,发现词汇的过程其实就是一个寻找多字词的过程,我们只需要关注抽取到的信息中的多字词,例如,对于原文"邾子克也",我们抽取出了"王命以爲邾子"这样的信息,就可以对它进行正确分词了,而不用理会"邾""克"也被抽取到。但如果可以抽取到"此傳言'爲魯夫人'者",也可以抽取到"其'友'及'夫人'",那么"魯夫人"和"夫人"都是经过严式原则的筛选,是最终"合格"的候选词。如何在这两个"合格"的候选词之间进行取舍,需要挖掘更多注疏内部的知识来支撑。

找到解决上述问题的有效办法是提高《左传》及其注疏文献词汇对齐效果的关键所在,也是本研究需要进一步探索和改进的方向。

参考文献:

[1] 陈小荷,冯敏萱,徐润华.先秦文献信息处理[M].北京:世界图书出版公司,2013:13-15
[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):50-55
[3] 张丽娟.八行本《春秋左传正义》版刻辨析[J].清华大学学报,2018(3):115-122,192
[4] 徐润华,陈小荷.一种利用注疏的《左传》分词新方法[J].中文信息学报,2012(2):13-17
[5] 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(24):100-102

(责任编辑:刘 鑫)