针对古代经典文献的引用查找问题的数据构建与匹配方法

李炜，邵艳秋
北京语言大学
信息科学学院
北京市海淀区学院路15号，100083
liweitj47@blcu.edu.cn，yqshao163@163.com

毕梦曦*
复旦大学
哲学学院
上海市杨浦区邯郸路220号，200433
1207950557@qq.com

摘要

中国古代思想家的思想建构往往建立在对更早期经典的创造性诠释中，将这些诠释中包含的引用查找出来对思想史研究意义重大。但一些体量较大的文献如果完全依靠手工标记引用将耗费大量时间与人力成本，因此找到一种自动化的方法辅助专家进行引用标记查找非常重要。以预训练语言模型为代表的自然语言处理技术的发展提升了计算机对于文本处理和语义理解的能力。据此，本文提出多种利用专家知识或深度学习语义理解能力的无监督基线方法来自动查找古代思想家著作中对早期经典的引用。为了验证本文提出的方法的效果并推动自然语言处理技术在数字人文领域的应用，本文以宋代具有重大影响力的理学家二程（程颢、程颐）对早期儒家经典的引用为例进行研究，并构建和发布相应的引用查找数据集[1]。实验结果表明本文提出的基于预训练语言模型和对比学习目标的复合方法可以较为准确地判断是否存在引用关系。基于短句的引用探测ROC-AUC值达到了87.83，基于段落的引用探测ROC-AUC值达到了91.02。进一步的分析表明本文的方法不仅有利于自动化找到引用关系，更能够有效帮助专家提高引用查找判断效率。本方法在注释整理、文本溯源、重出文献查找、引用统计分析、索引文献集制作等方面具有广阔的应用前景。

关键词： 引用查找；数字人文；古代文献

Data Construction and Matching Method for the Task of Ancient Classics Reference Detection

Wei Li, Yanqiu Shao
Beijing Language and Culture University
School of Information Science
15 Xueyuan Rd., HaiDian District,
Beijing, 100083
liweitj47@blcu.edu.cn，yqshao163@163.com

Mengxi Bi
Fudan Univeristy
School of Philosophy
Handan Rd., Yangpu District
Shanghai, 200433
1207950557@qq.com

Abstract

The idea construction of ancient Chinese ideologists tend to be built on the basis of explaining early ideological claims. Therefore, finding out the references owes great significance for the research on ideological history. However, it would be much too expensive for both time and man power if we only fully depend on human experts to label the literature especially when the literature is of large amount. Hence force, it is of great importance to develop an automatic method to facilitate experts looking for the reference items. With the development of natural language processing technologies typified by pre-trained language models, the ability for processing and understanding

* 通讯作者 Corresponding Author
[1]数据可在https://github.com/liweitj47/classic_reference_detection找到

第二十一届中国计算语言学大会论文集，第600页-第610页，南昌，中国，2022年10月14日至16日。

natural language has improved a great deal. Based on these observations, we propose several unsupervised baseline methods to automatically detect the references to early literature, which use either expert knowledge or deep language understanding technologies. To testify the effectiveness of our proposed method as well as promote the application of natural language processing techniques to the field of Digital Humanities, we take the literature of Ercheng referencing early Confusion classics as example, and construct the corresponding labelled dataset for reference detection. The experiment results show that our ensemble method based on pretrained language model and contrastive objective can accurately detect whether there exists reference. Sentence level reference detection achieves 87.83 on ROC-AUC, while paragraph level reference detection achieves 91.02 on ROC-AUC. Further analysis show that this work can not only help automatically find reference, but also improve the efficiency for the expert finding references. This model has great prospects on organizing annotation, text tracing, duplicate literature detection, reference statistical analysis and reference index generation.

Keywords: Reference Detection , Digital Humanity , Ancient Classics

1 引言

中国古代思想家的思想建构往往不是从零开始自己创建一套思想体系，而是在诠释早期经典的过程中完成自己思想的构建，中国古代思想的发展演变在这一方面具有一定的特殊性。黄俊杰先生就曾提出(黄俊杰, 2018)，“儒家经典解释者每一次所提出的新解释，都是一次通过解释者个人的思想系统和生命体验而完成的再创造。但同时解释者也不是完全自由的，诠释性的发挥需要在原典文本的印可范围之内。”所以在中国古代思想研究这一领域中，无论是哲学研究、思想史研究还是经学研究，思想家们对早期经典的引用与诠释都是研究中重要的问题。宋元明清思想史研究中，材料体量普遍比较大，给查找工作带来了难度。

传统人文研究者面对体量比较大的文献材料时，引用的查找统计就会变得非常困难。以本文所举的二程文献为例，中华书局点校本《二程集》总体字数约50万字左右，按照点校者的分段计数，共有语录、文章、书信等不同体裁的文字6000多段，在宋元明清思想研究中属于中等的文献量。即便是在准确识别引用的前提下，统计和标注也需要耗费比较大的工作量，这些重复工作会给人文领域的研究带来巨大的时间和人力成本。

利用自然语言处理技术来处理规模较大的文本具有天然的优势，对处理经典古籍等文本也具有很强的借鉴意义，为数字人文领域提供了重要的技术基础。但是传统的基于字符匹配的方法存在依赖专家经验，灵活度不足，覆盖范围不全，难以准确把握上下文语义内涵等问题。这些问题使得人文领域学者难以在实际中使用这类工具得出可靠的结论，限制了相关技术在数字人文领域的应用。

随着预训练模型等深度学习方法的发展(Devlin et al., 2019; Qiu et al., 2020)，其被应用在了自动问答(Yang et al., 2019)、阅读理解(Zhang et al., 2020)、机器翻译(Conneau et al., 2020)等众多任务中，并取得了巨大的成功，而这与预训练语言模型能够在一定程度上能够更好地把握上下文语义有关。此外，基于对比学习提出的SIMCSE(Gao et al., 2021)在预训练语言模型的基础上，通过对句中字符表示随机加入dropout(Srivastava et al., 2014)噪音的方式来获得对比学习(Jaiswal et al., 2020)中的正例，因而能够通过自监督方式学习到包含语义信息的更有区分度的向量表示。

在本文中，我们结合人文领域专家对文本本身及引用现象的观察和经验以及预训练语言模型和对比学习目标，提出了三种基础的引用查找和探测方法，并基于这三种基础方法，组合出不同的复合方法作为此任务的基线模型。为了能够对相关任务和方法做出可以量化的评价，我们首先给出了引用问题的明确定义，之后筛选出二程文本中一些较有代表性的文本，并标注了它们中对古代儒家最重要的十三经文本是否存在引用关系。最终得到了含822句的开发集

和2484句的测试集。我们在构建的数据集上进行了大量实验。结果表明，我们提出的利用预训练语言模型和对比学习目标的字符片段和句子语义相结合的复合引用查找方法在测试集上的效果最好，基本达到了人文领域实际应用的要求，并且泛化性也相较于基于规则的方法更好。

我们将本文的贡献总结如下：

- 我们提出一种考虑人文领域实际研究应用需求的古代经典引用定义。并据此以二程文本为例，构建并发布二程对早期儒家经典文本（十三经）的引用关系数据集；
- 我们结合人文领域专家经验和预训练语言模型以及对比学习目标，提出了多种引用查找和探测的基线方法，并进行了大量的实验和分析。结果表明，基于预训练语言模型目标和对比学习的字符片段和句子两种级别的复合方法可以较为有效地探测出是否存在引用关系，并能为人文学者的实际研究提供有力支持。

2 相关工作

国内学者对于古文引用相关的研究已经取得了一些成果。黄水清等人(黄水清et al., 2021)和周好等人(周好et al., 2021)针对规整古代文献中出现的论著名（明引）采用基于序列标注的方法进行了识别和统计学分析。尽管该方法可以识别带有明确书名的相关表述，后世文献（如宋代及之后）对早期经典文献的引用大多只引用早期经典的只言片语，并且是基于语义的引用，因此该方法并不适用于广泛存在于后世文献中对早期经典文献的引用识别（暗引）。本文的研究方法可以同时识别有书名标引的**明引**和没有书名标引而只有语义关联的**暗引**情况，因此具有更广泛的应用场景以及对人文领域更实际的应用价值。

对于在数字人文领域利用预训练语言模型来说，也有许多工作进行了尝试。耿云冬等人(耿云冬et al., 2022)、刘江峰等人(刘江峰et al., 2021)、胡昊天等人(胡昊天et al., 2022)和徐润华等人(徐润华et al., 2022)利用在四库全书语料上训练的siku-bert(王东波et al., 2022)模型分别尝试了词性标注、实体识别、文本分类和自动摘要的任务，俞敬松等人(俞敬松et al., 2019)将自行训练的古文BERT应用于古文的自动断句，均取得了良好的效果。然而这些工作大多仍然局限在自然语言处理的传统任务范式中，与直接的人文领域研究有一定距离。本文将siku-bert作为基础的预训练模型来辅助进行语义匹配，进而和专家知识、对比学习等相结合，以查找后世（以二程为例）文献对早期儒家经典论述（十三经）的引用。

3 引用查找数据构建

3.1 引用的定义

这里我们对何为引用进行定义。所谓引用需要与原文至少有两个字的重叠，作为引用的标识，而且重叠的部分具有辨识度，对原文的索引有提示作用，提示的作用指能够根据重合的文字找到原文出处之外，重合的文字还不能属于古文中反复出现的字词。以二程文献来说，首先二程文献如果与早期儒家经典有整段的连续重合，可以视为（规则的）引用：

不**甚，则身危国削**，名之曰幽厉，虽孝子慈孙，百世不能改也（《程氏遗书·卷23》）
不甚，则身危国削。（《孟子》）

此外，还存在大量不规则的引用。不规则的引用有多种情况，有的时候二程提及的字数比较少，只提到了关键的两三个字，或者在引用中调换了字词的顺序等，在没有使用模型判断之前，无论是人工判断还是在数据库中搜索，这一部分引用都是查找的难点：

既为先觉之民，岂可不觉未觉者（《程氏遗书·卷1》）
予，**天民之先觉者也。**（《孟子》）

无论是在整段的连续引用还是在不规则的引用中，“引用”都不只是字符的简单重复，还要求重合的字符需要有一定的特殊性和辨识度，能够辨认出二程文献中所提及的词语是来自哪部经典。以二程文献和《孟子》举例说明，比如二程提到了“浩然之气”，这种说法就具有一定的辨识度，可以认为是在用《孟子》中的典故。而“仁义”这样的说法就比较宽泛，多种早期经典中都有出现，虽然《孟子》当中也有，但是无法根据语境认为这个说法是在诠释或者引用《孟子》，这样的重复就不算引用，而应视为二程对“仁”或者“义”这些概念的发挥诠释。

3.2 数据整理和标注

本研究采用中华书局2004年版由王孝鱼先生点校的《二程集》作为数据来源，是较为可靠全面的二程文献集。为了后续导出的结果可以有更多的分析维度，本研究在数据构建阶段即对材料进行了更细致的标注。

首先是书名和卷数的标记。《二程集》是二程的几种文献集合在一起形成的，一共分为63卷，包括《程氏遗书》25卷，《程氏外书》12卷，《程氏文集》12卷，《伊川易传》4卷，《经说》8卷，《粹言》2卷，汇集了二程的全部语录和著作。首先需要对每段文字所在的书名以及卷数进行标记，这样引用查找结果导出后就可以进一步分析引用在二程文献中的分布情况，展开下一步的研究。

其次是作者的标记，即该段文字属于程颐或属于程颢。《程氏文集》《伊川易传》等著作文献作者归属是明确的，根据署名标记作者即可。但二程文献的语录部分（《程氏遗书》《程氏外书》《粹言》）有一部分记录和整理者已经标注了语录属于程颐或程颢，有一部分语录尚不明归属，学生记录的时候没有标明作者。所以需要以程颐、程颢、未知三类对每一段语录的作者进行标记。

第三是对注释性质的文献进行清理。《伊川易传》将《周易》原文分成了1253段，并且按照原文的顺序逐一进行了注释，体例是一段原文，一段注释，所引用的经典原文独自成段。这1253处可以直接进行统计，不需要再进行匹配。在程颐《伊川易传》注释工作中使用的其他儒家经典，则需要通过匹配进一步统计。进行相同处理的还有《经说》当中的《春秋传》和《书解》。文献中的祭文、年谱，文献编辑者所写的序文等由他人完成的，程颐程颢自身没有参与的文献内容需要删除。诗歌部分引用不明显，而且与语录、书信等文献的语言情况差异比较大，所以也进行删除处理。《易序》因为作者归属尚且存疑，进行删除处理。《经说》部分的伊川与明道先生改正《大学》，因为不是二程自己的表述，所以也进行删除。根据葛瑞汉先生的考证(葛瑞汉, 2000)，《经说》当中《春秋解》程颐亲作的部分应当到桓公九年止，所以桓公九年之后的部分予以删除。《经说》部分《中庸解》作者存疑，删除处理。

因为原文本中的许多标点可能会对模型的自动判断起到误导作用，因此本研究首先去除了原文本中诸如“【】『』”等标点符号。为了保证匹配的粒度相对统一，减少句长差异带来的干扰，本研究将文本大致限定在8~30字的范围内。为了达到这样的目的，如果长度超过30，那么本研究按照较大语义停顿的标点符号（如，句号、感叹号、问号等）进行分句。经过分句处理后，部分句子长度会非常短，以致难以提供足够的信息进行匹配。因此本研究对长度太短（小于8字）的句子进行向后合并（与后一句合并）。因为段落中的匹配是建立在内部句子级别的匹配上的，因此本研究中经段落拆分后的句级文本可以较好地用于后续的引用判断。

需要在二程文献中进行查找的经典原文指早期儒家经典十三经，使用古籍文献ctext网站[1]的版本。其中《春秋》三传当中都对《春秋》原文有引用，其体例都是原文引用+注释，为了清楚地查找二程所引的文献是出自《春秋》原文还是传文，将三传中的《春秋》原文清理出去，只留下传文。同时将《春秋》原文单列为一个文件。因为《中庸》《大学》在中国古代的经典注释和传播中往往是独立于《礼记》而进行的，所以将这两篇文章从《礼记》里摘出来单列为两个文件，共有《论语》《孟子》《大学》《中庸》《诗经》《尚书》《周易》《春秋》《仪礼》《礼记》《周礼》《谷梁》《公羊》《左传》《尔雅》《孝经》16个经典文本。

有时候二程讨论早期儒家经典没有直接引述经典内容，而是只提及了书名（如《论语》）或者某个章节的名称（如《离娄》），所以十三经本身的标题以及经典原文当中的各个章节的名称，比如《论语》等，也需要纳入查找的范围。本研究将这些书名和章节名单独列为一个清单，加入所属的经典文件中。《谷梁传》常简称《谷梁》，两个说法都需要列出，其他经典的别称也做同样的处理。对早期经典文献的处理与二程文本基本相同。不同点在于早期经典相较于二程文本来说的表意普遍较为集中、凝练，因此本研究只对早期经典按照标点符号切分成小句，而不对小句进行聚合。

3.3 停用词集构建

停用词主要是指文本中基本不承担实际语义的字或词，在现代汉语中一般以虚词的成分出现。停用词主要用于检索系统中，在查询语句与目标语句进行匹配时，停用词不计入匹配结果或者给与停用词一定惩罚。

[1] https://ctext.org/zhs

古代汉语的虚词与实词的分界本身是一个比较复杂的问题，有些词语介于实词和虚词之间；同时古汉语中也会出现这样的情况：虽然是虚词，但是在很多句子中可以作为查找的标记。筛选标准过分严格会影响召回率。所以停用词的构建不能单纯以实词和虚词来进行区分，而是需要综合考虑该词是否可以作为查找的具有辨识度的依据。同时停用词构建不能只考虑二程的文献情况，而是需要具有一定的普适性，希望在通过少量修改后即可在其他宋元明清思想史文献中也能发挥比较好的效果。

4 引用查找方法

从对问题本身的观察和人文领域专家经验出发，本文认为古文中的大部分引用具有较为鲜明的规律和模式，最突出的特点就是候选文本和参考文本中匹配片段的长度越长或者非连续匹配的字词个数越多，那么候选文本中包含对参考文本中内容的引用的可能性就越大。另一方面，本研究观察到对于匹配片段长度较短且数量较少的情况下，仍有许多引用存在，并且对是否包含引用的判断应该结合具体语境中的上下文信息，通过语义级别的匹配进行判断。对于结合语义的匹配，近年来在许多场景中取得了巨大突破的基于预训练语言模型的深度学习方法可以起到良好的作用。基于以上两个观察，本文提出三种分别利用专家规则、字符片段粒度语义匹配和句子粒度语义匹配的引用查找基础方法，并对三种方法结合得到不同的复合判断方法。

4.1 结合专家知识的规则方法

在本小节中首先介绍结合专家知识的规则判断方法。本研究将候选文本（比如二程集）称为源文本，将参考文本（比如十三经）称为目标文本，源文本和目标文本中连续的字符字面匹配称为直接匹配。即，设源文本为S (source)，包含的字符串为$s_1, s_2, ..., s_n$，目标文本为T (target)，包含的字符串为$t_1, t_2, ..., t_m$ 。其中n和m分别为源文本和目标文本的长度。如果S和T中存在长度为k的连续片段$s_{i+1}, \cdots, s_{i+k} = t_{j+1}, \cdots, t_{j+k}$, 那么我们认为这是一个$k$元组的直接匹配。基于人文领域专家对数据集中划定的开发集的研究，本文认为是否存在引用的关键信息来自于两个方面：

- 源文本（候选文本）和目标文本（参考文本）片段中直接匹配连续文本的长度，匹配片段的长度越长，则包含对目标引用的概率越大。比如出现了匹配的四元组文本片段，则两句有引用的概率较大；
- 源文本（候选文本）和目标文本（参考文本）片段中直接匹配的k元组个数越多，则包含对目标引用的概率越大，如匹配的文本片段个数较多，则两句有引用的概率较大。

除了以上两个原则外，本研究还发现直接匹配的文本片段中如果出现了在古文文本中大量出现的包含较少实际语义的辅助性字词，会对结果产生较大干扰。本研究将此类字词归为停用词（具体的停用词构建参见3.3小节）。

由于古文经常以单字成词，因此本文不对文本进行分词处理，而直接使用字作为单位进行匹配，匹配片段长度范围定义在$1 \sim 4$之间。本研究定义单字匹配的个数为unigram（一元组），双字匹配的个数为bigram（二元组），三字匹配的个数为trigram（三元组），四字匹配的个数为quadgram（四元组）。对这些匹配本文均只考虑不重复的组合，也就是说如果一个字符片段在源文本或目标文本中出现多次，我们只认为它起一次匹配作用。此外，对于较长片段中所包含的较短匹配（比如四元组内部一定会包含三元组），我们不进行重复计数。

以图1所举例子加以说明，源文本S为“既为先觉之民”，目标文本T为“天民之先觉者也”。S和T之间最长的匹配为双字匹配“先觉”，此时trigram和quadgram皆为0，而$bigram = 1$。单字匹配的个数除去“先”和“觉”（已经在二元组中体现）外，还有“民”和“之”，$unigram = 2$。根据本研究的经验和对开发集的观察，本研究通过以下规则来对源文本和目标文本的匹配度进行打分：

$$score_{rule} = unigram \times 0.4 + bigram \times 0.6 + trigram \times 1.4 + quadgram \times 2 \quad (1)$$

$$k-gram = k-gram \times 0.5 \quad if \quad words_{stop} \ \ in \ \ k-gram, \ n = 1, 2, 3, 4 \quad (2)$$

其中，k-gram代表连续片段的长度可以为1,2,3,4，如果k-gram中出现了停用词，那么该匹配片段的实际分数减半。如果多个停用词出现，本研究只对该k-gram分数做一次减半惩罚。

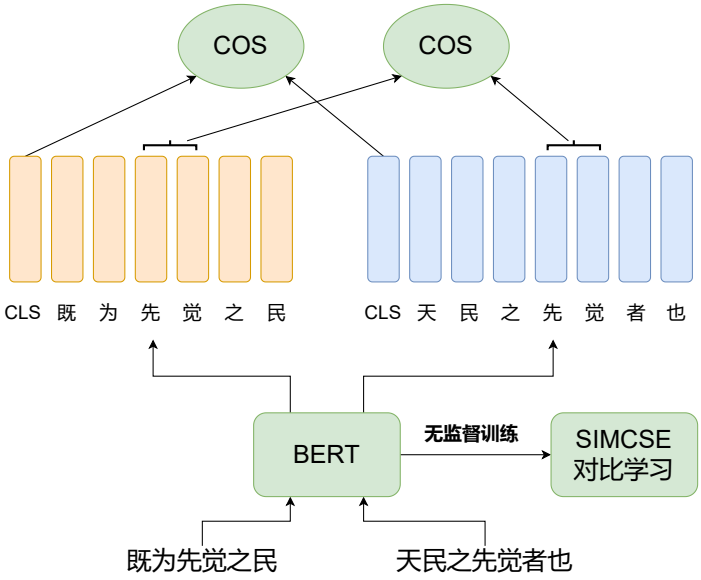

Figure 1: 基于预训练语言模型的字符片段语义和句子语义匹配方法示意图。源文本和目标文本分别经过SIMCSE模型（BERT经过对比学习无监督训练）编码后取重合的片段（字符片段）或CLS（句子）对应的向量表示进行余弦相似度计算

在上面的例子中，unigram中的“之”为停用词，也就是“之”只能算半个匹配，因此$unigram = 1.5$。这样得到$score_{rule} = 1.5 \times 0.4 + 1 \times 0.6 = 1.2$。在实际的匹配判断中，只要score大于等于1，模型就判定引用关系成立。

4.2 基于预训练语言模型的字符片段语义匹配方法

基于预训练语言模型的深度学习方法可以对上下文信息进行更深入地建模，从而更好地把握文本的语义信息。本文提出利用基于预训练语言模型和对比学习目标的深度学习方法来学习源文本和目标文本中匹配文本片段的语义表示，进而通过带有上下文语义的文本语义匹配来判断是否存在引用。相比于上文提到的基于规则的字符匹配判断方法，基于预训练语言模型的深度学习方法可以将上下文信息引入到每个共现k元组的匹配中，从而在匹配时考虑到带有上下文语义的信息。

本文采用基于对比学习的SIMCSE模型，以在古文语料上进行预训练的Sikubert（base）(王东波et al., 2022)为基础，在源文本和目标文本集上均采用掩码语言模型（masked language model）目标和SIMCSE中的对比学习目标进行适应性训练，来学习得到更贴合研究对象文本的模型参数。

如图1中所示，对任意一对待判断的源文本S和目标文本T，本研究首先在S和T的句首加入特殊标签CLS，得到$S' = [CLS, s_1, s_2, ..., s_n]$和$T' = [CLS, t_1, t_2, \ldots, t_m]$作为输入，分别通过SIMCSE模型得到S’和T’各自对每个字的表示$h^s_{CLS}, h^s_1, h^s_2, \cdots, h^s_n$和$h^t_{CLS}, h^t_1, h^t_2, \cdots, h^t_m$。其中，CLS为插入在句首的特殊符号，其对应的向量h_{CLS}代表整个句子的语义。以上图1为例加以解释，源文本S为“既为先觉之民”，目标文本T为“天民之先觉者也”。送给SIMCSE模型的输入分别为S’=[CLS，既，为，先，觉，之，民]，T’=[CLS，天，民，之，先，觉，者，也]。经过SIMCSE模型的编码，对于S’和T’中任意一个字符，都会得到一个768维（维度与所选预训练模型设定有关）的向量。该向量的每一维度均以一个实数表示。

之后，对于S和T中直接匹配的k-gram（k元组，长度为k的片段），本研究通过并联的方式得到k元组在S和T中分别的向量表示$h^s[i:i+k-1]$和$h^t[j:j+k-1]$，并计算$h^s[i:i+k-1]$和$h^t[j:j+k-1]$之间的余弦相似度。同直接匹配判断中的计算方法一样，如果k元组中出现了停用词，那么本研究对该k元组计算得到的相关度分数做减半惩罚。为了避免因为目标文本长度长而带来的匹配概率过高的问题，本研究对所有的k元组的余弦相似度求和并通过目标文本k元组数量

（$m-k-1$）进行规范化得到匹配的分数。具体计算方法如下：

$$score_k = \sum_i \frac{cos([h_i^s, h_{i+1}^s, \cdots, h_{i+k-1}^s] : [h_j^t, h_{j+1}^t, \cdots, h_{j+k-1}^t]) + 1}{2 \times (m-k-1)} \quad if\ s[i:i+k-1] = t[j:j+k-1] \tag{3}$$

注意，这里m指的是目标文本的长度，$m-k-1$代表目标文本中k元组的个数。

在上面的例子中，“先觉”是直接匹配的最长片段，此时本研究从S’和T’对应的向量中取出各自“先”和“觉”对应的向量h_3^s，h_4^s和h_4^t，h_5^t。本研究把维度均为768的两个向量h_3^s，h_4^s合并成一个1536维的向量，把同样维度均为768的两个h_4^t，h_5^t也合并成1536维的向量（本研究称这个操作为向量并联）。这样“先觉”在S’和T’两端均有了一个1536维的向量。这两个向量尽管均是代表“先觉”一词的含义，但是因为结合了S和T中不同的上下文，因此它们的向量数值并不相同。为了衡量在两种语境下“先觉”代表的语义的相关性，本研究采用余弦相似度来对这两个向量进行比较。为了使余弦相似度的范围保持在0 ∼ 1这个范围内，本研究对余弦相似度数值做了加1后除以2的规范化操作。因为只有一个二元组的匹配，这样就得到了$score_2$为上面计算得到的分数除以6，其中6为目标文本T中二元组的个数。

最后，对于不同长度的k元组得到的$score_k$本研究根据经验设定权重并进行加权求和，得到最终的基于预训练模型的字符片段语义方法判断是否存在引用的分数。在实际的设定中，一元组对应的权重为0，即因为一元组匹配带来的噪声太大，本研究不考虑一元组的语义匹配，二元组的权重为0.2，三元组的权重为0.3，四元组的权重为0.5:

$$score_{ngram} = 0 \times score_1 + 0.2 \times score_2 + 0.3 \times score_3 + 0.5 \times score_4 \tag{4}$$

4.3 基于预训练语言模型的句子语义匹配方法

在本节中，本文提出直接使用代表句子粒度语义含义的CLS特殊符号对应的经SIMCSE模型编码后的隐向量h_{CLS}匹配的方式来判断是否存在引用关系（如图1所示）。Gao等人(Gao et al., 2021)指出，通过对比学习目标获得的句子级别的表示能够更好地区分和表达不同句子之间的语义关系。具体来说，我们使用在4.2节中提到的h_{CLS}^s和h_{CLS}^t表示，并计算它们之间的余弦相似度作为句子级别的语义相似度分数:

$$score_{sentence} = \frac{cos(h_{CLS}^s, h_{CLS}^t) + 1}{2} \tag{5}$$

4.4 复合判断方法

为了结合前面提到的三种基本判断方法的优点，本文提出使用基于三种判断方法的复合判断方法。按照组合关系，我们共得到四种复合模型：规则+字符，规则+句子，字符+句子，规则+字符+句子。其中，“规则”指4.1节中提到的结合专家知识的文本匹配规则方法，“字符”指4.2节中提到的基于预训练语言模型的字符片段语义匹配方法，“句子”指4.3节中提到的基于预训练语言模型的句子语义匹配方法。在复合判断方法中，本研究将三种方法得到的判断分数进行加权平均。在实际的计算中，为了简单起见，本研究将不同方法各自的分数均按照0.5的权重进行平均。并根据在开发集上的实验结果设定一个阈值来实际判断是否存在引用关系。

5 实验

5.1 实验设定

本文采用的BERT模型是在四库全书古文语料上进行预训练的Sikubert[2]（base）。其中的隐藏层向量维度是768，模型层数是12层，多头注意力机制（multi-head attention）的头数是12。在本文相关数据集上继续训练（SIMCSE方法和掩码语言模型）的轮数是5轮。

[2]https://github.com/hsc748NLP/SikuBERT-for-digital-humanities-and-classical-Chinese-information-processing

5.2 评价指标

本文采用正确率（accuracy）、准确率（precision）、召回率（recall rate）、F1值以及ROC-AUC值作为实验结果的评价指标。并以ROC-AUC作为主要评价标准。

正确率（accuracy）是衡量模型表现的一个常用指标，它的定义为：

$$ACC = \frac{right\ number}{total\ number}$$

F1值的计算依赖于准确率和召回率，F1值是对准确率和召回率的几何平均，反映了在某个阈值下相对平衡的模型表现。这三个指标的计算方式如下：

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \qquad F1 = \frac{2 \times Precision \times Recall}{Precision + Recall}$$

其中，TP代表True Positive，FP代表False Positive，FN代表False Negative。

但是，无论是正确率还是F1值等都会受到分类器中阈值选取的影响，因而，ROC-AUC指标常被用来衡量模型在不同阈值下的实际表现。ROC-AUC值根据预测分数计算接收器工作特性曲线下的面积。这个计算指标可以对于二值分类中不同的阈值取值下，得到不受阈值影响的分类器性能，本研究选取ROC-AUC这个指标来作为衡量模型实际分类性能的主要指标。理论上，以上指标都是越高越好。

5.3 判断的两种粒度：切分后的单句和自然段落

考虑到人文领域研究者对于结果的实际需求，除了上面介绍的对拆分后的单句级别文本进行是否包含引用的判断，本研究还提出按照通行本文献的分段方式进行引用判断。为了避免段落长度对判断模型的干扰，本研究采取一种简单直接的方法来判定段落中是否含有引用，即如果段落中任意一个单句级别的文本被判断为包含引用，那么本研究判定整段文本包含引用。

5.4 实验结果

在表1中，本文给出单一基线方法在测试集上以单句和段落为单位的模型表现。段落级别也就是判断一段中至少存在一处引用，这种设定更加接近人文领域的实际需求。从表中数据可以看出，基于预训练语言模型不同语义粒度的两种方法（字符、句子）单独使用均好于基于规则的方法，且“字符”和“句子”两种关注不同语义粒度的方法表现大体相当。在以段落为单位进行判断时，各项指标相比以单句为单位时均有所上升，尤其是F1值上升更为明显。单一方法的F1值即可达到80左右。

在表2中，本研究进一步给出使用复合方法得到的结果。当将三种单一基线方法结合，得到复合基线方法后，可以看到，效果均有较大幅度提升。说明三种单一方法之间关注的语义层次均具有互补之处。但是将三种方法同时结合在一个复合模型中的效果则不如只使用两种基于预训练语言模型的方法。此外可以看出，在以段落为单位时，将两种基于预训练语言模型的方法复合起来可以获得最好的效果，达到超过90的ROC-AUC值（91.02），且正确率达到了83.31，F1也超过了86，达到了能够实际帮助人文学者得出可靠结论的水平。而再结合规则后，效果反而会有所下降，这与基于预训练语言模型的方法具有更好的泛化性和迁移性有关。

5.5 实验分析

5.5.1 使用对比学习目标的效果

在图5.5.1中我们展示了使用不同模型在测试集上两种粒度（单句、段落）下的ROC-AUC值。其中，Sikubert指直接使用原始Sikubert预训练语言模型得到两种粒度的表示，Fine Tune指在领域内（二程和十三经文本）使用掩码语言模型进行领域适应性训练5轮的模型，SIMCSE指同时使用掩码语言模型和基于dropout的对比学习在领域内文本进行适应性训练5轮的模型。目标匹配方法使用表现最好的“字符+句子”方法。

从图中可以看出，在领域内进行适应性训练在单句级别上的效果提升尤其明显，因为这种复合方法中的字符片段级别匹配依赖于每个字符的正确学习和表示，而在领域内进行继续训练可以使得对字符的表示更贴近领域真实的分布。此外，可以看出，使用带有对比学习目标的SIMCSE模型相比单纯使用掩码语言模型作为目标的模型又有显著提高，这也验证了对比学习目标的有效性。

单句	Acc	Precision	Recall	F1	ROC-AUC
规则	56.40	45.20	90.75	60.34	64.01
字符	**76.49**	**68.79**	65.31	**67.01**	**81.65**
句子	72.83	60.26	**75.33**	66.96	81.57
段落	Acc	Precision	Recall	F1	ROC-AUC
规则	64.93	63.47	96.92	76.70	57.41
字符	**76.72**	**79.70**	81.75	80.71	**84.93**
句子	76.11	75.83	**87.92**	**81.43**	83.52

Table 1: 单句和段落级别**单一**基线方法对引用判断的结果。字符代表基于预训练语言模型字符片段语义的判断方法，句子代表基于预训练语言模型句子语义的判断方法。

单句	Acc	Precision	Recall	F1	ROC-AUC
规则+字符	75.56	64.43	74.01	68.89	82.54
规则+句子	75.64	63.37	79.07	70.36	83.61
字符+句子	**80.64**	**72.40**	75.99	**74.15**	**87.83**
规则+字符+句子	79.15	68.61	**79.19**	73.52	85.99
段落	Acc	Precision	Recall	F1	ROC-AUC
规则+字符	77.64	76.59	89.97	82.74	86.18
规则+句子	77.18	75.53	91.26	82.65	85.63
字符+句子	**83.31**	**82.56**	91.26	**86.69**	**91.02**
规则+字符+句子	81.01	79.51	**91.77**	85.20	89.99

Table 2: 单句和段落级别**复合**方法对引用判断的结果。字符代表基于预训练语言模型字符片段语义的判断方法，句子代表基于预训练语言模型句子语义的判断方法。

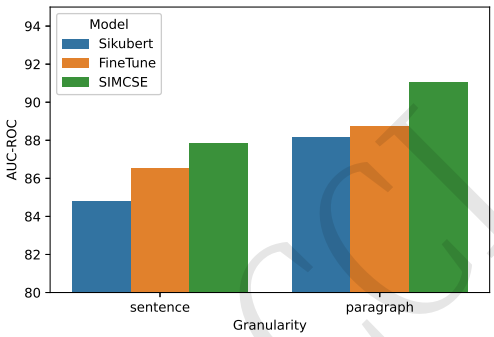

Figure 2: 不同模型在句子级别和段落级别的ROC-AUC值，SIMCSE为加入对比学习训练目标后的方法，匹配方法使用表现最好的“字符+句子”方法。

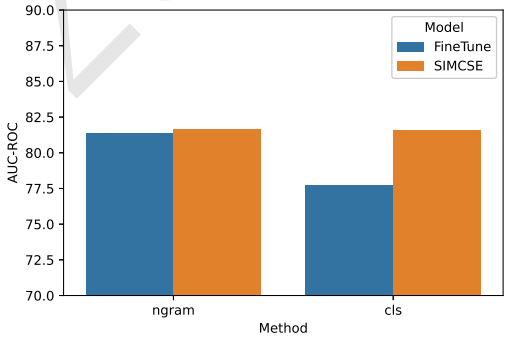

Figure 3: 使用对比学习目标在两种方法下的差别。ngram代表基于预训练语言模型的字符片段语义匹配方法，cls代表基于预训练语言模型的句子语义匹配方法。

从图5.5.1中可以进一步看出，加入对比学习目标后，相比于只做掩码语言模型训练，提升主要体现在句子级别的语义提升上。而在字符片段级别的语义上基本没有改变。这是因为SIMCSE中的对比学习目标主要针对的是整句文本表示的学习，而对字符级别影响不大。我们还可以看出，加入对比学习目标前，使用句子级别的语义进行判断效果与使用字符片段相比有较大差距，而在SIMCSE加入对比学习目标后，使用两种级别的语义判断效果基本相当。

5.5.2 使用停用词的效果

在图5.5.2中，我们展示了是否使用停用词惩罚的ROC-AUC效果差别。可以看出，尽管差距不大，但是使用停用词惩罚后，无论在单句粒度上还是段落粒度上均有稳定的提升。这说明引入带有人文专家经验的停用词可以在一定程度上提升引用查找的效果。在图5.5.2中，我们进

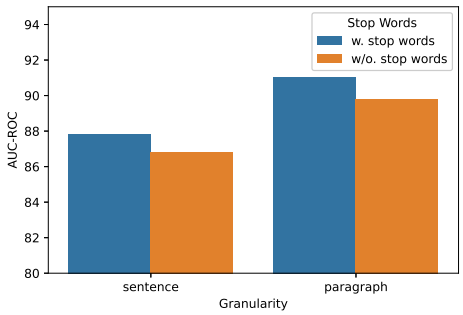

Figure 4: 是否使用停用词惩罚的ROC-AUC效果。引用查找方法选择的是基于预训练语言模型的“字符+句子”方法。

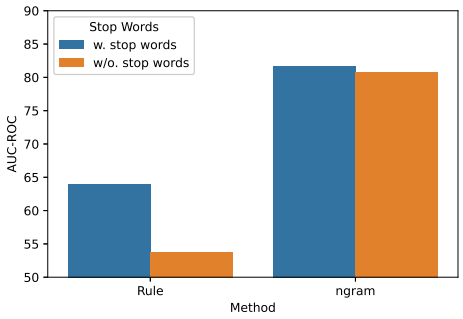

Figure 5: 对于字符片段方法和规则方法是否使用停用词ROC-AUC效果。

二程（正确）	得此道而不者，仁者之事也；因其不，故曰此仁也
经典（论语）	子曰：君子道者三，我能焉：仁者不，知者不惑，勇者不
二程（错误）	古人祭祀用尸，有深意，不可不深思
经典（礼记）	子云：祭祀之有尸也，宗之主也，示民有事也
二程（错误）	畏天命，可以不失付畀之重
经典（论语）	孔子曰：君子有三畏：畏天命，畏大人，畏人之言

Table 3: 字符片段和句子语义复合引用方法判断的代表性样例

一步给出了规则方法（Rule）中和基于预训练语言模型的字符片段语义匹配方法（ngram）中是否使用停用词的差别，从图中可以看到，规则方法对停用词更加敏感，而尽管影响较小，字符片段匹配方法也会在一定程度上受到是否停用词的影响。

5.6 样例分析

在表3中，我们给出了几个具有代表性的例子。在第一个例子中，经典（论语）原文是“仁者不（忧）”，而二程的解释里并没有出现这四个字，而是分开解释了“忧”与“仁者”之间的关系。单看“仁”这个概念的话，实际上会在经典文献和二程文本中大量出现，属于儒家思想的核心概念。无论对于专家人工检索（关键词“仁”和“忧”均存在大量干扰项）还是使用规则方法均存在很大的挑战。我们提出的基于预训练语言模型和对比学习目标的方法因为能够对整个句子的语义进行建模而很好地解决了这个问题。

对于后面两个例子来说，模型给出了错误的预测。对于二程文献中提到的“祭祀用尸”在先秦两汉文章中多次提到过，这里二程可能是针对某一处经典所发的议论，也有可能泛泛地使用“祭祀用尸”这个说法。即便在人工的引用查找过程中，这类句子也要结合上下文来综合判断，对模型来说难度也非常大。这部分引用查找模型只能给出可能存在的引用，而需要继续由专家校对完成判断。对于第三个例子来说，关键信息“天命”在多数句子中都能担任辨识的依据，所以不能作为停用词处理，但是“天命”这种说法在二程文献和经典文献中都非常多，在很多情境中其存在可以视为噪音，也容易给判断造成干扰。

6 结论

本文主要研究了如何自动化探测中国古代思想家对早期经典文献的引用，并给出了引用的明确定义。本文提出了多种结合专家知识和基于预训练语言模型和对比学习目标的无监督基线方法来自动查找中国古代思想史中思想家在阐发思想时对早期文献的引用。为了验证方法的有效性，本文以二程对早期儒家经典文献的引用为例，构建并发布二程对早期儒家经典引用的数据集并在该数据集上进行了大量实验。实验结果表明本文提出的基于预训练语言模型的字符片段和句子复合方法可以有效地找到大多数引用，并且能够为提高专家人工精确查找效率提供有效帮助。本文的研究成果在集注集释整理、文本生成溯源、重出文献查找、引用统计分析、索引文献集制作等方面具有广阔的应用前景。

致谢

本成果受国家自然科学基金项目（61872402）,教育部人文社科规划基金项目(17YJAZH068)，中央高校基本科研业务费(北京语言大学梧桐创新平台，21PT04)，模式识别国家重点实验室开放课题基金资助。

参考文献

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. SimCSE: Simple contrastive learning of sentence embeddings. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Ashish Jaiswal, Ashwin Ramesh Babu, Mohammad Zaki Zadeh, Debapriya Banerjee, and Fillia Makedon. 2020. A survey on contrastive self-supervised learning. *CoRR*, abs/2011.00362.

Xipeng Qiu, Tianxiang Sun, Yige Xu, Yunfan Shao, Ning Dai, and Xuanjing Huang. 2020. Pre-trained models for natural language processing: A survey. *CoRR*, abs/2003.08271.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with bertserini. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 72–77.

Zhuosheng Zhang, Yuwei Wu, Hai Zhao, Zuchao Li, Shuailiang Zhang, Xi Zhou, and Xiang Zhou. 2020. Semantics-aware bert for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 9628–9635.

俞敬松, 魏一, and 张永伟. 2019. 基于bert的古文断句研究与应用. 中文信息学报, 33(11):7.

刘江峰, 冯钰童, 王东波, 胡昊天, and 张逸勤. 2021. 数字人文视域下sikubert增强的史籍实体识别. 图书馆论坛.

周好, 王东波, and 黄水清. 2021. 古籍引书上下文自动识别研究——以注疏文献为例. 情报理论与实践, 44(9):7.

徐润华, 王东波, 刘欢, 梁媛, and 陈康. 2022. 面向古籍数字人文的《资治通鉴》自动摘要研究——以sikubert预训练模型为例. 图书馆论坛.

王东波, 刘畅, 朱子赫, 刘江峰, 胡昊天, 沈思, and 李斌. 2022. Sikubert与sikuroberta:面向数字人文的《四库全书》预训练模型构建及应用研究. 图书馆论坛.

耿云冬, 张逸勤, 刘欢, and 王东波. 2022. 面向数字人文的中国古代典籍词性自动标注研究——以sikubert预训练模型为例. 图书馆论坛.

胡昊天, 张逸勤, 邓三鸿, 王东波, 冯敏萱, 刘浏, and 李斌. 2022. 面向数字人文的《四库全书》子部自动分类研究——以siku bert和siku ro berta预训练模型为例. 图书馆论坛.

葛瑞汉. 2000. 二程兄弟的新儒学:中国的两位哲学家. 二程兄弟的新儒学:中国的两位哲学家.

黄俊杰. 2018. 东亚儒家经典诠释史中的三个理论问题. 山东大学学报：哲学社会科学版, (2):8.

黄水清, 周好, 彭秋茹, and 王东波. 2021. 引书的自动识别及文献计量学分析. 情报学报, 40(12):13.