数字人文视域下简帛文献的分词研究 *

——以《里耶秦简牍》为例

刘 铭[1,2] 冯慧敏[1,2] 陈镱文[1,2]

（1. 西北大学科学史高等研究院 西安 710127；
2. 陕西省文化遗产数字人文重点实验室 西安 710127）

[摘要] 简帛文献是一类不同于传世典籍的传统文化载体。本文以两卷里耶秦简为例，结合数字人文的文本数据计算及分析方法，对其进行自动分词研究。基于经过人工标注的里耶秦简文本构建里耶秦简语料库，分别使用 3 类分词方法进行实验，对比并讨论其结果。实验显示，Bi-LSTM-CRF 模型的分词效果最佳，准确率达到 94.54%，召回率 94.82%，*F* 值为 94.68%。实验结果不仅验证了深度学习的分词方法在里耶秦简等简帛文献中的有效性和泛化能力，还表明其可应用于简帛词汇研究、语料库深加工以及文本分析等多元任务中。

[关键词] 数字人文；简帛文献；里耶秦简；自动分词；深度学习

[中图分类号] G255.1 [文献标识码] A [文章编号] 1003–5397（2024）03–0132–13

DOI:10.16499/j.cnki.1003-5397.2024.03.001

Research on Word Segmentation of Liye Qin Bamboo Slips from the Perspective of Digital Humanities

LIU Ming, FENG Huimin, CHEN Yiwen

Abstract: Bamboo slips and silk manuscripts are kinds of traditional Chinese culture and thought, which is different from ancient Chinese classics. Taking the two volumes of Liye Qin Slips published now as an example, this paper conducts research on automatic word segmentation. The Liye-Text-Corpus is constructed based on the artificially annotated Liye Qin Slips text, and the word segmentation experiments are carried out on three word-segmentation methods, and the comparison and discussion of its influence are carried out. The Bi-LSTM-CRF model works best, with an accuracy rate of 94.54%, a recall rate of 94.82%, and a *F* value of 94.68%. This result confirms the effectiveness and generalization ability of Deep Learning Word-Segmentation-Method for word segmentation on Bamboo slips and silk documents such as Liye Qin bamboo

[收稿日期] 2024-04-10

[作者简介] 刘铭，西北大学科学史高等研究院博士生，主要研究数字人文；冯慧敏，西北大学科学史高等研究院博士生，主要研究数字人文；陈镱文（通讯作者），西北大学科学史高等研究院教授，博士生导师，主要研究数字人文。

*本研究得到陕西省重点研发计划项目“数字化文化资源平台的智能分析与利用研究”（2019ZDLGY17-03）和陕西省秦创原队伍建设项目“数字人文视域下文化遗产人工智能核心技术研发与应用‘科学家 + 工程师’队伍”（2022KXJ-143）的资助。衷心感谢曲安京教授在论文写作中的指导，以及审稿专家提出的宝贵意见！

slips, and can serve downstream tasks such as vocabulary research, corpus deep processing, and text analysis on Bamboo Slips.

Keywords: digital humanities; bamboo slips and silk manuscripts; Liye Qin bamboo slips; automatic word segmentation; deep learning

一　引言

基于语言学和数字人文之间的内在联系，运用自然语言处理等计算机技术可实现文本的深度分析（刘炜等，2017；王军，2020；魏晓萍，2018；徐彤阳等，2021）。中文分词（Chinese word segmentation，CWS）作为自然语言处理、数据挖掘等信息处理的基础，是提升文本分析能力的重要指标（黄水清等，2017）。古籍文本中没有词界，需进行词语的切分，才能实现古籍文本深层次的语义理解问题（李斌等，2020；李明杰等，2020）。目前，分词研究在古汉语典籍信息处理中主要以传世文献文本语料为主，并取得了丰硕成果（周好等，2021）。

自然语言处理技术中，常见的中文分词方法有三种：词典分词、传统机器学习的统计分词和神经网络的深度学习分词（王佳楠等，2021）。随着古籍数字化的迅速发展，传世文献文本信息处理中的分词方法也基本围绕其展开（邓三鸿等，2021）。词典分词的特点是构建领域词典，专业领域词典对于领域分词重要且必要。邱冰、皇甫娟（2008）基于《汉语大词典》通过词典和统计相结合的方法对《国语》《论语》《商君书》等先秦至唐代的21 种古代汉语语料进行分词。与词典分词不同，传统的机器学习及深度学习模型无需构建领域词典，且能够提高模型的适用性，已被引入古籍文本信息处理中，成为常规方法。钱智勇等（2014）利用 HMM 模型对《楚辞》进行自动分词标注实验，最终得到一个分词标注辅助软件。杨新生、胡立生（2020）将改进的 HMM 模型与维特比算法（Viterbi）结合后，对《论语》《道德经》等古籍文本进行词性标注，使用的 bigram 和 trigram 模型的准确率分别为 94.9% 和 96.5%。Fu 等（2019）基于 HMM 模型进行词性标注构建中医术语词库和中医专用词性标注方法，其 F 值达到 90.65%。然而，传统的机器学习分词在古籍文本分词应用中仍显不足，对不同文本语料处理结果差异性大。深度学习方法可以根据训练语料自动学习特征，规避复杂的特征工程，实现模型泛化能力的提高。程宁等（2020）基于 Bi-LSTM-CRF 模型在《左传》《梦溪笔谈》《阅微草堂笔记》《清史稿》四部不同语料上进行断句、分词、词性标注测试，并在此基础上实现了一体化标注体系。王莉军等（2020）基于 Bi-LSTM-CRF 的模型对中医文言文文献进行分词研究，证明了此模型在分词任务上取得了比通用分词器更优的效果。刘畅等（2022）结合预训练和词典信息融合两种策略，运用四种深度学习模型在六部春秋至魏晋时期的官修史籍文本上进行自动分词实验。王东波等（2022）基于《四库全书》无监督训练集，构建了 SikuBERT 预训练语言模型及先秦典籍分词词性一体化标注平台。

然而，具有重要史料价值的出土文献，其分词研究鲜少见到。陈寅恪先生认为“一

时代之学术，必有新材料与新问题”[①]。百年来大量简帛文献的出土，如《清华大学藏战国竹简》《北京大学藏秦简牍》《里耶秦简牍》《居延汉简》《马王堆帛书》《长沙走马楼三国吴简》等，不仅弥补了传世文献资料的不足和缺失，还勘正了传世文献的误记和谬篡（李学勤，2013；李均明等，2019）。同时，简帛也是研究中国古代政治学、法制史、经济史、军事史以及语言学、文字学、哲学、医学、地理学乃至天文学等学科的重要史料，并极大地促进了中国多种学科的发展（林剑鸣，1984；骈宇骞等，2006；王彦辉，2011）。由于简帛语料属于封闭语料，具有残断性强、格式多、符号多等特征，需要预先进行数字化及大量的标注工作。其次，这类语料的词法、句法与传世文献及自然语言存在较大差异，故在保留其文本特征的基础上，实现对简帛文献的自动分词尤为重要。并且，在分词语料库的基础上，还能够利用文本挖掘等方法进行数字人文的研究（朱琳等，2022）。

综上所述，相较于自然语言领域，古籍文本分词研究虽已取得了较好效果，但对实验数据的选择尚不具备普适性。本文拟以《里耶秦简牍》（湖南省文物考古研究所，2012、2017）（以下简称为“里耶秦简”）为基础，首先基于自然语言和传世文献文本处理规范，根据专家意见及简文内容，编纂了《里耶秦简词典》，并在此基础上构建了分词语料库的加工规范；然后，在验证传世文献文本分词方法对简帛文献有效性的基础上，采用人机协助技术，深加工里耶秦简语料库，并探索不同分词模型类别对简帛文献分词的影响。本文将现有三种分词方法中 FMM、BMM、BM、HMM、LSTM、Bi-LSTM-CRF 及 BERT 七种应用较广泛的算法模型，在出版的两卷里耶秦简文本上进行自动分词实验。通过数据扩张的方式，将自然语言处理技术在传世文献上的应用方法移植至简帛这类出土文献文本中，并探究不同方法的分词模型对简帛文献分词的影响。研究显示，该语料库的深加工不仅可以完善古汉语语料库的加工规范，更可以实现出土文献语料的信息处理或数据挖掘等任务。

二　里耶秦简语料库的加工规范

秦汉简牍具有如下特征：第一，词汇及词性研究的基础资源匮乏；第二，区别于现代汉语语料不断扩张的发展态势，秦汉简牍属于封闭性语料；第三，秦汉简牍内容繁杂，每部简牍之间存在较大差异性，因此文献之间的同质性较低；第四，不同于传世文献文本，秦汉简牍具有其独特的“残篇断简居多”的文本特征（陈伟，2012、2018）。这就决定了现代汉语、传世文献文本分词及词性标注处理的一般模式不完全适用于秦汉简牍文本分词及词性标注的处理，需在现有研究的基础上，探索创新一种适应秦汉简牍文本分词的处理规范及方式。

里耶秦简包含 3.8 万余枚简牍，出土时大多残断，且无字简数量占总数一半以上。《里耶秦简牍》前两卷包含四层简文，第五层有 35 条简文，第六层 40 条，第八层 2552 条，第九层 3423 条，共计 6050 条简文，总字数约为 12.25 万，且已对其中出现的集外字进行信息处理（唐杰等，2022）。构建中文古籍标注语料库的基础工作之一是语料的分词和词性标注，但古汉语信息处理领域尚未像现代汉语领域一样形成较为科学的、通用的

分词与词性标注规范体系，这不利于古汉语信息处理的发展，对于出土文献这类特殊文献更是如此（郑童哲恒等，2023）。

（一）词典构建的规范

词的切分标准是建立简牍文本语料库的关键问题之一。词粒度划分直接影响机器分词结果的准确率和一致性（陈小荷等，2013；化振红，2014）。本实验采用两卷《里耶秦简牍校释》（陈伟，2012、2018），制定了简牍文本的分词规范，并且在结合其词法特征及特殊用语的基础上，编纂了《里耶秦简词典》，从而保证了模型的适用性。

该词典遵循的原则是：第一，组合词尽量拆分，使得词粒度为最小。如简文中常出现的“迁陵守丞”“迁陵丞”“洞庭尉”等形式为“地名＋职官名”的组合，切分为“迁陵\守丞”，即将地名与职官名切分开；第二，特殊用法的组合词以词的形式出现，如简文中出现“以邮行”等特殊用法的组合词时，不做切分；第三，“省称”词为单独的词，如简文中出现的“贰春乡”也被称为“贰乡”；第四，习语及固定搭配短语为词，如简文中“以律令从事”等，其他词的切分都以最小词粒度为标准进行切分。此想法在构建里耶秦简词典和人工校对里耶秦简语料库的步骤中都有所体现。

《里耶秦简词典》共包含 2985 个词条，其中单字词 1134 个，占比 42%；二字词 1148 个，占比 43%；三字词 278 个，占比 10%；四字词 92 个，占比 1%，其余词的个数占比均低于 1%。词典示例见图 1。

图 1　《里耶秦简词典》示例

（二）构建分词语料库的规范

下面从文本语料含量、文本格式、符号等方面详述本实验的分词语料库构建标准。

首先，简牍文本的内容形式相对复杂。简牍内容具有残断性（如图 2 所示），极少的简片保存完好（如图 3 所示），且存在“无字简”（如图 4 所示）。无字简总计有 253 条，在文本中的分布为第五层 7 条、第六层 3 条、第八层 75 条、第九层 168 条。由于无字简对分词结果无影响，故在语料预处理时，清洗了所有无字简的信息。

图 2　简 8–78（正、背）

图 3　简 9–1（正、背）

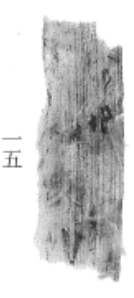

图 4　简 5–15

其次，简牍的简文格式有其独特性。简文内容分栏显示，如“第一栏”“第二栏”等；简有正背面，其中部分简文正背面文义相连，但也有部分简文的相邻两栏之间、正背面文义不相连。例如，简 8-458 是迁陵县库记录兵器的簿籍类文书。其简文与两种不同的分词结果如表 1 所示。

表 1　简 8–458 去除分栏格式前后的分词结果对比

原简文	标准切分结果	去除分栏格式后切分结果
遷陵庫真□ 甲三百冊九 甲宪廿一 鞮瞀卅九（第一欄） 胄廿□ 弩二百五十一 臂九十七 弦千八百一（第二欄） 矢四萬九百□ 戟二百五十（第三欄）	遷陵 \ 庫 \ 真 \ □ 甲 \ 三百冊九 甲宪 \ 廿一 鞮瞀 \ 卅九 胄 \ 廿 \ □ 弩 \ 二百五十一 臂 \ 九十七 弦 \ 千八百一 矢 \ 四萬九百 \ □ 戟 \ 二百五十	遷陵 \ 庫 \ 真 \ □ \ 甲 \ 三百冊九 \ 甲宪 \ 廿一 \ 鞮瞀 \ 卅九 \ 胄 \ 廿 \ □ \ 弩 \ 二百五十一 \ 臂 \ 九十七 \ 弦 \ 千八百一 \ 矢 \ 四萬九百 \ □ \ 戟 \ 二百五十

通过初步分析发现：第一，由于简文中无断句，全部去除格式后，会使得理解文义有误或是计算机过度理解，导致分词准确率下降。在此结果中，去除分栏格式切分结果比标准切分结果多出 9 组词组合，如“三百冊九 \ 甲宪”“千八百一 \ 矢”等。根据机器学习自动分词的原理，增加或减少词组合会影响分词准确率。第二，简牍文书中的簿籍多以分栏书写，栏数以三、四栏居多，也有多达八栏者。若去除分栏格式，在后续的文本分类中，簿籍类将缺少文本特征。第三，简牍文本中分栏的方式灵活，秦吏在记录时通常根据不同情况采用不同方法（李学勤，2013）。去除分栏特征，则会对簿籍类文书的具体分类产生影响。

此外，简牍文本中存在一定数量的符号。这些符号分为两类（如表 2 所示）。第一类是简文原有符号，部分符号进行了替换；第二类是后人校释简牍文字时使用的释文符号。所有符号在数据预处理中均被保留。

表 2　里耶秦简语料库中内含的符号举例

序号	符号	符号类型	是否有替换符号	含义
1	·	简文符号	无	
2	ㄴ	简文符号	无	
3	ノ	简文符号	有（/）	
4	少	简文符号	无	
5	丨	简文符号	有（|）	
6	=	简文符号	无	
7	□	释文符号	无	表示简文残泐无法辨认的字，一字一“□”。
8	……	释文符号	无	简文漫漶，表示不确定字数的。
9	⧄	释文符号	无	表示简文残断处，或一角残断。
10	☒	释文符号	无	简文被削除处。

在简牍文本中，最常用的符号是“□”和“⊠”（如表 3、表 4 所示）。表 3 原简文中的“士五”后通常情况下会出现人名，原简中未识别出其后二字具体为何，故采用符号“□”代替。去除符号后，“士五”与“若”形成词组合，给文本增加特征，不利于准确分词。简文中还出现了符号“/”，删除这类符号后，会丢失部分文本特征，也不利于分词。

表 3　简 9–1874 背面简文删除符号前后分词结果对比

原简文	标准切分结果	删除符号后切分结果
敢告主毋公印以私印印章曰 李志十一月甲午銷士五□□ 若思以來 / □□但手	敢告主 \ 毋 \ 公印 \ 以 \ 私印 \ 印章 \ 曰 \ 李志 十一月 \ 甲午 \ 銷 \ 士五 \ □ \ □ \ 若 \ 思 \ 以來 \ / \ □ \ □ \ 但 \ 手	敢告主 \ 毋 \ 公印 \ 以 \ 私印 \ 印章 \ 曰 \ 李志 十一月 \ 甲午 \ 銷 \ 士五 \ 若 \ 思 \ 以來 \ 但 \ 手

表 4　简 9–1117 与简 9–1194 拼连缀合中符号的作用

简 9–1117	简 9–1194	拼连缀合后的简文
徑廩粟米　四斗泰半斗　　卅 一年六月壬午朔朔日田官守敬佐郤 稟人婬出資居貲士五巫庫□⊠ 令史逐視平⊠	⊠處闌叔五月乙亥以盡辛 巳七日食 ⊠郤手	徑廩粟米　四斗泰半斗　　卅一年六月壬 午朔朔日田官守敬佐郤 稟人婬出資居貲士五巫 庫處闌叔五月乙亥以盡辛巳七日食 令史逐視平 郤手

再如简 9-1117 与简 9-1194，校释第二卷将该两简拼合后缀连（如表 4 所示）。

符号“⊠”意为简 9-1117 的下半段残断，要将该简进行拼合缀连，需首先查找首字符为“⊠”的简文。因此，符号不仅在分词任务中对文本正确分词有意义，在后续的文本挖掘任务中也具有重要价值。

综上，本文尝试基于简文内容、格式、符号等特征对语料进行预处理，建成里耶秦简语料库，并在此基础上进行分词实验。

三　简帛文献分词实验

（一）分词模型简介

词典分词也称为机械分词，其核心是创建分词词典，保存尽可能多的词汇集合，通过读取输入的文本内容进行匹配切分，常用的有正向最大匹配算法（Forward Maximum Matching，FMM）、逆向最大匹配算法（Backward Maximum Matching，BMM）和双向最大匹配算法（Bi-directction Matching，BM）。

统计分词的实验选择了隐马尔科夫模型（Hidden Markov Model，HMM）。该模型是一种基于统计的分词模型，也是关于时序的概率模型。在 HMM 中有一个经典假设，即每个状态只与它前面 n 个状态有关（n=1 时为一阶马尔科夫模型，且 n 是影响状态转移的数目）。本实验 HMM 模型的预测算法为 Viterbi 算法。

深度学习（Deep Learning）是机器学习重要分支领域，深度学习使用神经网络模拟人类智能。长短期记忆网络（Long Short-Term Memory，LSTM）通过引入控制门和记忆单元解决其他循环神经网络在训练过程中梯度消失和梯度爆炸的问题。单向的 LSTM 无法保留后文信息，因此本实验同时选择了双向循环神经网络模型（Bidirectional Long Short-Term Memory，Bi-LSTM）。Bi-LSTM-CRF 模型将上下文信息进行编码，最大限度地保留了整个信息序列的训练结果，从而提升预测效果（Schuster et al.，1997）。CRF 模型是一

个无向图模型的框架，Bi-LSTM-CRF 模型将 Bi-LSTM 的输出序列作为 CRF 的观测序列，通过 CRF 将局部归一转换为全局归一，从而获得全局最佳的标签序列。

BERT 模型是一种将双向 Transformer 作为基础网络结构构建的神经网络模型。它是完全基于自注意力机制对文本进行建模，且具有强大的特征提取能力。通过海量语料的预训练，BERT 模型能够获得序列当前最全面的局部和全局特征表示的结果。在针对具体任务的应用中，BERT 模型首先通过在大规模数据集上进行预训练来生成语言模型，随后对模型的顶层参数微调（Fine-Tuning），从而能够重新优化高层特征提取器的参数，最终使得模型适应并高效地完成新数据集上的特定任务（耿云冬等，2022）。

（二）语料预处理与实验设置

词典分词以《里耶秦简词典》为切分标准，对里耶秦简文本进行词切分，同时比较三种不同方法的分词效果。

基于统计的分词方法及深度学习的分词方法中，本文均采用四词位标注 4-tag 方法，输入层文本序列为“七月 \ 癸亥 \ 旦 \ 士五 \ 臂 \ 以來”，其对应的标签是“BE\BE\S\BE\S\BE”。

在深度学习中，本文采用 Word2vec 模型训练的词向量（Word Embedding）作为 LSTM 及 Bi-LSTM-CRF 模型的输入层（Torregrossa et al.，2021；Song et al.，2019）。如果将“词”看作文本的最小单元，可以将词向量理解为一种映射，其过程是将文本空间中的某个词，映射或嵌入（Embedding）到另一个数值向量空间中，实现文本数据到数值型数据的转换。这种方法在利用数学向量解释潜在语义信息的同时，也利用向量数值上的相似性解释了文本语义的关联性。目前深度学习主要利用 Word2vec 模型进行训练。以 Bi-LSTM-CRF 模型架构为例，深度学习的工作框架如图 5 所示。

训练集与测试集将里耶秦简语料按照 4:1 的比例随机划分，并按三类分词方法的七种分词模型依次实验，对比并分析其分词效果。本研究的技术路线如图 6 所示。

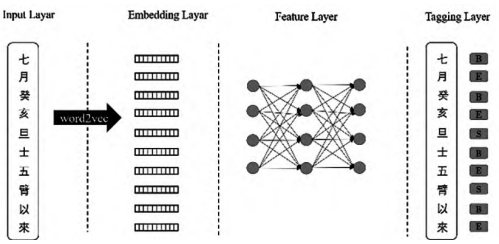

图 5　深度学习的工作框架

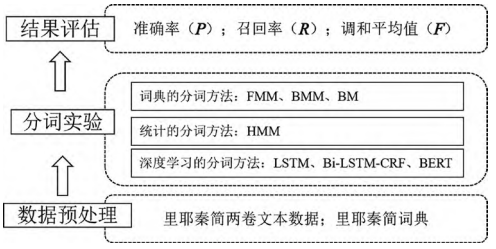

图 6　技术路线

LSTM 及 Bi-LSTM-CRF 模型中神经网络训练具体的实验参数如表 5 所示。在 LSTM 及 Bi-LSTM-CRF 结构中，通过在验证集上进行实验，发现层数对实验效果的影响较弱，故将层数设置为 2。由于 80% 左右的简文长度在 200 字以内，故将词向量维度设置在 100，适当缩小维度，提高词与词之间的相关性。在序列标注任务中，隐藏节点数通常取 200~600 之间，这里取 200 作为参数。学习率设置为 0.005 时，模型收敛最优。在模型的优化中，采用序列标注问题中效果最好的 ADAM 算法，最小样本数设为 32。权重衰

减设为 0.0001，弱化各个特征之间由于数据量太小导致过多的相互作用问题，在一定程度上减少模型过拟合问题。

表 5　LSTM 及 Bi-LSTM-CRF 的实验参数

超参数	LSTM	Bi-LSTM-CRF
层数	2	2
embedding_dim	100	100
hidden size	200	200
learning rate	0.005	0.005
batch size	32	32
weight_decay	0.0001	0.0001

在 BERT 模型中，本实验选取的实验工具为 Python 3.8，环境为 Transformers 4.5.0 与 Pytorch 1.6.0，并选取了 Hugging Face 提供的 12 层、768 个隐藏单元，1.1 亿参数量，及用于 Pytorch 框架的 BERT-base-Chinese 预训练模型。所有 BERT 模型的超参数均调整至最佳状态。模型在任务环境下的主要超参数设置如表 6 所示。

表 6　BERT 的实验参数

超参数	值
max_seq_length	512
learning rate	3e-5
batch_size	16
dropout	0.3

（三）实验结果及分析

实验结果以验证集数据结果为衡量标准，评价指标为准确率、召回率和 *F* 值。准确率（*Precision*，*P*）表示分词结果中的正确率；召回率（*Recall*，*R*）表示分词结果中正确分词数与标准分词结果的比率；*F* 值是这两个指标的综合评价。三者的计算公式分别为：

$$P=\frac{RW}{AW} \quad (1)$$

$$R=\frac{RW}{SW} \quad (2)$$

$$F=\frac{2PR}{P+R} \quad (3)$$

其中，*RW* 表示分词结果中正确分词的词目数，*AW* 表示分词结果的总词数，*SW* 表示人工分词结果中的总词数。*P*、*R* 的取值范围在 0 和 1 之间，*F* 的取值范围在 *P* 和 *R* 之间，这三者数值越接近 1，代表分词效果越好。

经过语料预处理及七种分词模型构建，基于里耶秦简语料库训练得到的深度学习模型测试结果最佳，实验的评价指标如表 7 所示。

实验结果表明，与统计的分词方法相比，深度学习的分词方法准确率、召回率及 *F* 值较高，这说明深度学习的分词方法在里耶秦简这类简帛文献分词中具有较强的适应性。

（1）词典的分词方法简单快捷，准确率最高可达 83.50%，验证了构建《里耶秦简词典》在封闭语料库上的有效性及必要性。实验中采用了三种词典分词方法，效果最好的

表 7　基于里耶秦简语料库得到七个模型的分词效果

分词方法	模型	准确率（*P*）/ %	召回率（*R*）/ %	调和平均值（*F*）/ %
词典的分词方法	FMM	83.50	88.85	86.09
	BMM	82.93	80.96	81.94
	BM	83.01	81.03	82.01
统计的分词方法	HMM	91.91	89.96	90.91
深度学习的分词方法	LSTM	94.52	89.01	91.68
	Bi–LSTM–CRF	94.54	94.82	94.68
	BERT	88.38	88.03	88.20

是 FMM，其 *P* 值、*R* 值及 *F* 值在三种算法里均为最高，这说明了 FMM 可以利用词典切分得到更多准确的词。这一结果与传统的词典分词中 BM 的效果更好的情况不同。其主要原因是BM将前向最大匹配法得到的分词结果和后向最大匹配法得到的结果进行比较，按照最大匹配原则，选取词数切分最少的作为结果。由于书写材料的限制，简牍语言以简短为主要特征，其单字词占比较大。“词数切分最小”算法机制在这里适用性不强，故其效果较低。然而，《里耶秦简词典》中包含 2985 个词条，该方法对词典的依赖性较大。

（2）统计的分词方法实验效果较优于词典的分词方法。在利用 HMM 进行的自动分词实验中，取最大概率作为最后的分词结果，并在其中使用全切分和加值平滑算法。其 *P* 值、*R* 值及 *F* 值均有提升，尤其是 *P* 值从 83.50% 提升至 91.91%，较为显著。这与现代汉语、传世文献语言应用效果的结论相同，统计的分词方法在里耶秦简这类简牍语言中的效果优于词典的分词方法。HMM 模型的局限性在于未登录词的预测问题，以往研究表明对未登录词标注的准确率明显低于已知单词（杨新生等，2020），但相较于词典的分词方法，HMM 模型的效果较好。

（3）与 HMM 模型相比，LSTM 模型的 *P* 值、*R* 值及 *F* 值均有提升，*P* 值提升至 94.52%。将两个传统深度学习 LSTM 模型上下叠加在一起，并加入 CRF 层后，相较于单个 LSTM 模型，其 *R* 值和 *F* 值均有明显提升。由于单个 LSTM 无法对从后到前的信息进行编码，而 Bi-LSTM 模型可以更好地捕捉双向的语义依赖，叠加 CRF 层后可找到更优的表达。实验结果证明了将 Bi-LSTM 与 CRF 结合，在针对简牍语言的分词任务上具有一定的突出性。

（4）实验结果表明，BERT 模型的效果不如 Bi-LSTM-CRF 模型。里耶秦简语料库的规模仅包含了 12 万个字符数，远少于传世文献千万级以上文本量，而 BERT 预训练语言模型则是针对大数据构建而成的。BERT-Base-Chinese 是 BERT 模型开发的古籍文本智能处理模型，它能对文本进行预训练，也能够学习到其语义和语法信息。有研究表明，通过对此模型进行微调，可以使其更加专注于特定领域或特定任务，从而提高结果的准确性和效率（俞敬松等，2019）。然而，即使本次训练语料采用了里耶秦简语料库，同时进行了微调，但在此类特定的简帛文献分词任务上，Bert-Base-Chinese 表现不如其他模型。在后续工作中，可尝试增加数据集和调整此模型来提高分词效果。

综上，里耶秦简语料库的应用效果显示，深度学习的分词方法效果最佳。相较于传统的机器学习，深度学习属于端到端（end-to-end）的学习，在优化损失函数和挖掘数据的

潜在特征基础上，可以自动从数据中学习其特征表示。里耶秦简语料库属于小型语料库，且具有文本残断性强的特征。在本次分词实验中，实验效果说明了深度学习的分词方法在里耶秦简这类简牍文本上具有一定的适用性及有效性。

（四）其他简牍文本上的分词应用

简牍作为出土文献，其语言特征与成书年代紧密相关，且内容繁杂。根据实验结果，选择了 Bi-LSTM-CRF 模型及泛化能力较好的 BERT 模型，测试在睡虎地秦简、岳麓秦简和张家山汉简文本上的分词能力，并与里耶秦简进行比较。这三部简的成书年代均是秦至汉初时期，与里耶秦简的成书年代相近；其次，在文书内容上，三部简的大部分内容是律令、算术书及日书等，与里耶秦简中的部分公文文书内容较为相近。测试中未选择清华大学藏战国竹简（以下简称“清华简”）。清华简是战国时期楚国境内的竹书，其内容以书籍为主，且清华简中存在较多的集外字（李学勤，2010）。这与里耶秦简文本的体例、内容及文字上均存在较大差异。故基于里耶秦简语料库的分词模型不能直接应用于清华简这类简帛文献上。

里耶秦简、睡虎地秦简、岳麓秦简与张家山汉简的部分简文及其分词结果如表 8 所示。在两条里耶秦简的简文测试结果中，BERT 模型没有 Bi-LSTM-CRF 模型效果好，但在另外三部简牍文本上，BERT 模型展示出了优于 Bi-LSTM-CRF 模型的分词能力。这是预训练语言模型泛化能力的优势所在。然而，观察这三部简牍文本的分词结果，尚有部分切分不准确。这是由于里耶秦简的文本数据量及内容均有局限性，在其他简牍文本中出现了较多的未登录词，如“人奴妾”“二千石官”“县道官”“变事”等，此类未登录词多是专有名词或是特殊用语。因此，在后续工作中，应当增加更多的简牍文本作为训练样本，如清华简、居延汉简、敦煌汉简及长沙走马楼三国吴简等文本量较大的简牍文本。具体思路如下：

（1）增加同一时代不同内容的简牍文本为训练样本，从现有的秦代扩展至战国时期、汉代及魏晋时期等。

（2）同一时代但出土地域不同也在一定程度上影响着分词结果。如居延汉简是出土于居延烽燧遗址中的简牍，其内容主要涉及汉代西北边塞的屯戍活动；而张家山汉简则是在湖北江陵张家山出土的，其内容主要涉及汉初律令、历谱及遣策等。增加不同出土地域的简牍，也有助于丰富训练样本涉及的内容。

（3）现已出土的简帛文献中还有部分简文未被释读或等待刊出，如另三卷里耶秦简、悬泉汉简等，这部分简文也需要不断补充进来。

（4）在增加语料库内容及数量的基础上，词表和语料库规范也需要同时扩充。如汉简中常见的“叩头死 = 罪 = 敢言之”等特殊用语；《长沙走马楼三国吴简 – 嘉禾吏民田家莂》中常见的合同符号“  ”在秦简中也未见到；《长沙五一广场东汉简牍》中常见到的“君教若”简文格式与目前所见的秦简也有部分差异。因此，词表的扩充还需要收纳更多的特殊用语及专有名词；语料库构建规范还需要根据收入简牍文本的特征，扩充符号、语序和格式。

表 8 Bi-LSTM-CRF 与 BERT 实验结果及校验[②]

简牍名称	Bi-LSTM-CRF 测试结果	BERT 测试结果	正确分词结果
里耶秦简《里耶秦简牍》	元年\七月\庚子\朔\丁未\倉守\陽\敢言之\獄佐\辨\ㄴ\平\ㄴ\士吏\賀\具獄\縣官\食\盡\甲寅\謁\告\過所\縣鄉\以\次\續食\雨留\不能\投\宿\齎\來復傳\零陽\田\能\自食\當\騰\期\卅日\敢言之\/\七月\戊申\零陽\襲\移\過\所\縣鄉\/\齮\手\/\七月\庚子\朔\癸亥\遷陵\守丞\固\告\倉嗇夫\以律令從事\/\嘉\手\	元年\七月\庚子\朔\丁未\倉守\陽\敢言之\獄佐\辨\ㄴ\平\ㄴ\士吏\賀\具獄\縣官\食\盡\甲寅\謁\告\過所\縣鄉\以\次\續食\雨留\不能\投\宿\齎\來\復\傳\零陽\田\能\自食\當\騰\期\卅日\敢言之\七月\戊申\零陽\襲\移\過\所\縣鄉\/\齮\手\/\七月\庚子\朔\癸亥\遷陵\守丞\固\告\倉\嗇夫\以律令從事\/\嘉\手	元年\七月\庚子\朔\丁未\倉守\陽\敢言之\獄佐\辨\ㄴ\平\ㄴ\士吏\賀\具獄\縣官\食\盡\甲寅\謁告\過所\縣鄉\以次\續食\雨留\不能\投宿\齎\來復傳\零陽\田\能\自食\當騰\期\卅日\敢言之\/\七月\戊申\零陽\襲\移\過\所\縣鄉\/\齮\手\/\七月\庚子\朔\癸亥\遷陵\守丞\固\告\倉嗇夫\以律令從事\/\嘉\手
	士五\巫南\就\曰\路\娶\貲\錢\二千六百\☑	士五\巫南\就\曰\路\娶\貲\錢\二千六百\☑	士五\巫\南就\曰\路\娶\貲錢\二千六百\☑
睡虎地秦简《司空律》	隸臣\妾\、\城旦\舂\之\司寇\、\居貲\贖責\罷\城旦\舂\者\，\勿責\衣食\；\其\與\城旦\舂\作者\，\衣食\之\如\城旦\舂\。\隸臣\有\妻\，\妻\更\及\有外妻\者\，\責\衣\。\人\奴妾\罷\城旦\舂\，\貳\衣食\公\，\日未備\而\死者\，\出\其\衣食\。\司\空	隸臣\妾\、\城旦\舂\之\司寇\、\居貲\贖責\罷\城旦\舂\者\，\勿責\衣食\；\其\與\城旦\舂\作者\，\衣食\之\如\城旦\舂\。\隸臣\有\妻\，\妻\更\及\有外妻\者\，\責\衣\。\人\奴妾\罷\城旦\舂\，\貳\衣食\公\，\日未備\而\死者\，\出\其\衣食\。\司空	隸臣妾\、\城旦舂\之\司寇\、\居貲贖責\罷\城旦舂\者\，\勿責\衣食\；\其\與\城旦舂\作者\，\衣食\之\如\城旦舂\。\隸臣\有\妻\，\妻\更\及\有外妻\者\，\責衣\。\人奴妾\罷\城旦舂\，\貳\衣食\公\，\日未備\而\死者\，\出\其\衣食\。\司空
睡虎地秦简《置吏律》	除\吏\、\尉\，\已\除\之\，\乃\令\視事\及\遣\之\；\所\不當除\而\敢\先見事\，\及\相聽\以\遣\之\，\以律\論之\。\嗇夫\之\送\見\它官\者\，\不得\除\其\故\官佐\、\吏\以\之\新官\。\置吏\律	除\吏\、\尉\，\已\除\之\，\乃\令\視事\及\遣\之\；\所\不當除\而\敢\先見事\，\及\相聽\以\遣\之\，\以律\論之\。\嗇夫\之\送\見\它官\者\，\不得\除\其\故\官佐\、\吏\以\之\新官\。\置吏\律	除吏\、\尉\，\已\除之\，\乃令\視事\及\遣之\；\所\不當\除\而\敢\先見事\，\及\相聽\以\遣之\，\以律\論之\。\嗇夫\之\送見\它官\者\，\不得\除\其\故官佐\、\吏\以\之\新官\。\置吏律
岳麓秦简《田律》	•\田律\曰\:\吏\歸休\，\有\縣\官吏\乘\乘馬\及\縣官\乘馬\過\縣\，\欲貣\芻稾\、\禾\、\粟\、\米\及\買菽者\，\縣\以\朔日	•\田律\曰\:\吏\歸休\，\有\縣\官吏\乘\乘馬\及\縣官\乘馬\過\縣\，\欲貣\芻稾\、\禾\、\粟\、\米\及\買菽者\，\縣\以\朔日	•\田律\曰\:\吏\歸休\，\有\縣官\吏乘\乘馬\及\縣官\乘馬\過\縣\，\欲貣\芻稾\、\禾\、\粟\、\米\及\買菽\者\，\縣\以\朔日
岳麓秦简《数》	租\誤券\。\田\多若\少\，\耤\令\田十\畝\，\稅田\二百\冊\步\，\三\步一\斗\，\租八\石\。\• 今\誤券\多\五斗\，\欲益\田\。\其\述\曰\:\以\八\石五\斗\為八\百\。	租\誤券\。\田\多\若\少\，\耤\令\田\十\畝\，\稅田\二百冊\步\，\三\步\一\斗\，\租\八\石\。\•\今\誤\券\多\五\斗\，\欲\益田\。\其\述曰\:\以\八\石\五\斗\為\八百\。	租\誤券\。\田\多\若\少\，\耤\令\田\十\畝\，\稅田\二百冊\步\，\三\步\一\斗\，\租\八\石\。\•\今\誤券\多\五\斗\，\欲\益田\。\其\述曰\:\以\八\石\五\斗\為\八百\。
张家山汉简《盜律》	盜\臧直\過\六百\六十\錢\，\黥為\城旦\舂\；\六百\六十\到二\百廿\錢\，\完為\城旦\舂\；\不盈\二百\廿到\百一\十\錢\，\耐為\隸臣\妾\；\不盈\百一\十\錢\到\廿二\錢\，\罰金\四\兩\；\不盈\廿二\錢到\一錢\，\罰\金\一兩\。	盜\臧\直\過\六百六十\錢\，\黥\為\城旦\舂\；\六百六十\到\二百廿\錢\，\完\為\城旦\舂\；\不盈\二百廿\到\百一十\錢\，\耐\為\隸臣妾\；\不盈\百一十\錢\到\廿二\錢\，\罰金\四\兩\；\不盈\廿二\錢\到\一\錢\，\罰金\一\兩\。	盜臧\直\過\六百六十\錢\，\黥\為\城旦舂\；\六百六十\到\二百廿\錢\，\完\為\城旦舂\；\不盈\二百廿\到\百一十\錢\，\耐\為\隸臣妾\；\不盈\百一十\錢\到\廿二\錢\，\罰金\四\兩\；\不盈\廿二\錢\到\一\錢\，\罰金\一\兩\。

张家山汉简《置吏律》	郡\守二\千\石官\、\縣\道\官言\邊\變\事急\者\，\及吏\遷徙\、\新\為官\屬尉\、\佐\以上\毋\乘馬\者\，\皆\得為\駕傳\。\縣\道官\之\計\，\各關\屬所\二千\石官\。\其\受恒\秩氣\稟及\求財\用年\輸\，\郡關\其守\，\中\關內\史\。\受爵\及除\人關\於尉\。\都\官自\尉\、\內史\以下\毋\治獄\，\獄\無輕\重關\於正\；\郡\關其\守\。	郡守\二千\石\官\、\縣道\官\言\邊變事急者\，\及\吏\遷徙\、\新\為\官\屬尉\、\佐\以上\毋\乘馬\者\，\皆\得\為\駕傳\。\縣道\官之\計\，\各\關\屬所\二千\石\官\。\其\受\恒秩氣\稟\及\求財\用年\輸\，\郡關\其守\，\中關\內史\。\受\爵\及\除\人\關\於\尉\。\都官\自\尉\、\內史\以下\毋\治獄\，\獄\無\輕重\關\於\正\；\郡\關\其守\。	郡守\二千\石\官\、\縣道官\言\邊\變事急者\，\及\吏\遷徙\、\新\為\官\屬\尉\、\佐\以上\毋\乘馬\者\，\皆\得\為\駕傳\。\縣道官\之\計\，\各關\屬所\二千\石\官\。\其\受\恒秩\氣稟\及\求財用\年輸\，\郡關\其守\，\中關\內史\。\受爵\及\除人\關於\尉\。\都官\自尉\、\內史\以下\毋\治獄\，\獄\無\輕重\關於\正\；\郡關\其守\。

四　结语

本文在一个新的领域——简帛类出土文献的分词研究进行了探索。首先，由于简帛具有残篇断简居多、缺字、漏字的文本特征，在数据预处理时对其进行不同于传世文献及自然语言的加工；其次，实验验证了在传世文献文本上应用的深度学习分词方法，对里耶秦简这类简帛文献的分词研究具有一定优越性，证实了深度学习的分词方法在简帛文献上的有效性及较好的泛化能力。该实验结果可服务于里耶秦简语料库的深加工、出土文献词法研究和文本分析等，为简帛文献的数字人文研究提供了良好的信息处理基础。

［附　注］

① 参见陈寅恪《敦煌劫余录序》。

② 表中 Bi-LSTM-CRF 模型及 BERT 模型分词结果与正确结果有区别之处均用下划线标注。

［参考文献］

［1］陈　伟 . 里耶秦简校释（第一卷）［M］. 武汉：武汉大学出版社，2012.

［2］陈　伟 . 里耶秦简校释（第二卷）［M］. 武汉：武汉大学出版社，2018.

［3］陈小荷，冯敏萱，徐润华等 . 先秦文献信息处理［M］. 北京：世界图书出版公司北京公司，2013.

［4］程　宁，李　斌，葛四嘉等 . 基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究［J］. 中文信息学报，2020，（4）.

［5］邓三鸿，胡昊天，王　昊等 . 古文自动处理研究现状与新时代发展趋势展望［J］. 科技情报研究，2021，（1）.

［6］耿云冬，张逸勤，刘　欢等 . 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［J］. 图书馆论坛，2022，（6）.

［7］湖南省文物考古研究所 . 里耶秦简（壹）［M］. 北京：文物出版社，2012.

［8］湖南省文物考古研究所 . 里耶秦简（贰）［M］. 北京：文物出版社，2017.

［9］化振红 . 深加工中古汉语语料库建设的若干问题［J］. 西南大学学报（社会科学版），2014，（3）.

［10］黄水清，王东波 . 古文信息处理研究的现状及趋势［J］. 图书情报工作，2017，（12）.

［11］李　斌，王　璐，陈小荷等 . 数字人文视域下的古文献文本标注与可视化研究——以《左传》知识库为例［J］. 大学图书馆学报，2020，（5）.

［12］李均明，刘国忠，刘光胜等 . 当代中国简帛学研究［M］. 北京：中国社会科学出版社，2019.

［13］李明杰，张纤柯，陈梦石 . 古籍数字化研究进展述评（2009-2019）［J］. 图书情报工作，2020，（6）.

[14] 李学勤 . 清华简九篇综述 [J]. 文物, 2010,（5）.
[15] 李学勤 . 简帛佚籍的发现与重写中国古代学术史 [J]. 河北学刊, 2013,（1）.
[16] 林剑鸣 . 简牍概述 [M]. 西安：陕西人民出版社, 1984.
[17] 刘　畅, 王东波, 胡昊天等 . 面向数字人文的融合外部特征的典籍自动分词研究——以 SikuBERT 预训练模型为例 [J]. 图书馆论坛, 2022,（6）.
[18] 刘　炜, 叶　鹰 . 数字人文的技术体系与理论结构探讨 [J]. 中国图书馆学报, 2017,（5）.
[19] 骈宇骞, 段书安 . 二十世纪出土简帛综述 [M]. 北京：文物出版社, 2006.
[20] 钱智勇, 周建忠, 童国平等 . 基于 HMM 的楚辞自动分词标注研究 [J]. 图书情报工作, 2014,（4）.
[21] 邱　冰, 皇甫娟 . 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）.
[22] 唐　杰, 刘　铭, 陈镱文 . 基于出土文献数据库的集外字数字化处理方法研究 [J]. 商洛学院学报, 2022,（2）.
[23] 王东波, 刘　畅, 朱子赫等 .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 [J]. 图书馆论坛, 2022,（6）.
[24] 王佳楠, 梁永全 . 中文分词研究综述 [J]. 软件导刊, 2021,（4）.
[25] 王　军 . 从人文计算到可视化——数字人文的发展脉络梳理 [J]. 文艺理论与批评, 2020,（2）.
[26] 王莉军, 周　越, 桂　婕等 . 基于 BiLSTM-CRF 的中医文言文文献分词模型研究 [J]. 计算机应用研究, 2020,（11）.
[27] 王彦辉 . 对简牍与秦汉史研究的几点思考 [J]. 史学月刊, 2011,（5）.
[28] 魏晓萍 . 数字人文背景下数字化古籍的深度开发利用 [J]. 农业图书情报学刊, 2018,（9）.
[29] 徐彤阳, 王　霞 . 语料库语言学视域下数据驱动的数字人文研究——以《数字人文季刊》为例 [J]. 图书馆论坛, 2021,（10）.
[30] 杨新生, 胡立生 . 基于隐马尔科夫模型的古汉语词性标注 [J]. 微型电脑应用, 2020,（5）.
[31] 俞敬松, 魏　一, 张永伟 . 基于 BERT 的古文断句研究与应用 [J]. 中文信息学报, 2019,（11）.
[32] 郑童哲恒, 李　斌 . 上古汉语分词与词性标注加工规范——基于《史记》深加工语料库的标注实践 [J]. 语言文字应用, 2023,（4）.
[33] 周　好, 王东波, 黄水清 . 古籍引书上下文自动识别研究——以注疏文献为例 [J]. 情报理论与实践, 2021,（9）.
[34] 朱　琳, 冯慧敏, 刘　铭等 . 数字人文视域下秦汉简牍文本挖掘研究——以里耶秦简（一、二卷）为例 [J]. 渭南师范学院学报, 2022,（6）.
[35] Fu, X., Yuan, T., Li, X., et al. Research on the method and system of word segmentation and POS tagging for ancient Chinese medicine literature[A]. *2019 International Conference on Bioinformatics and Biomedicine*[C]. 2019.
[36] Schuster, M. & Paliwal, K.K. Bidirectional recurrent neural networks[J]. *Transactions on Signal Processing*, 1997,（11）.
[37] Song, B., Chai, B., Zhang, Q., et al. A Chinese word segment model for energy literature based on neural networks with electricity user dictionary[A]. *2019 International Conference on Asian Language Processing (IALP)*[C]. 2019.
[38] Torregrossa, F., Allesiardo, R., Claveau, V., et al. A survey on training and evaluation of word embeddings[J]. *International Journal of Data Science and Analytics*, 2021,（2）.

（责任编辑　刘琪）