注疏文献中的注释语句自动分析

马创新[1] 陈小荷[1] 曲维光[2]
(南京师范大学文学院 南京 210097)[1] (南京师范大学计算机科学与技术学院 南京 210097)[2]

摘 要 注疏文献中蕴含着丰富的知识,并且它们的行文方式具有半结构化特征。研究了经典古籍与其注疏文献句子对齐的方法,以及注疏文献中注释语句的自动分析方法。该项研究成果能为古籍语料库精加工提供便捷的途径,也能为语言研究者提供更为智能的检索模式。

关键词 古籍数字化,注疏文献,句子对齐,注释
中图法分类号 TP391 **文献标识码** A

Automatic Analysis of Comments in Commentary Literatures

MA Chuang-xin[1] CHEN Xiao-he[1] QU Wei-guang[2]
(College of Liberal Arts, Nanjing Normal University, Nanjing 210097, China)[1]
(College of Computer Science and Technology, Nanjing Normal University, Nanjing 210097, China)[2]

Abstract The commentary literatures contain a wealth of knowledge, and they have the characteristics of semi-structured. This paper researched sentences alignment between classic original and their commentary literatures, and automatic analysis of comments in commentary literatures. The study can provide a convenient way to build the ancient corpus, and more intelligent retrieval mode for language researchers.

Keywords Digitization of ancient books, Commentary literatures, Sentence alignment, Comments

1 引言

我国的古籍数字化研究开始于20世纪80年代,发展到今天,已经产生了大量成果,但是以往偏重于研究古籍内容再现的方法和技术,而对古籍文献做精加工和知识挖掘的研究还不够充分,古籍文献的检索方式也不够智能。注释语句自动分析就是利用计算机使用相关算法自动发现注疏文献中的注释语句,分析找出每条注释语句的"注释对象",并把"每条注释语句"与引文中相应的"被注释成分"联系起来。在做注释语句自动分析之前,要先实现注疏文献与其经典原文的句子对齐,句子对齐能够在注疏文献与其经典原文之间建立起双向的联系,并且能把注释文献中的"引文"与"注释"区分开来。

2 研究内容及意义

在早期出现的大部分注疏文献中,"注"是单独行文自成一书的。约从东汉开始,为方便阅读,出现了经与注合一的经注本,一般的行文体例是先引用原文中的一句或几句话,其注文紧随所要注释的文字。

"经典古籍与其注疏文献的句子对齐"是指利用计算机使用句子对齐算法把经典古籍中的各个句子,与其注疏文献中"引文"(即引用经典古籍的内容)的相应句子进行自动对应的过程。比如在《论语》和《论语集注》中有这样两段内容:

子曰:'笃信好学,守死善道。危邦不入,乱邦不居。天下有道则见,无道则隐。邦有道,贫且贱焉,耻也;邦无道,富且贵焉,耻也。"(《论语》)

子曰:"笃信好学,守死善道。好,去声。笃,厚而力也。不笃信,则不能好学;然笃信而不好学,则所信或非其正。不守死,则不能以善其道;然守死而不足以善其道,则亦徒死而已。盖守死者笃信之效,善道者好学之功。危邦不入,乱邦不居。天下有道则见,无道则隐。见,贤遍反。君子见危授命,则仕危邦者无可去之义,在外则不入可也。乱邦未危,而刑政纪纲紊矣,故洁其身而去之。天下,举一世而言。无道,则隐其身而不见也。惟笃信好学、守死善道者能之。邦有道,贫且贱焉,耻也;邦无道,富且贵焉,耻也。"世治而无可行之道,世乱而无能守之节,碌碌庸人,不足以为士矣,可耻之甚也。晁氏曰:"有学有守,而去就之义洁,出处之分明,然后为君子之全德也。"(《论语集注》)

所指的句子对齐就是把《论语》中的"子曰:'笃信好学,守死善道。'"等3个句子与《论语集注》中相应的3句引文对应起来。

该项研究意义重大。以《论语》为例,《论语》是儒家经典著作,它的注疏文献种类繁多,如果能够把《论语》的各部注疏文献分别与《论语》实现句子对齐,那么它的各部注疏文献也

到稿日期:2011-11-01 返修日期:2012-03-25 本文受国家社科基金重大项目(10&ZD117),江苏高校重点研究基地重大项目(2010JDXM023),江苏省普通高校研究生科研创新计划项目(CXZZ12_0357)资助。

马创新(1980—),男,博士生,主要研究方向为计算语言学,E-mail:mchuangxin@gmail.com;陈小荷(1952—),男,博士,教授,博士生导师,主要研究方向为计算语言学;曲维光(1964—),男,博士后,教授,博士生导师,主要研究方向为计算语言学和人工智能。

就可以通过《论语》这个枢纽实现相互之间的“引文对齐”，因此就在这些著作之间构建起相互联系的关系网。

如果一部注疏文献与其经典原文实现了句子对齐，那么在这部注疏文献中无法形成对齐关系的句子就是对引文的注释，由此就把注疏文献中的内容分成了两类：“引文”和“注释”。注释语句自动分析就是设计相关算法自动发现注疏文献中的“注释语句”，分析找出“注释对象”，把“注释语句”与引文中“被注释成分”联系起来。比如朱熹的《论语集注》有下面一段内容：

有子曰：“其为人也孝弟，而好犯上者，鲜矣；不好犯上，而好作乱者，未之有也。弟、好，皆去声。鲜，上声，下同。有子，孔子弟子，名若。善事父母为孝，善事兄长为弟。犯上，谓干犯在上之人。鲜，少也。作乱，则为悖逆争斗之事矣。此言人能孝弟，则其心和顺，少好犯上，必不好作乱也(《论语集注》)。

注释语句自动分析就是在这段内容中自动发现诸如“有子，孔子弟子，名若。”之类的注释语句，判断出该条语句的“注释对象”是“有子”，并把该条语句与引文中的“有子”联系起来，形成如表1所列的对应关系。

表1　注释语句自动分析示例表

注释对象	注释语句
有子	有子，孔子弟子，名若。
孝	善事父母为孝
弟	善事兄长为弟 弟、好，皆去声。
好	弟、好，皆去声。
犯上	犯上，谓干犯在上之人。
鲜	鲜，少也。鲜，上声，下同。
作乱	作乱，则为悖逆争斗之事矣。

如果在一部经典古籍与其多部注疏文献之间实现了句子对齐，并且这多部注疏文献又各自实现了注释语句的自动分析，那么就会在这些注疏文献的所有注释之间建立起相互联系的网络，通过该网络，可以把某一注释对象在各部注疏文献中的注释汇集起来，便于综合比较。该项研究成果可应用于古籍语料库建设之中，实现了句子对齐和注释语句自动分析的古籍语料库，能为语言研究者提供智能的检索模式，还能为编纂古汉语词典提供素材。

3　存在的主要困难

在句子对齐过程中，最常见的对齐情况是“1∶1句子对齐”，即经典古籍中的一个句子与其注疏文献中引文的一个句子相对应。但由于经典原文与其注疏文献中的引文在断句位置上时常不一致，因此还会出现一对多，或者多对多的句子对齐。古籍中还经常有字形变异、词语省略等现象，这也加大了句子对齐的难度。

训诂学家通常会从多个角度对一段引文进行注释，采用的训诂术语也是多种多样的。注释语句的自动分析分为4个步骤：

1. 自动区分注疏文献中的“引文”与“注释”；
2. 自动判断出注释部分中的注释语句；
3. 自动分析出每条注释语句中的“注释对象”；
4. 自动构建注释语句与引文部分中“被注释成分”的联系。

4　经典古籍与其注疏文献的句子对齐

句子对齐是当今基于实例的机器翻译研究中的重要技术，其发展已经较为成熟。句子对齐的方法可以分为3大类：基于长度的方法[1]、基于词汇信息或字信息的方法[2]、长度和词汇相结合的方法[3]。考虑到古代汉语单音节词占绝大多数，并且如果分词错误会造成错误的扩散，本文使用基于字信息的方法实现句子对齐。

本文的句子对齐是在篇章对齐的基础上进行的。用 $Chapter\langle O,C\rangle$ 表示包含 m 句经典原文和 n 句注疏文献的一对篇章，其中 O 为 m 个经典原文句子($SO_1\cdots SO_i\cdots SO_m$)组成的句子序列，$C$ 为 n 个注疏文献句子($SC_1\cdots SC_j\cdots SC_n$)组成的句子序列，$SO_i$ 表示第 i 个经典原文句子，SC_j 表示第 j 个注疏文献句子。$a=\langle O_a,C_a\rangle$ 构成一个对齐的句对，根据含有的经典原文句子个数和注疏文献句子个数，对齐句对分为1∶1,1∶0,0∶1,2∶1,1∶2,2∶2,1∶3,3∶1等8种类型。经典原文和注疏文献的一对篇章中存在多种句对组合序列，每种句对组合序列就是一种对齐方式。句子对齐就是要寻找一个最佳的句对组合序列 $A=a_1\cdots a_h\cdots a_r$($r$ 为该组合序列中的句对个数)。假设存在 k 种句对组合序列，那么句子对齐问题实质上就是在这 k 种句对组合序列中寻找最优句对组合序列的问题，见式(1)。

$$A=\arg\max_{1\leqslant g\leqslant k}S(A_g) \tag{1}$$

这里用评价函数 S 来评估一个句对序列，每个可能的句对序列都有一个评估值 $S(A_g)$，它是该序列中各个句对评价函数值之和。比如在第 g 个句对序列中共有 r 个句对，即 A_g($a_1\cdots a_r$)，那么该句对序列的评价值计算方法见式(2)。

$$S(A_g)=\sum_{h=1}^{r}Score(a_h) \tag{2}$$

式中，$Score(a_h)$ 为句对 a_h 的评价函数，在本文研究中该函数的主要作用是计算两个句子的相似度。

目前句子相似度计算的方法一般可以概括为4大类：基于字面特征或词面特征的方法[4]、基于词义特征的方法[5]、基于句法特征的方法以及基于语用特征的方法。这些方法各有优点和缺点，比如：基于字面特征或词面特征的方法简单易行、可操作性强，但没有考虑到词的同义替换和词义距离；基于词义特征的方法考虑到了词义因素，但没有考虑到词义之间相互作用所形成的关系；基于句法特征的方法考虑到了句法因素对句子相似度的影响，但往往无法单独使用，需要结合其他方法；基于语用特征的句子相似度计算一直是人们的目标，但实现难度相当大，目前这方面的研究还没有达到实用水平。这些方法之间并没有优劣之分，要根据应用的领域选用适当的方法。本文使用基于字面特征的句子相似度计算方法，见式(3)。

$$Score(a_h)=Score(SO_i,SC_j)=\vec{V_1},\vec{V_2}=\frac{\sum_{i=1}^{n}W_i\cdot\phi_i}{\sqrt{\sum_{i=1}^{n}\phi_i^2\cdot\sum_{i=1}^{n}W_i^2}} \tag{3}$$

假设句对 a_h 中有句子 SO_i 和 SC_j，它们的所有汉字构成的向量空间为 $V=\{X_1,X_2,X_3,\cdots,X_n\}$。句子 SO_i 的向量 $V_1=\{W_1,W_2,W_3,\cdots,W_n\}$，其中 W_i 为 X_i 在句子 SO_i 中出现的频次。句子 SC_j 的向量 $V_2=\{\phi_1,\phi_2,\phi_3,\cdots,\phi_n\}$，其中 ϕ_i 为 X_i 在句子 SC_j 中出现的频次。

经典原文和注疏文献的一对篇章中存在多种句对组合序列，其中最优句对序列的各个句对的相似度之和在所有句对

序列中是最大的。我们使用动态规划算法搜索最优句对序列[6,7]，其核心递推公式见式(4)。

$$S(i,j)=\max\begin{cases}S(i-1,j-1)+Score(SO_i,SC_j) & 1:1\text{型}\\ S(i-1,j)+Score(SO_i,\text{null}) & 1:0\text{型}\\ S(i,j-1)+Score(\text{null},SC_j) & 0:1\text{型}\\ S(i-2,j-1)+Score(SO_{i-1}+SO_i,SC_j) & 2:1\text{型}\\ S(i-1,j-2)+Score(SO_i,SC_{j-1}+SC_j) & 1:2\text{型}\\ S(i-2,j-2)+Score(SO_{i-1}+SO_i,SC_{j-1}+SC_j) & 2:2\text{型}\\ S(i-1,j-3)+Score(SO_i,SC_{j-2}+SC_{j-1}+SC_j) & 1:3\text{型}\\ S(i-3,j-1)+Score(SO_{i-2}+SO_{i-1}+SO_i,SC_j) & 3:1\text{型}\end{cases}\tag{4}$$

$S(i,j)$表示经典原文的前 i 个句子和注疏文献的前 j 个句子组成的所有对齐序列中最优序列的得分。$Score(SO_{i-1}+SO_i,SC_j)$中的＋号表示把其两边的字符串 SO_{i-1}和 SO_i 连接起来作为一个整体，计算与句子 SC_j 之间的相似度。

采用这种方法做《论语》与其注疏文献的句子对齐实验，注疏文献选用朱熹的《论语集注》和邢昺的《论语注疏》。这两部著作版本很多，我们对多个版本进行比较甄选，《论语集注》采用中华书局 1957 年出版的《四书集注》中的版本，《论语注疏》采用上海古籍出版社 1990 零出版的《十三经注疏》中的版本。《论语》经典原文的版本也很多，后人所加的标点也不一致，我们选用中华书局 1980 年出版的，近人杨伯峻的《论语译注》中所引用的《论语》原文。对实验结果进行评测时，采用正确率(P)、召回率(R)和 F 值(F) 3 个评价指标：

P=(正确对齐的句珠数÷对齐的句珠数)×100%

R=(正确对齐的句珠数÷实际存在的句珠数)×100%

F=((正确率×召回率×2)÷(正确率＋召回率))×100%

《论语》分别与《论语集注》、《论语注疏》的句子对齐实验结果见表 2。其中，RS 表示实有的句珠数，AS 表示机器对齐的句珠数，CS 表示机器正确对齐的句珠数。

表 2 《论语》与《论语集注》、《论语注疏》句子对齐实验结果

	RS	AS	CS	P(%)	R(%)	F(%)
《论语集注》	2780	2772	2750	99.21	98.92	99.06
《论语注疏》	2780	2772	2692	97.11	96.83	96.97

从表 2 中可以看到，与《论语集注》句子对齐实验的 3 项评测指标都接近百分之百，而与《论语注疏》句子对齐的实验结果要稍差一些。产生这种情况的主要原因是《论语注疏》总体篇幅约是《论语集注》的两倍，引文之后的注释语句非常多，而且有些注释语句与引文的句子相似度很大，这对句子对齐造成干扰，使产生错误的概率增加。对机器未能对齐的句子进行分析，发现这些本应对齐的句子之间相似度极低，只有很细的规则才能覆盖到这些句珠。

5 注释语句的自动分析

作为一门历史久远的学科，训诂学有很多常用术语[8]，用正则表达式对这些具体的训诂术语做抽象化表述，通过建立训诂术语的模式库来实现注释语句的自动分析。由于训诂术语的发展有一个从零散到系统的过程，因此不同历史时期的训诂著作在术语种类上会有所不同。把普遍使用的训诂术语用正则表达式表述，并存入一个模式库中。在做注释语句自动分析时，根据特定注疏文献的训释特点，增加或减少模式库中的模式[9]。

训诂学的主要任务是解释词义，释义的方法多种多样，训诂术语和相关的正则表达式如表 3 所列。

表 3 解释词义的训诂术语以及相关正则表达式

类别	训诂术语		正则表达式
人名、物名	x,名 y x,y 名(也)	x,字 y x,姓 y	(.＋),[名字姓].＋ (.＋),.{1,5}名也?
动词	x,y 之	x,为 y	(.＋),.＋之 (.＋),为.＋
形容词	x,y 貌 x,y 的样子	x,y 之状 x,y 意也	(.＋),.＋(貌|之状|的样子|意也)
代词	x,身也 x,指 y 之辞	x,自称之词 x,指 y 辞也	(.＋),(身也|自称之词) (.＋),指.＋(之辞|辞也)
助词	x,y 辞也 x,y 语助(也)	x,y 辞 x,发声也	(.＋),.＋(辞也|辞|语助也|语助) (.＋),发声也
叹词	x,y 叹辞(也) x,y 嗟叹辞也	x,y 之声	(.＋),.＋(嗟叹辞也|叹辞) (.＋),.＋之声
象声词	x,y 声(也)		(.＋),.＋(声也|声)
通用术语	x,y 也 x 者,y 也	x 者,y x,谓 y	(.＋)者,.＋也,(.＋)者,.＋,(.＋),.＋也,(.＋),谓.＋

在表 3 中，“训诂术语”列中，用 x 表示注释对象，用 y 表示注释语句中的主体部分。在“正则表达式”一列中，所有的正则表达式都使用了分组保存运算符。能够与各个正则表达式中的第一组模式相匹配的字符串就是注释语句中的注释对象，其会被保存到缓冲区 1 中，可以使用“向后引用”原子 \1 来引用它，接下来再判断这个注释对象是否出现在与注释语句相同文本块的引文中。

实现句子对齐后，在注疏文献中，能与经典原文句子构成对齐关系的句子属于“引文”，无法构成对齐关系的句子属于“注释”；由此注疏文献被分成为多个“文本块”，每个文本块是由“连续的引文和紧随其后的注释”组成。注释语句自动分析的算法可以描述为：

步骤 1 循环处理注疏文献的每个文本块，处理方法为：

步骤① 循环处理此文本块注释中的每个句子，处理方法为：

步骤 i 判断该句是否与模式库中的任一个模式相匹配；

步骤 ii 如果相匹配，判断该句的注释对象是否在此文本块的引文中出现；

步骤 iii 如果上一条件为真，就把该句与引文中的“被注释成分”联系起来。

为了验证该算法的可行性，选取了朱熹的《论语集注》进行实验。实验中只分析注释对象为“字”、“词”和“短语”的 3 类注释，以人工分析的注释语句作为标准答案，考察其正确率(P)、召回率(R)以及 F 值(F)。

P=正确分析的注释数÷分析的注释数×100%

R=正确分析的注释数÷文献中的注释数×100%

F=((正确率×召回率×2)÷(正确率＋召回率))×100%

实验发现：在《论语集注》中共有注释对象为“字”、“词”和“短语”的注释 2042 条，使用这种方法可以分析找出注释 1962 条，其中 1835 条注释得到正确分析，正确率为 93.53%，召回率为 89.86%，F 值为 92.29%。

在实验中，发现正则表达式的灵敏性和精确性是一对矛盾，过分追求正则表达式的灵敏性，就会降低精确性；反之，着力提高它的精确性，就会降低灵敏性。正则表达式的精确性和灵敏性这一对矛盾反映到注释语句自动分析的实验上，就是正确率和召回率的矛盾。

结束语 在研究句子对齐方法时，结合古籍文献独有的特点，计算经典原文与其注疏文献的句子的相似度，使用动态规划算法实现了句子对齐。注释语句自动分析研究是在实现了句子对齐的基础上进行的。用正则表达式对常用的训诂术语做抽象化概括，通过建立训诂术语的模式库来实现注释语句的自动分析。

以上述两项研究成果所构建的语料库[10]为基础，设计了经典古籍与其注疏文献对齐语料库检索平台(见图1)。利用该平台，语言研究者可以很方便地检索到经典古籍中的某一注释对象，在与其相关的各个注疏文献中的注释，便于语言研究者进行综合比较。

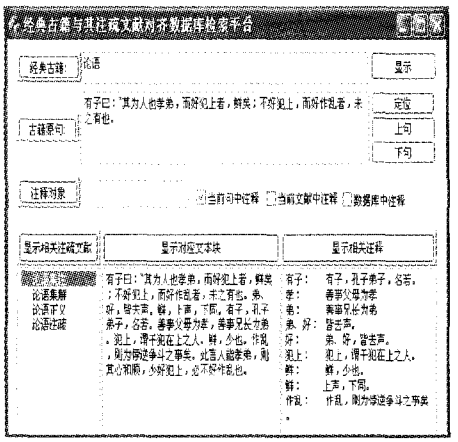

图1 经典古籍与其注疏文献对齐数据库检索平台

参 考 文 献

[1] Gale W A, Church K W. A Program for Aligning Sentences in Bilingual Corporal[J]. Computational Linguistics, 1993, 19(1): 75-90

[2] Champollion M X. A Robust Parallel Text Sentence Aligner [C]// Proceedings of LREC-2006: Fifth International Conference on Language Resources and Evaluation. 2006: 489-492

[3] Moore R C. Fast and Accurate Sentence Alignment of Bilingual Corpora [C]// Proceedings of AMTA. Springer-Verlag, 2002: 135-144

[4] Nirenburg S. Two Approaches of Matching in Example-Based Machine Translation[C]// Proc. TMT-93. Kyoto, Japan, 1993

[5] Li S, Zhang J, et al. Semantic Computation in Chinese Question-Answering System[J]. Journal of Computer Science and Technology, 2002, 17(6): 933

[6] 郭锐，宋继华，廖敏. 基于自动句对齐的相似古文句子检索[J]. 中文信息学报，2008，22(2)：87-91

[7] 于新，吴健，洪锦玲. 基于词典的汉藏句子对齐研究与实现[J]. 中文信息学报，2011，25(4)：57-62

[8] 许威汉. 训诂学读本[M]. 上海：上海交通大学出版，2010：48-84

[9] Watt A. 正则表达式入门经典[M]. 李松峰，李丽，译. 北京：清华大学出版社，2008：156-178

[10] 丁溪源，黄河燕，张海军，等. 基于大规模语料划分的频繁模式查找算法[J]. 计算机科学，2012，39(3)：149-152

(上接第219页)

量之间的关系，为解决非线性、多变量耦合问题提供了新的思路。目前，离散型分布估计算法研究已经比较成熟，而连续域分布估计算法的研究比较缓慢，这方面的资料很少。

本文根据连续域优化问题的特征以及分布估计算法的内涵，探索研究了采用均匀分布作为概率模型、保持概率模型不变缩小采样领域以及保留优势个体确保进化方向等思想设计了针对连续域函数优化的分布估计算法。

连续域分布估计算法的难点一方面是变量的取值问题，即由于变量是连续域，在取值范围内有无限种取值方法，编码困难，并且使得优化算法的搜索空间非常大，因此获得更好的采样区间是下一步的研究方向；另一方面，种群的规模、优势种群的规模、更新时保留的个体规模等也是下一步要研究的方向，其它的诸如连续域分布估计算法的采样算法、算法的收敛特性也比较有研究的价值。

参 考 文 献

[1] 周树德，孙增圻. 分布估计算法综述[J]. 自动化学报，2007，33(2)：113-124

[2] 许昌，常会友，徐俊. ASON网中基于分布估计的恢复容量优化算法[J]. 计算机科学，2010，37(7)：183-185

[3] Salimas-Gutierrez R, Hernandez-Aguirre A, Villa-Dihasrce E R. Dependence Trees with Copula Selection for Continous Estimation of Distribution Algorithms[C]//GECCO '11

[4] Marti L, Garcia J, Berlanga A, et al. On the Computational Properties of the Multi-Objective Neural Estimation of Distribution Algorithm[C]// Nature Inspired Cooperative Strategies for Optimization(NICSO 2008). 2009: 239-251

[5] Godingho P, Meiguins A, Oliveira R, et al. An Estimation of Distribution Algorithms Applied to Sequence Pattern Mining[C]// Innovations in Computing Sciences and Software Engineering. 2010: 589-593

[6] Salinas-Gutierrez R, Hernandez-Aguirre A, Villa-Diharrce E R. Estimation of Distribution Algorithms based on Copula Functions[C]//GECCO '11

[7] Salinas-Gutierrez R, Hernandez-Aguirre A, Villa-Diharce E R. Dvine EDA: A new Estimation of Distribution Algorithms based on Regular Vineas[C]//GECCO '10

[8] Lima C, Pelikan M, Goldberg D, et al. Influence of selection and replacement strategies on linkage learning in BOA[C]// Evolutionary Computation, 2007 (CEC 2007). IEEE Congress, Washington DC: IEEE, 2008: 1083-1090

[9] Naeem M, Lee D. Estimation of Distribution algorithm for sensor selection problems [C] // Radio and Wireless Symposium (RWS), 2010 IEEE. Washington DC: IEEE, 2010: 388-391

[10] Suganthan P N, Hansen N, Liang J J, et al. Problem Definitions and Evaluation Criteria for the CEC 2005 Special Session on Real-Parameter Optimization[R]. 2005