数字人文视角下的《史记》《汉书》比较研究

邓泽琨[1,2], 杨浩[2], 王军(✉)[1,2]
[1] 北京大学信息管理系/ 北京市
[2] 北京大学数字人文研究中心/ 北京市
{dzk,yanghao2008,junwang}@pku.edu.cn

摘要

《史记》和《汉书》具有经久不衰的研究价值。尽管两书异同的研究已经较为丰富，但研究的全面性、完备性、科学性、客观性均仍显不足。在数字人文的视角下，本文利用计算语言学方法，通过对字、词、命名实体、段落等的多粒度、多角度分析，开展对于《史》《汉》的比较研究。首先，本文对于《史》《汉》中的字、词、命名实体的分布和特点进行对比，以遍历穷举的考察方式提炼出两书在主要内容上的相同点与不同点，揭示了汉武帝之前和汉武帝到西汉灭亡两段历史时期在政治、文化、思想上的重要变革与承袭。其次，本文使用一种融入命名实体作为外部特征的文本相似度算法对于《史记》《汉书》的异文进行自动发现，成功识别出过去研究者通过人工手段没有发现的袭用段落，使得我们对于《史》《汉》的承袭关系形成更加完整和立体的认识。再次，本文通过计算异文段落之间的最长公共子序列来自动得出两段异文之间存在的差异，从宏观统计上证明了《汉书》文字风格《史记》的差别，并从微观上进一步对二者语言特点进行了阐释，为理解《史》《汉》异文特点提供了新的角度和启发。本研究站在数字人文的视域下，利用先进的计算方法对于传世千年的中国古代经典进行了再审视、再发现，其方法对于今人研究古籍有一定的借鉴价值。

关键词： 数字人文；《史记》；《汉书》；命名实体；异文；文本相似度

A Comparative Study of *Shiji* and *Hanshu* from the Perspective of Digital Humanities

Zekun Deng[1,2], Hao Yang[2], Jun Wang(✉)[1,2]
[1] Department of Information Management, Peking University / Beijing
[2] Research Center for Digital Humanities of PKU / Beijing
{dzk,yanghao2008,junwang}@pku.edu.cn

Abstract

Shiji and *Hanshu* have been studied extensively throughout the past centuries. Although the similarities and differences of the two works have been researched in numerous literatures, these studies are limited in their collectiveness, comprehensiveness, objectiveness and rigor. Under the sight of digital humanities, this paper attempts to adopt computational linguistic methods to compare *Shiji* and *Hanshu* in a multi-scale and multi-perspective way by analyzing the characters, words, named entities and paragraphs in the books. Firstly, this paper compares the distribution and characteristics of the characters, words and named entities in *Shiji* and *Hanshu*, finding out the major similarities and differences of their contents by an exhaustive enumeration, revealing the significant political, cultural and ideological transformations from pre-2th

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。

century B.C. to the rest of Western Han dynasty. Secondly, this paper adopts a text similarity metric incorporating named entity as an external feature to automatically probe variant readings in *Shiji* and *Hanshu*. We manage to discover variant readings that have not been found by past researchers who rely solely on manual approaches, obtaining much richer knowledge of *Hanshu*'s inheritance of *Shiji*. Thirdly, this paper derives the differences between the variant readings of *Shiji* and *Hanshu* automatically by computing their longest common subsequences. Based on the results, this paper rigorously proves the writing style discrepancies of the two books through a macroscopic statistical analysis and interprets their linguistic features respectively with microscopic example texts, providing new perspective and insights on the variant readings of the two books. In conclusion, under the sight of digital humanities, this paper employs advanced computational methods to reexamine and reexplore centuries-old ancient Chinese classics, bringing enlightenment to moderners about new approaches for studying ancient literatures.

Keywords: digital humanities , *Shiji* , *Hanshu* , named entity , variant reading , text similarity

1 引言

《史记》和《汉书》是中国古代史籍的经典之作，在文学、历史学、语言学等领域具有宝贵且不可替代的研究价值。《史记》和《汉书》有诸多相似之处。从二者记载的历史时段上看，《史记》记载的历史横跨三皇五帝到汉武帝时期，而《汉书》则写的是整个西汉的历史，同时也包括秦朝末年至西汉建立之前的部分事件，因此两书记录的历史在时间上存在大幅重合。从二者的体例上看，二者都是以人物传记为主的纪传体史书。当然，《史记》和《汉书》也有很多方面的差异。两书作者不同的写作动机、行文风格，两书所载历史时期的差异，后世文学家和历史学家对两书的不同看法等等，也都是有价值的研究主题，使得两书的相同和相异之处交织，让《史记》和《汉书》的比较研究成为了有意义的研究问题。

尽管前人对这两部经典的研究已经相当丰富，但是，这些研究仍然存在一些不足。一方面，过往的研究在对两书文本进行分析时，往往通过举例的方法进行论证，采用"以点带面"的模式，用个别的例子来分析得出结论。这种研究方法虽然能够得到可信的结论，但是在全面性和完备性上有所欠缺，有可能忽略某些重要的文本细节。另一方面，人文学者在研究《史记》和《汉书》时采用的定性方法往往具有较强的主观性，其结论往往受制于学者自身的知识储备，并且其分析过程时常具有随意性和偶然性，研究的客观性和科学性有所不足。

近年来，数字人文对于人们阅读古代经典文献的方式产生了巨大的改变。一方面，数字人文使得我们可以采用量化计算手段对于大量文本进行提炼、抽象和概括，从而使得人们可以在不完整阅读全文的情况下把握一本书的主线和要旨。另一方面，数字人文将信息抽取、信息检索等现代计算机科学技术引入人文研究，使得机器可以自动发现卷帙浩繁的古代文献之间潜藏的关系和知识，可以帮助我们做出以前人的能力无法达成的新发现、新洞察。

在数字人文的视域下，本文试图利用计算语言学方法，通过对字、词、命名实体、段落等的多粒度、多角度分析，开展对于《史记》和《汉书》的比较研究。本研究的整体流程如Figure 1所示。

本文的主要贡献如下：

1. 本文利用基于深度学习的古汉语分词和命名实体识别模型对《史记》《汉书》进行处理，对于《史记》《汉书》中的字、词、命名实体的分布和特点进行对比，以遍历穷举的考察方式提炼出两书在主要内容上的相同点与不同点，通过对典型实例的深入分析，揭示了汉武帝之前和汉武帝到西汉灭亡两段历史时期在政治、文化、思想上的重要变革与承袭。

2. 本文利用以命名实体作为外部特征的文本相似度计算方法对于《史记》《汉书》的异文进行自动发现，其结果表明本算法的发掘结果不但与过往学者人工发现的结果相吻合，并且能

本研究得到国家自然科学基金国际重点合作项目" 中国儒家学术史知识图谱构建研究"(项目号:72010107003)的支持

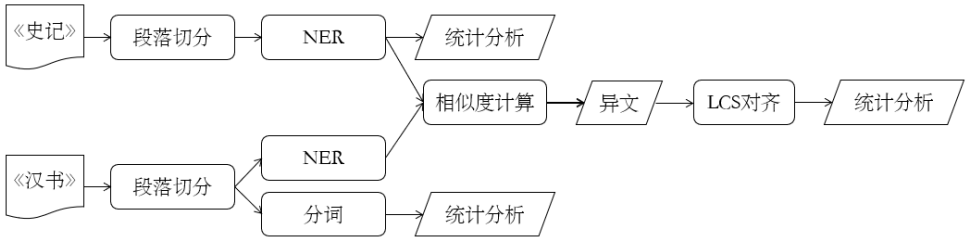

Figure 1: 研究整体流程

够发现过去研究者通过人工手段没有发现的袭用段落。这使得我们对于《史记》《汉书》的承袭关系形成更加完整和立体的认识。

3. 本文通过计算异文段落之间的最长公共子序列来自动得出两段异文之间存在的差异，首先从统计上证明了《汉书》文字风格相比于《史记》的差别，随后选取了《史记》《汉书》的若干典型异文段落进一步对二者的语言特点进行了阐释，为理解《史记》《汉书》的异文特点提供了新的角度和启发。

2 文献综述

2.1 《史记》《汉书》比较研究

《史记》与《汉书》的比较研究具有悠久的历史。关于《史》《汉》二书的优劣问题，早在汉朝就有文人论述。东汉的王充[0]、晋人张辅(朱一玄, 刘毓忱, 2012)、南宋郑樵[1]、邵博[2]等皆从不同角度对于《史》《汉》做了比较。在清朝以前，《史》《汉》研究多停留在争论二者"孰优孰劣"的程度，而到了清及以后，《史》《汉》研究逐渐走向更加客观和科学的视角，"互有得失论"占据上风，且大量学者开始进行仔细的考据，着重于对文本进行更细致的分析。清代涌现出大量专门论述马、班异同的著作或著作章节，包括周中孚《补班马异同》、王筠《史记校》、蒋中和《马班异同议》等(曾小霞, 2009)。到了20世纪，吴福助(1975)著《史汉关系》一书对二者进行全面探讨。朴宰雨(1994)著《〈史记〉〈汉书〉比较研究》一书，详尽系统地对《汉书》袭用《史记》各卷的情况用表格进行了一一列举说明，具有很高的参考价值。

21世纪以来，《史》《汉》对比研究迈上了新的台阶。在文献学、语言学领域，沙志利(2005)遵循文献学方法，以《史》《汉》的详细比勘为依据，从史料、文字、思想等多个方面对两书进行了分析，从形式和内容两个角度分析了《史》《汉》异同产生的内因和外因。王海平(2003)总结了《史》《汉》异文在字、词、句三个层次上的表现形式，指出了其在文献学和语言学领域的运用。张明月(2021)研究了《史》《汉》重合篇章的文字差异和文学解读。张添雅(2021)梳理了《汉书》八表对于《史记》十表的承袭和创新之处。在文学领域，也有研究者(曾小霞, 2012; 诸雨辰, 2016; 夏德靠, 2019)从叙事、人物塑造、思想等角度对《史》《汉》进行了比较。

2.2 文本相似度算法和异文发现

在古典文献学中，"异文"有多种含义，一种是指一本书中的某一段文字因为传抄而产生的不同版本，另一种则是指记载同一件事但措辞有差异的字句。由于异文存在字词和语义上的相似性，因此我们可以利用文本相似度算法对古书中的异文进行自动发现。传统上的语义文本相似（Semantic Textual Similarity，STS）(Agirre et al., 2013)任务采用词袋模型（Bag of words）或者TF-IDF(Ramos, 2003)方法将文本转化成实值向量，通过计算向量之间的接近程度来判断文本语义的相似性。例如，肖磊和陈小荷(2010)用bigram计算句子相似度在春秋三传中进行异文寻找。李越(2014)利用改进的编辑距离算法与事件信息标注相结合，利用事件数据库对语料进行人物、地点、时间标注，加权计算文本相似度，实现《左传》《史记》异文发现。然而，该论文并未明确给出事件相似度的计算公式，且其方法对于数据库的依赖过强。近年来，神经网络开始成为STS的主流方法。梁媛等(2021)用《春秋》和春秋三传建立了异文平

[0] 《论衡》
[1] 《通志》
[2] 《邵氏闻见后录》

行语料，训练了BERT模型判断两句话是否为异文，但该方法的语料切割粒度过小，准确率不高，且需要大量人工标注数据。

本文提出以命名实体为外部特征的文本相似度计算方法，该方法不需要人工监督数据，且不存在输入长度限制，能够适应长文本比较的需要。

3 《史》《汉》字、词和命名实体的统计对比

3.1 《史》《汉》字频分布对比

欲全面了解《史记》和《汉书》的异同，两本书在用字和用词上的差异是必须注意的。本文首先对两书用字总数进行了统计。在将异体字合并后，本文统计得到《史记》全书共使用了4619个不同的字，而《汉书》则用了5343个，与《史记》相比增加了15.7%。

本文对两书中频率最高的15个字进行了统计和对比，其结果见附录A Table 7。本文发现，在这些字中，有3个字在两书中的频率差异较大，它们分别是“王”、“子”和“公”。具体而言，三个字在《汉书》中的频率与《史记》中相比分别低了6.28‰、5.13‰和5.50‰。Table 1展示了两书中包含这三个字的最常见词语（分词的方法见3.2节），按照出现频率由高到低排序。综合这些数据可以看出，以上三个字在两书中的频率的差异显然不只是偶然造成的。为了更充分地理解这种差异出现的原因，本文进一步统计了这三个字在命名实体内出现的频率（即在命名实体内出现的次数除以全书总字数），如Table 2所示。可以看出，这三个字的频率差异很大程度上可以由包含这三个字的命名实体占全文比例的降低来解释。结合以上信息，本文猜测，这一变化可能是来源于两书所叙历史时期的政治形势、政治制度和所涉人物的差异。

字	包含该字的最常见的词	
	《史记》	《汉书》
王	王 汉王 大王 秦王 赵王 齐王 项王 楚王	王 汉王 王莽 大王 王者 齐王 赵王 淮南王
子	子 太子 天子 孔子 公子 君子 子孙 弟子	子 天子 太子 孔子 子孙 父子 君子 弟子
公	公 公子 沛公 太史公 桓公 文公 周公 景公	公 公卿 沛公 周公 公主 三公 安汉公

Table 1: 《史记》《汉书》中包含“王”、“子”、“公”三个字的最常见词

字	在全文中			在命名实体中		
	《史》	《汉》	Δ	《史》	《汉》	Δ
王	16.36	10.08	-6.28	11.09	6.65	-4.43
子	12.96	7.83	-5.13	6.22	3.00	-3.22
公	9.62	4.12	-5.50	7.85	2.91	-4.94

Table 2: 《史记》《汉书》中“王”、“子”、“公”在全文和在命名实体中的词频(‰)及差值(Δ)

3.2 《史》《汉》词频分布对比

本研究对《史记》和《汉书》的词频分布进行了对比。本文中《史记》的词频统计利用的是台湾“中央研究院”的《史记》人工标注语料，该语料已经由专家进行分词，可以直接进行统计。《汉书》的词频统计则利用了Tang and Su(2022)开发的古汉语分词模型，人工检验表明其准确率较高，可以用于统计分析。

本文首先统计了两书词语的平均词长和不同长度词语的比例，如附录A Table 8所示。可以发现，《汉书》相比于《史记》平均词长更大，单音节词的占比更低，而2字、3字、4字或以上的词的占比均更高。之后，本文分别统计了两书中频率最高的单音节词、双音节词、三音节词和四音节词（详细结果见附录A Table 9）。由于时间词不具有实际意义（例如“二年”、“六月”等），因此没有列入此表。

在高频词方面，《史》《汉》二书的高频词总体较为接近，但是又呈现出几处明显的不同。单音节词频率的分布和字频的分布比较相似，《史》《汉》的高频单音节词基本一致，说明《史》《汉》从语言学角度看大致处在同一时代，用字上的总体差异不大。双音节词中，《汉书》与《史记》相比，“诸侯”、“孔子”两个词的排名大幅下降，而“匈奴”、“丞相”、“单

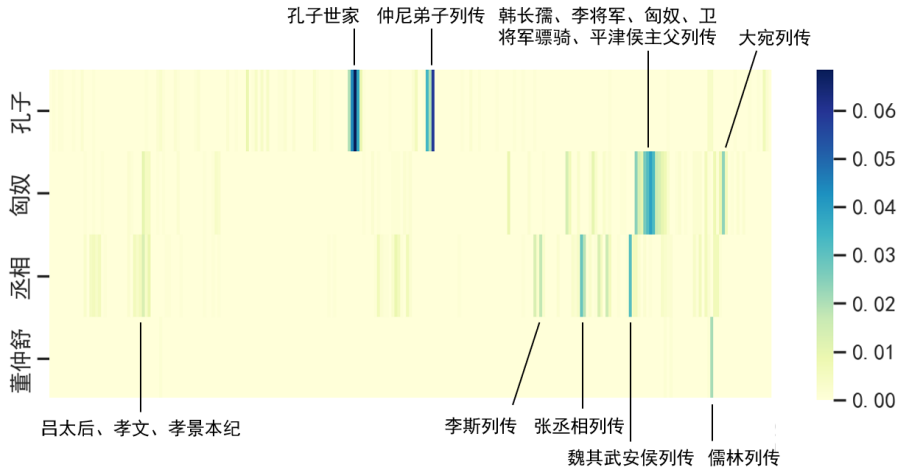

Figure 2: 四个典型词在《史记》全文的分布密度图

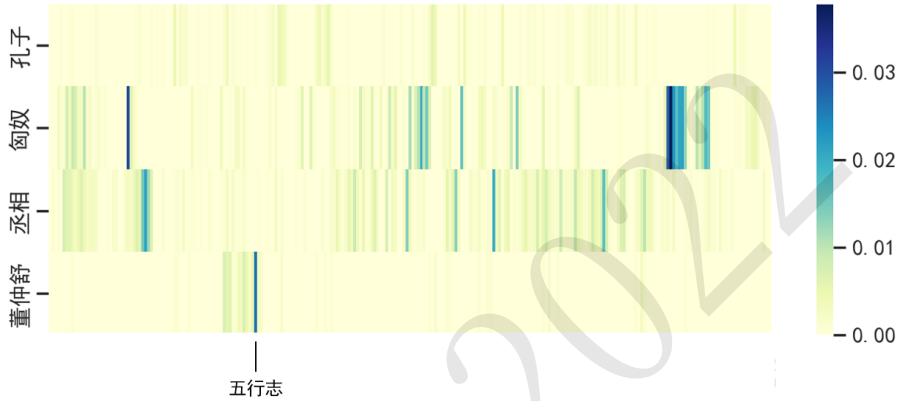

Figure 3: 四个典型词在《汉书》全文的分布密度图

于”三个词的排名大幅上升。结合历史来看，“丞相”和“诸侯”词频的一升一降很大程度上可以归因于汉朝的大一统政治体制和地方分封势力的大幅削弱。长度为4的词中，仅有5个在《史记》和《汉书》中词频均排前十，其余仅在《史记》或《汉书》中进入前十的词均体现出明显的时代特征，反映了先秦和西汉两个时期主要历史人物和官职设置的差异。

以上结果中有几个词比较值得注意，即“孔子”、“匈奴”、“丞相”和“董仲舒”。为了更深入地理解这几个词的词频所体现的文本含义，本文绘制了它们在《史》《汉》两书全文中出现的密度图，如Figure 2和3所示。Figure 2和3将每本书分别按照字数从前到后均分成250份，然后将每个词在每一份中出现的密度按照高低不同赋以不同颜色。此处，记一段文字$\boldsymbol{t} = [a_1, ..., a_N]$，表示$\boldsymbol{t}$由$a_1, \ldots, a_N$这$N$个词组成，则一个词$w$在文字$\boldsymbol{t}$中的密度由以下公式定义：

$$\rho(w, \boldsymbol{t}) = \frac{\boldsymbol{t}.\text{count}(w) \cdot w.\text{length}}{\sum_{i=1}^{N} a_i.\text{length}},$$

其中$\boldsymbol{t}.\text{count}(w)$是$w$在$\boldsymbol{t}$中出现的次数，$w.\text{length}, a_i.\text{length}$ 分别是w和a_i的长度。不同颜色所对应的密度值可见图右侧的图例。

从Figure 2和3可以看出，所关注的四个词在两书中的分布各有特点。具体而言，“孔子”在《史记》中集中分布在《孔子世家》和《仲尼弟子列传》中，在该书的其他篇章中频率极低，而在《汉书》中出现次数寥寥。这表明尽管孔子是中国文化史上最重要的人物之一，但其在《史》《汉》两书中的出现仍然比较局限。在《史记》中，“匈奴”一词主要集中在《韩长孺列传》、《李将军列传》等6篇列传中，而在《汉书》中“匈奴”一词的分布较为分散，在纪、传中均大量出现，存在多个分布高峰。“丞相”一词在《史记》中集中在《吕太后本纪》、《孝文本纪》等6篇中，而在《汉书》中的分布状况与“匈奴”类似，在全书各个部分都有高频出现。在

《史记》中，“董仲舒”绝大多数出现在《儒林列传》中且总的出现次数不多，而在《汉书》中则大量集中于《五行志》。这表明董仲舒的“五行”思想在当时已经有了很大的影响。

以上四个词在《史》《汉》二书中的分布具有很大的启示性，因为它们很好地勾勒了两个不同历史时期之间政治、文化、思想领域的重要变化。从政治的角度看，“匈奴”分布的变化很大程度上呼应了匈奴在战国兴起、在西汉强盛并与中原政权展开持久拉锯的事实，其在《汉书》中频繁而分散的出现忠实地体现出匈奴势力在西汉政治舞台上所扮演的重要角色。“丞相”一词在《史记》中的集中出现几乎仅限于记述秦汉两朝历史的三篇本纪和三篇列传中，与丞相这一官职的历史沿革高度吻合：丞相一职初创于战国时期，当时各国虽有与丞相地位等同的“相国”一职，但是名称有所不同，而到了战国中后期，秦国才率先设立“丞相”一职，这一官职也在秦汉两朝得以沿袭(袁祖亮, 1988)。这也就很好地解释了为何“丞相”在《史记》中的出现非常局限，而在《汉书》中多有分布。从文化的角度看，“孔子”和“董仲舒”均是历史上重要的儒学家，而二者在《史》《汉》二书中分布频次的巨大差异，无疑暗示了二者在不同的时代所具备的影响力大不相同。

3.3 《史》《汉》命名实体频率分布对比

命名实体识别（Named Entity Recognition, NER）是一项重要的NLP任务。本文使用基于四库BERT的NER模型来对《史记》《汉书》中的命名实体进行识别。该模型采用BERT+BiLSTM+CRF（Conditional Random Field，条件随机场）架构，其中BERT参数通过《四库全书》语料预训练初始化，之后利用《资治通鉴》NER数据集对模型进行微调。该模型在《资治通鉴》测试集上的F1约98%，在《史记》上测试的F1值约为90%，表现出较高的精度和可用性。该模型的超参数和《四库全书》语料的详细信息见附录C。

在数字人文研究中，针对中国古代典籍的特点，我们一般将古汉语文本中的命名实体分为人物、地点、时间、书籍、官职等类别。命名实体区别于文本中其他字词和短语的最重要特点之一是它们都对应于现实世界中的一个真实存在的实体。因而，对语料中命名实体出现的频率和分布特点进行分析，能够在很大程度上反映出语料所述历史时期中的人、地、时等实体的特征，有助于我们透过表面洞察文字背后的历史脉络。因此，对于史料中的命名实体分布进行分析，对于我们更深入地发掘和理解史料具有重要的意义。

利用前文所述的命名实体识别模型，本文对《史》《汉》二书中每个实体类型频次最高的10个实体进行了统计，结果列在附录A Table 10中。该结果有丰富的解读空间。从人物来看，《史记》和《汉书》的高频人物实体存在一些差异，例如前者包含有“项羽（项王）”、“赵王”等仅在汉朝建立以前存在的人物，而后者则包含有“莽（王莽）”、“光”等主要活动于汉武帝以后的人物。以上差异并不令人惊奇，然而《汉书》排名前十的高频人物中还包含了“禹”、“汤”、“武”等先秦时期的人物，其生活时间与《汉书》所叙时间毫无重合。经过查找原始语料，本文发现，《汉书》中的“禹”、“汤”、“武”指的不全是三代的三位君王，实际上有时指的是汉朝的同名人物。客观上说，《汉书》中“禹”、“汤”、“武”三个人物实体的频繁出现，是《汉书》反复引用夏禹、商汤、周武三位君王的事迹和其他重名人物的存在这两个因素共同作用的结果。

从地点来看，《史记》中出现频次最高的十个地点实体中，除了“汉”以外，其余均为春秋和战国时期力量最强大的诸侯国的国名，而在《汉书》中则出现了“匈奴”、“长安”、“河”等与西汉政治关联更紧密的地点实体。

两部书频次最高的十个书籍实体差异总体不明显，都包含了《春秋》、《诗》、《书》、《易》等先秦时期的经典著作，且在两部书中都排名前四，表明了这些典籍在西汉建立前后均保持着重要的地位。但是，在《史记》中排名靠前的《老子》在《汉书》中并未进入前十，而《汉书》中《五经》和《左氏传》均出现在前十，这一结果与司马迁和班固二位作者本身的思想倾向不无关系，同时也某种程度上说明了西汉建立前后文人思想观念的微妙变化。

排名靠前的时间和官职实体均未显示太多区别。《史记》和《汉书》频次前十的官职实体有9个是相同的，除了这9个外，《史记》有“夫人”，而《汉书》有“太守”。这种现象的原因也容易解释：根据唐朝杜佑《通典》记载，“太守”原名“郡守”，秦朝行郡县制，在郡一级设立郡守一职，汉景帝年间更名为“太守”。因而，“太守”这一官职名实际上在西汉时期才广泛被使用，从而在《汉书》中出现的频次相比《史记》大幅上升。

总的来说，《史记》和《汉书》中出现的命名实体有同有异，呈现出了各自时代的独有特点，并在某种程度上反映了从秦之前到西汉时期历史发展的轨迹。我们使用模型以可观的准确

率对古代史籍中的绝大多数命名实体进行识别，用一种精确和客观的方式来描述《史》《汉》中的最主要人物、地点、官职等等历史主体，为探究人文问题提供了一种新的手段。这些结果有助于我们从一个新的角度来认识和理解这段历史。

4 计算语言学视角的《史》《汉》异文研究

4.1 《史》《汉》文本袭用的自动发现

4.1.1 方法

本文使用一种基于TF-IDF和命名实体的文本相似度算法来实现《史》《汉》二书异文的自动发现。对于两个字符串$\boldsymbol{p}_i = a_{i,1}...a_{i,N_i}, \boldsymbol{p}_j = a_{j,1}...a_{j,N_j}$，该方法可以输出二者的相似度$u(\boldsymbol{p}_i, \boldsymbol{p}_j) \in \mathbb{R}$。本方法描述如下：首先，利用3.3节介绍的NER模型得到每个字符串中包含的所有实体提及$\boldsymbol{e}_i$和$\boldsymbol{e}_j$，其中

$$\boldsymbol{e}_i = \{(a_{i,b}...a_{i,c}, h_{i,b,c}) \mid a_{i,b}...a_{i,c}\text{是一个类型为}h_{i,b,c}\text{的命名实体}\},$$

$h_{i,b,c} \in \boldsymbol{H}$，$\boldsymbol{H}$是所有实体类型标签的集合。然后，选取一个字符串集合$\mathcal{D}$，记$\mathcal{D}$中出现的所有n-gram为$\boldsymbol{g} = [g_1, ..., g_M]$，定义$g_i$的逆文档频率（Inverse Document Frequency, IDF）为$\text{idf}(g_i) = \log \frac{|\mathcal{D}|}{|\{\mathbf{d}|g_i\text{在}\mathbf{d}\text{中出现},\ \mathbf{d}\in\mathcal{D}\}|}$，其中$|\mathcal{D}|$是集合$\mathcal{D}$中元素的数量。记$n$-gram g_j在字符串$\boldsymbol{p}_i$中出现的频次为$\text{tf}(\boldsymbol{p}_i, g_j)$，则字符串$\boldsymbol{p}_i$的TF-IDF向量为

$$\mathbf{v}_i = [\text{tf}(\boldsymbol{p}_i, g_1) \cdot \text{idf}(g_1), ..., \text{tf}(\boldsymbol{p}_i, g_M) \cdot \text{idf}(g_M)]^{\mathrm{T}},$$

因此，字符串$\boldsymbol{p}_i, \boldsymbol{p}_j$的相似度可定义为

$$u(\boldsymbol{p}_i, \boldsymbol{p}_j) = \lambda \frac{\mathbf{v}_i^{\mathrm{T}} \mathbf{v}_j}{\|\mathbf{v}_i\| \|\mathbf{v}_j\|} + (1-\lambda)\frac{|\boldsymbol{e}_i \cap \boldsymbol{e}_j|}{|\boldsymbol{e}_i \cup \boldsymbol{e}_j|}, \tag{1}$$

其中λ是人为设定的常数且$\lambda \in [0, 1]$。

从公式(1)容易看出，这一相似度定义着重强调了两段文本中命名实体的重合度。当两段文字中所提到的人物、地点、官职等命名实体高度重合时，两段文字很有可能在讲述同样的历史事件，语义相似的可能性也更高。因而，与不考虑命名实体的方法相比，本方法相比更加适用于人、地、时等对象较为密集的史部文献。

本文将此方法与古籍异文发现的现有最好方法(梁媛等, 2021)的性能进行了对比。该论文提出使用预训练语言模型的有监督方法进行异文识别，具体而言是将每组待检测文本对同时输入BERT编码器，对输出隐向量进行分类以确定该文本对是否为同事异文（下称基线方法）。为了对比基线方法和本文方法的性能，本文遵循梁媛等(2021)的描述在《史记》《汉书》语料上复现了该方法。由于本文方法是无监督的，因此为了进行更好的对比，训练基线模型时正例和负例均只提供了1或3个训练样本（1-shot/3-shot）。模型的其余设定与原论文完全相同。测试使用的是人工标注的《史记》《汉书》异文段落数据集，总数约100条。测试结果如Table 3。可以看到，在少样本条件下，本文方法的P、R、F1明显高于基线方法。

此外，基线方法的时间复杂度也值得注意。异文发现场景要求在大量语料中寻找相似文本。若待发现的文本条数为n，则基线方法需要使用BERT进行$n(n-1)/2$次推理，其时间成本相当高昂。相比之下，本文的方法只需要少量的向量内积操作和n次BERT推理，时间成本低，在文本数量较大时仍能保持较高的可用性。

方法		P	R	F1
BERT	1-shot	57.01(± 18.45)	90.61(± 17.93)	66.29(± 9.03)
	3-shot	83.73(± 15.39)	91.47(± 8.71)	86.51(± 9.04)
本文方法		84.21	96.97	90.14

Table 3: 本文方法与现有方法的性能比较（括号内为10次重复实验的标准差）

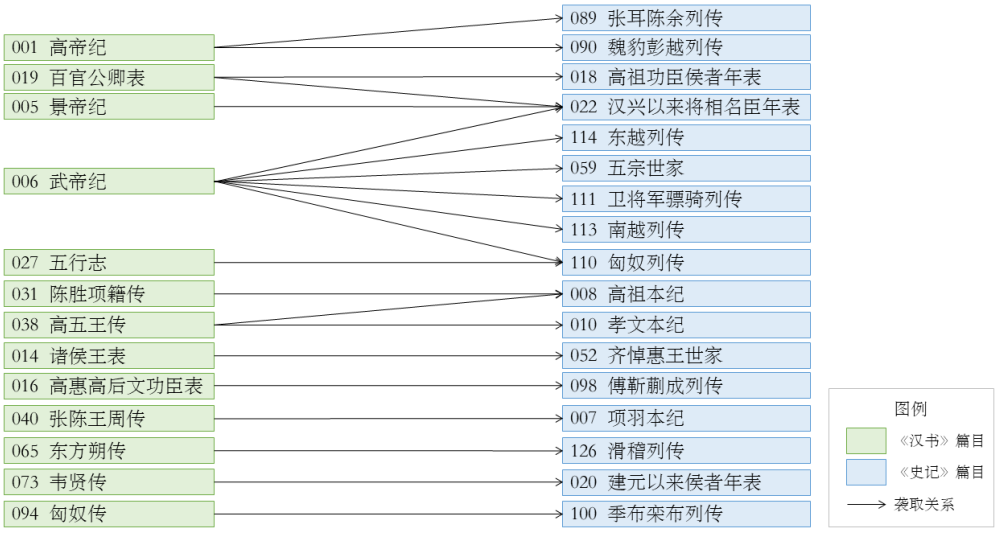

Figure 4: 本文新发现的《汉书》各篇承袭《史记》篇目（仅展示朴未记载的，其余省略）

4.1.2 结果

利用本文方法，将《史记》和《汉书》按照中华书局标点本的分段切割成段落，将得到的所有段落作为字符串集合$\mathcal{D}$，计算其中所有n-gram的IDF。对于《汉书》中的每一段文本$\boldsymbol{p}$，利用公式(1)计算其与《史记》中所有段落的相似度，并选出其中相似度最高的一段（记为$\boldsymbol{q}$），记二者的相似度为$u(\boldsymbol{p},\boldsymbol{q})$。设定阈值$\mu$，当$u(\boldsymbol{p},\boldsymbol{q}) > \mu$时，认为$\boldsymbol{p}$和$\boldsymbol{q}$是相似的。在本研究中，公式(1)中的加权参数取$\lambda = 0.7$，阈值$\mu = 0.25$。通过这一方法，本文发现了《汉书》共2082段文本与《史记》存在高度相似（长度不超过10个字的段落被排除了）。经过逐一人工比对确认后，本文对于《汉书》对《史记》可能的袭用情况进行了全面梳理。为了将结果与之前学者的研究作比较，本文以朴宰雨(1994)在《〈史记〉〈汉书〉比较研究》第三节第三部分“《汉书》各篇承袭《史记》之情况概述”为对照对象。经对照发现，本文的方法不但能够识别出朴宰雨书中所记录的所有袭用篇目，还发现了多达21对本书中没有记载的袭用篇章（见Figure 4）。

可以看到，Figure 4所呈现出的《汉书》袭用《史记》的情况较为复杂，下面本文对其中列出的几组典型的袭用案例进行详细阐释。

案例1：《高帝纪》袭用《张耳陈余列传》、《魏豹彭越列传》

朴宰雨在其书中指出，《高帝纪》“增补与改写之处甚多……为汉书中最大规模者”，并指出班固将《项羽本纪》、《留侯世家》、《韩信卢绾世家》等篇中有关刘邦之事移入了《高帝纪》。但是，朴的列举仍然不够全面，有所遗漏。事实上，《高帝纪》还袭取了《魏豹彭越列传》中五年冬十月张良为汉王献计使韩信、彭越引兵会师之事以及《张耳陈余列传》中高祖逃脱贯高谋杀阴谋之事。Table 4列出了《高帝纪》、《高祖本纪》、《张耳陈余列传》对逃脱谋杀之事的不同记载。虽然《高祖本纪》也记录了同一件事，但是并未记载高祖询问县名的对话。综合三段文本判断，《高帝纪》的这一段文字更可能是从《张耳陈余列传》袭取而来的。

篇名	内容
《汉书·高帝纪》	八年冬，上东击韩信馀寇于东垣。还过赵，赵相贯高等耻上不礼其王，阴谋欲弒上。上欲宿，心动，**问“县名何？”曰：“柏人。”上曰：“柏人者，迫于人也。”**去弗宿。
《史记·高祖本纪》	高祖之东垣，过柏人，赵相贯高等谋弒高祖，高祖心动，因不留。
《史记·张耳陈余列传》	汉八年，上从东垣还，过赵，贯高等乃壁人柏人，要之置厕。上过欲宿，心动，问**曰：“县名为何？”曰：“柏人。”“柏人者，迫于人也！”**不宿而去。

Table 4: 《史记》和《汉书》对于高祖逃脱贯高谋杀阴谋之事的不同说法

案例2：《武帝纪》袭用《汉兴以来将相名臣年表》、《五宗世家》、《匈奴列传》、《卫将军骠骑列传》、《南越列传》、《东越列传》

朴宰雨在书中指出，由于《史记》有录无书，且所补者“言辞鄙陋，非迁本意”，因此《武帝纪》并未袭用而是进行了重新创作。本文的结果表明，这一说法不够准确。事实上，本文发现《武帝纪》中有若干段落与《汉兴以来将相名臣年表》、《五宗世家》、《匈奴列传》等共6卷存在高度相似，经分析极有可能为班固袭用。为了更清楚地说明袭用情况，Table 5中列出了《武帝纪》和《匈奴列传》的四个相似段落中的一段。可以看到，《武帝纪》和《匈奴列传》的这些段落高度雷同，基本可以确信二者之间存在袭取关系。

《汉书·武帝纪》内容	《史记·匈奴列传》内容
夏五月，贰师将军三万骑出酒泉，与右贤王战于天山，斩首虏万馀级。又遣因将军出西河，骑都尉李陵将步兵五千人出居延北，与单于战，斩首虏万馀级。陵兵败，降匈奴。	其明年，汉使贰师将军广利以三万骑出酒泉，击右贤王于天山，得胡首虏万馀级而还。匈奴大围贰师将军，几不脱。汉兵物故什六七。汉复使因将军敖出西河，与强弩都尉会涿涂山，毋所得。又使骑都尉李陵将步骑五千人，出居延北千馀里，与单于会，合战，陵所杀伤万馀人，兵及食尽，欲解归，匈奴围陵，陵降匈奴，其兵遂没，得还者四百人。

Table 5: 《汉书·武帝纪》和《史记·匈奴列传》部分雷同段落

案例3：《东方朔传》袭用《滑稽列传》

朴宰雨在其书中认为，《东方朔传》为班固“根据另外的资料有意新创”，因而未将该传列入其袭用情况概述之中。本文认为，《东方朔传》并非由班固完全新创，而是有所承袭《滑稽列传》。现学界普遍认为，今本《史记》的《滑稽列传》只有一部分是司马迁原作，另一部分是由褚少孙补写的，而司马迁原稿中并未写东方朔，东方朔的事迹是褚少孙增补的(李林晓, 2020)。然而值得注意的是，褚少孙是西汉人，而班固是东汉人，因此在班固写作《汉书》时，他显然能够接触到褚少孙增补后的《史记》版本，从而在写作时加以袭用在逻辑上是完全可能的。事实上，本文发现《东方朔传》和《滑稽列传》存在如Table 6所示的雷同段落。综合以上事实，我们有理由认为《汉书·东方朔传》袭用了《史记·滑稽列传》。

《汉书·东方朔传》内容	《史记·滑稽列传》内容
客难东方朔曰：“苏秦、张仪一当万乘之主，而都卿相之位，……同胞之徒无所容居，其故何也？”	时会聚宫下博士诸先生与论议，共难之曰：“苏秦、张仪一当万乘之主，而都卿相之位，……官不过侍郎，位不过执戟，意者尚有遗行邪？其故何也？”
东方先生喟然长息，仰而应之曰：“是固非子之所能备也。彼一时也，此一时也，岂可同哉？……使苏秦、张仪与仆并生于今之世，曾不得掌故，安敢望常侍郎乎！故曰时异事异。	东方生曰：“是固非子所能备也。彼一时也，此一时也，岂可同哉！……使张仪、苏秦与仆并生于今之世，曾不能得掌故，安敢望常侍侍郎乎！传曰：‘天下无害灾，虽有圣人，无所施其才；上下和同，虽有贤者，无所立功。’故曰时异则事异。

Table 6: 《汉书·东方朔传》和《史记·滑稽列传》雷同段落

总的来说，过去的学者在研究《汉书》对《史记》的承袭情况时常常只关注明显的大段袭用而忽略较为零散琐碎的小段袭用，并且限于各种主、客观原因，遗漏了部分袭用关系。而本文利用算法自动地发现了若干过往学者未发现的袭用篇章，有效深化了我们对史汉关系的认知，使我们对于班固创作过程中对《史记》材料的剪裁、重排、整理工作有了更全面的认识。

4.2 基于最长公共子序列的《史》《汉》异文对比分析

最长公共子序列（Longest common subsequence, LCS）算法被广泛地用于解决文本比对问题。本文通过计算异文的LCS来分析《史记》《汉书》异文段落的差异，从宏观统计和微观案例两个角度进行分析阐释。

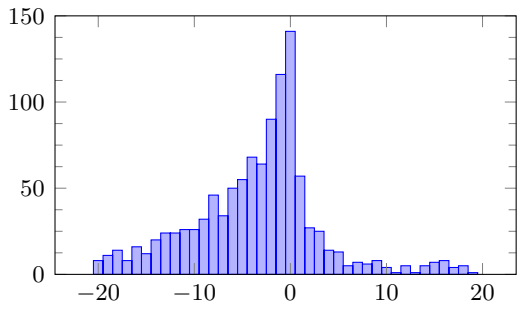

Figure 5: 《汉书》《史记》异文净变化字数直方图（绝对值大于20的部分省略；横轴为净变化字数，纵轴为频数）

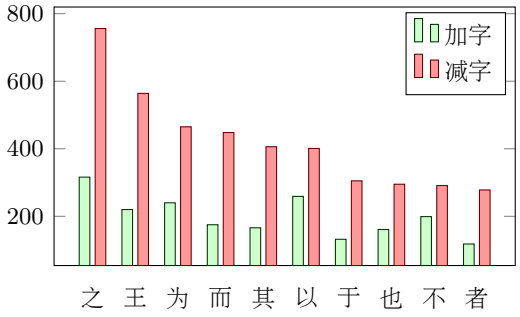

Figure 6: 《汉书》相比《史记》异文删除次数最高的十个字的加字与减字数量（纵轴为加字或减字的频数）

4.2.1 宏观统计分析

此处本文继续利用4.1节提出的文本相似度算法来筛选可以进行比对的文本。在4.1节中，算法的阈值 $\mu = 0.25$，这导致了一些实际并不相关的文本被算法筛选出来。因此，本节将算法的阈值 μ 提高到0.5，从而得到了1364对异文。本文分别计算出这些异文的LCS，将两段文本分别与LCS对齐之后，统计了《汉书》文段在《史记》的基础上增加和删除了哪些字。本文将增加的字数与删除的字数之差称为这对异文的“净变化字数”。Figure 5是算法得出的所有异文的净变化字数的直方图，其中均值为−8.37，中位值为−3。为了从统计上证明《汉书》相比《史记》存在明显的“删字”现象，本文对于异文的净变化字数进行了单尾单样本T检验。检验的原假设为“异文净变化字数的均值大于等于0”。经计算，T统计量的值为−8.950，$p < 0.0001$，故原假设不成立，表明在统计意义上，《汉书》对于《史记》有显著的“减字”现象。

Figure 6中画出了《汉书》异文相比于《史记》删除的字中频率最高的十个字的加字和减字的数量（异体字已去除）。虽然这些字多为无意义的虚词，但是总体上仍能看出删除字数比增加字数更多的趋势，这也与前人(朴宰雨, 1994; 沙志利, 2005)所认为《汉》比《史》异文的文字更加简短凝练的研究结论高度相符。

4.2.2 微观典型例子分析

本文从《史》《汉》异文中选取了两个典型例子，对LCS对齐结果进行对比分析。限于篇幅限制，具体的例子及分析列入附录B。从例子可以看出，LCS可以有效帮助我们在微观层面上对两书异文进行剖析，与宏观分析相配合，对于文本进行更立体的解读。

总的来说，由以上分析可见，本节所使用的LCS算法能够有效地对《史记》《汉书》中的异文进行比对。通过利用LCS对于《史》《汉》异文进行宏观和微观对比，我们可以更好地把握异文在字词增删、用字用词、历史事实、叙事顺序、人物塑造上存在的差异。

5 结论

本文借助基于BERT的古汉语分词模型和命名实体识别模型，对于《史记》《汉书》的字、词、命名实体进行了全量统计对比，对其基本统计数据进行了比较，分析了高频字、高频词、高频命名实体的异同，对于典型词语在全书的分布密度进行了对比，从中挖掘了有关西汉及西汉之前我国历史、政治、文化、思想等方面的沿革，不但为我们理解这段历史提供了新的手段和视角，也为我们用大数据方法处理其他历史文献提供了借鉴。

本文利用一种以命名实体作为外部特征的基于TF-IDF的文本相似度算法对于《史记》《汉书》中的异文进行了自动发现，成功发现了过往学者所未能发现的异文片段，大大拓宽了我们对于《汉书》承袭《史记》规模与程度的认识。进一步，基于这些结果，本文利用最长公共子序列算法对特定异文对进行了对齐，从宏观统计和微观案例两个视角分析了异文加、减、改字的特点，利用更加科学严谨的手段对《史》《汉》写作风格和语言特点的异同作了新的阐释。

本研究站在数字人文的视域下，利用先进的计算方法对于传世千年的中国古代经典进行了再审视、再发现，其方法对于今人研究古籍有一定的借鉴价值，对于数字人文学科的发展做出了独特的贡献。

参考文献

Agirre E, Cer D, Diab M, et al. 2013. *SEM 2013 shared task: Semantic Textual Similarity. *Second joint conference on lexical and computational semantics (* SEM), volume 1: proceedings of the Main conference and the shared task: semantic textual similarity*, 32–43.

李林晓. 2020. 《史记·滑稽列传》选录标准及司马迁不载东方朔之原因. 太原学院学报*(*社会科学版*)*, 21(04):51–59.

李越. 2014. 《左传》《史记》同事异文自动发现及分析. 南京师范大学, 硕士学位论文.

梁媛, 王东波, 黄水清. 2021. 古籍同事异文的自动发掘研究. 图书情报工作, 65(09): 97-104.

朴宰雨. 1994. 《史记》《汉书》比较研究. 中国文学出版社, 北京.

Ramos J. 2003. Using tf-idf to determine word relevance in document queries. In *Proceedings of the first instructional conference on machine learning*, 29-48.

沙志利. 2005. 《史》《汉》比较研究. 北京大学, 博士学位论文.

Tang X and Su Q. 2022. That Slepen Al the Nyght with Open Ye! Cross-era Sequence Segmentation with Switch-memory. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7830–7840, Dublin, Ireland. Association for Computational Linguistics.

王海平. 2003. 《史记》《汉书》异文研究. 暨南大学, 硕士学位论文.

吴福助. 1975. 史汉关系. 文史哲出版社, 台北.

夏德靠. 2019. 从“帝王传记”到“帝王大事记”——《史记》《汉书》“本纪”叙事异同简论. 四川师范大学学报*(*社会科学版*)*, 46(06):115-126.

肖磊, 陈小荷. 2010. 古籍版本异文的自动发现. 中文信息学报, 24(05):50-55.

袁祖亮. 1988. 战国秦汉魏晋南北朝时期的相国与丞相. 郑州大学学报*(*哲学社会科学版*)*, 1988(06):58-65.

曾小霞. 2009. 明清《史记》《汉书》比较研究综述. 苏州大学学报*(*哲学社会科学版*)*, 30(02):71-73.

曾小霞. 2012. 从《史记》和《汉书》看汉代文学之演变. 山西师大学报*(*社会科学版*)*, 39(03):58-61.

张明月. 2021. 《史记》《汉书》重合篇章比较研究. 辽宁师范大学, 硕士学位论文.

张添雅. 2021. 《汉书》八表研究. 哈尔滨师范大学, 硕士学位论文.

朱一玄, 刘毓忱. 2012. 《三国演义》资料汇编. 南开大学出版社, 天津.

诸雨辰. 2016. 历史文本的独断读法——章学诚的《史记》《汉书》解读. 求索, 2016(10):142-148.

附录A. 统计结果

排序	《史记》		《汉书》	
	字	频率（‰）	字	频率（‰）
1	之	24.71	之	20.93
2	王	16.36	为	14.95
3	为	14.80	以	14.83
4	以	14.65	不	13.23
5	不	14.53	王	10.08
6	子	12.96	年	8.94
7	而	11.79	其	8.54
8	年	11.39	而	7.90
9	曰	10.99	十	7.87
10	其	10.12	子	7.83
11	公	9.62	人	7.66
12	于	9.50	大	7.62
13	人	8.90	侯	7.61
14	侯	8.61	于	7.48
15	也	8.26	曰	7.30

Table 7: 《史记》《汉书》高频字

统计量		《史记》	《汉书》
平均词长		1.247	1.304
不同长度词语占比(%)	1	78.56	74.14
	2	18.80	22.32
	3	2.10	2.53
	4及以上	0.54	1.00

Table 8: 《史记》《汉书》词长分布对比

附录B. 异文LCS微观分析举例

本附录列出两个使用LCS对《史》《汉》异文进行对比的案例，以帮助我们理解LCS方法对于异文研究的辅助作用。引文中，《汉书》相比于《史记》删除的字标红并缩小，增加的字标绿加下划线，替换视作先删除后增加。

例1：《西域传》与《大宛列传》（节选）

大月氏在大宛西可二三千里，居妫水北。其南则大夏，西则安息，北则康居。本行国也，随畜移徙，与匈奴同俗。控弦者可一二十余万。故时，轻匈奴。本居敦煌、祁连间，及至冒顿立，单于攻破月氏，至匈奴而老上单于，杀月氏王，以其头为饮器。始月氏居敦煌、祁连间，及为匈奴所败，乃远去，过大宛，西击大夏而臣之，遂都妫水北，为王庭。其余小众不能去者，保南山羌，号小月氏。

本段主要讲大月氏和小月氏的渊源，差异主要体现在历史事实的明确程度和叙事顺序上。在谈论“控弦者”（士兵）时，《史记》言“可一二十万”，而《汉书》言“十余万”，明确地说明大月氏的士兵少于二十万，比《史记》精准。另外“过大宛”一句《汉书》相比于《史记》明确指出是大宛而非小宛，颇为严谨。在叙述大月氏最开始居敦煌、祁连时，《史记》采用倒叙手法，而《汉书》将这一句提前，按照时间顺序叙述。

例2：《杜周传》与《酷吏列传》（节选）

杜周者，南阳杜衍人也。义纵为南阳太守，以周为爪牙，举为廷尉史。事荐之张汤，汤数言其无害，至御为廷尉史。使案边失亡，所论杀甚众多。奏事中上意，任用，与减宣相编，更为中丞者十余岁。其治与宣相放，然周少言重迟，外宽，而内深次骨。

词长	排序	《史记》		《汉书》	
		词	频率（‰）	词	频率（‰）
2	1	天下	3.01	天下	2.53
	2	诸侯	2.32	匈奴	1.63
	3	于是	2.06	以为	1.54
	4	太子	1.52	丞相	1.44
	5	天子	1.50	天子	1.22
	6	匈奴	1.03	诸侯	1.17
	7	孔子	0.97	陛下	1.07
	8	以为	0.96	于是	1.05
	9	丞相	0.88	太子	1.01
	10	将军	0.88	单于	0.98
3	1	太史公	0.39	大将军	0.62
	2	大将军	0.37	大司马	0.51
	3	平原君	0.29	二千石	0.45
	4	孟尝君	0.25	关内侯	0.25
	5	淮南王	0.20	京兆尹	0.22
	6	二千石	0.20	皇太后	0.20
	7	齐桓公	0.18	左将军	0.20
	8	孝文帝	0.17	光禄勋	0.19
	9	秦昭王	0.15	董仲舒	0.19
	10	春申君	0.13	大司农	0.18
4	1	御史大夫	0.25	御史大夫	0.66
	2	骠骑将军	0.10	光禄大夫	0.31
	3	越王勾践	0.07	车骑将军	0.22
	4	孝文皇帝	0.06	太皇太后	0.14
	5	车骑将军	0.05	票骑将军	0.12
	6	吴王夫差	0.04	太中大夫	0.11
	7	齐悼惠王	0.04	水衡都尉	0.09
	8	太中大夫	0.04	孝文皇帝	0.07
	9	公子弃疾	0.03	司隶校尉	0.06
	10	二师将军	0.03	孝武皇帝	0.06

Table 9: 《史记》《汉书》高频词

本段中，《汉书》对于《史记》做了两处重要的改动。第一是杜周任廷尉史一事，《史记》记载杜周是先当了廷尉史再为张汤做事，而《汉书》则记载杜周是先被推荐给张汤才当上了廷尉史。这一差异说明《汉书》对其先后顺序进行了考订，肯定了张汤对酷吏杜周任职的关键作用。第二，《汉书》相比《史记》突出描写了杜周“少言”之特点，更显得杜周为人冷酷，不近人情，其人物形象顿时丰满。可以说，尽管《汉书》这一段对《史记》的改动十分微小，但是“微言大义”，把杜周的人物形象刻画地更加准确、立体、生动。

附录C. NER模型超参数与《四库全书》数据集信息

本文使用的NER模型采用BERT+Bi-LSTM+CRF架构，其主要超参数如Table 11。

《四库全书》是中国古代现存规模最大的丛书，包含了从先秦到清朝的古籍超过三千种。预训练使用的数据集是现代人对大多数现存《四库全书》进行电子化的成果。该数据集质量高、覆盖广，适合于构建预训练模型。语料全部为繁体字，没有标点符号，总字数在6亿左右。

实体类型	排序	《史记》		《汉书》	
		提及	频次	提及	频次
人物	1	孔子	400	莽	734
	2	汉王	338	光	370
	3	项羽	273	汉王	337
	4	高祖	265	王莽	286
	5	秦王	246	汤	281
	6	沛公	242	武	269
	7	赵王	227	信	242
	8	齐王	216	高祖	234
	9	汤	208	禹	232
	10	项王	203	羽	210
地点	1	秦	2601	汉	1545
	2	楚	1681	匈奴	968
	3	齐	1622	秦	771
	4	汉	1041	楚	668
	5	赵	1030	齐	555
	6	魏	805	长安	407
	7	晋	746	赵	359
	8	燕	580	周	309
	9	韩	567	河	291
	10	吴	511	吴	258
书籍	1	春秋	81	诗	243
	2	诗	77	春秋	220
	3	书	46	易	210
	4	易	31	书	133
	5	尚书	23	易传	73
	6	老子	19	尚书	55
	7	颂	16	礼	43
	8	礼	13	五经	35
	9	武	11	论语	33
	10	雅	9	左氏传	32
官职	1	太子	693	丞相	859
	2	丞相	497	将军	682
	3	将军	446	陛下	639
	4	大夫	284	太子	605
	5	陛下	269	上	539
	6	太后	257	太守	499
	7	御史大夫	195	太后	496
	8	大将军	189	御史大夫	395
	9	上	160	大夫	380
	10	夫人	139	大将军	372

Table 10: 《史记》《汉书》高频命名实体

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。

超参数	值
编码器骨架	BERT-base
BERT编码器层数	12
BERT隐向量维度	768
BERT自注意力头数	12
Bi-LSTM层数	2
Bi-LSTM隐向量维度	100

Table 11: NER模型超参数

第二十一届中国计算语言学大会论文集，第656页-第670页，南昌，中国，2022年10月14日至16日。