基于CRF模型的《里耶秦简》自动断句与分词研究

冯慧敏[1,2],郭帅帅[2],刘铭[2]

1. 山东农业工程学院基础课教学部,济南 250100
2. 西北大学科学史高等研究院,西安 710127

摘要　里耶秦简的数量是之前出土秦简的10倍,填补了秦朝历史记载中的诸多空白。将《里耶秦简》作为实验语料,探索基于CRF(条件随机场)模型的里耶秦简自动断句与分词方法。结合简文的实际特点,通过设置不同的特征模板,面向不同的任务验证模型序列标注的泛化能力;通过设置断句、分词一体化的对比实验,以选取性能更优的处理方案;同时设计了深度学习方法与预训练模型的对比试验。实验结果表明,CRF模型一体化的标注方案在各任务中的整体性能均有所提升,其中自动断句、分词的F_1值分别达到75.79%与94.44%,且速度快用时少,更适用于里耶秦简。

关键词　CRF模型;里耶秦简;自动断句;自动分词

出土文献是中华文化与文明的重要组成部分,其内容记录的真实性不仅可以补充传世文献记录的不足,同时可与传世文献的内容互相补正,对于史学研究具有特殊的价值。对出土文献进行深层次加工可便于对文献内容进行深度挖掘与知识发现,辅助研究者在海量的文献中发现其中隐含的新知识与解决问题的新视角,从而充分发挥出土文献的史料价值[1]。

分词是中文信息处理中的重要研究方向,是对文本数据进行组织与挖掘的基础,在自然语言领域中的研究已经臻于成熟。古代汉语与现代汉语有显著差异,行文过程是连续书写,并无断句,因此,

收稿日期:2023-05-29;修回日期:2023-10-30

基金项目:陕西省重点研发计划科研项目(2019ZDLGY17-03);西北大学研究生创新项目(CX2023045);山东农业工程学院科研启动经费项目(2024GCCZR-17)

作者简介:冯慧敏,讲师,研究方向为数字人文,电子信箱:gfhm_2013@163.com

引用格式:冯慧敏, 郭帅帅, 刘铭. 基于CRF模型的《里耶秦简》自动断句与分词研究[J]. 科技导报, 2024, 42(23): 135-144;
doi: 10.3981/j.issn.1000-7857.2023.05.00812.

断句是古文信息处理中必不可少的环节[2]。在传世文献领域,自动断句与分词任务已经取得了比较丰硕的研究成果,而在出土文献领域,相关研究却鲜少涉及。

里耶秦简出土于湖南省里耶古城,主要是秦朝洞庭郡迁陵县遗留的簿籍档案。里耶秦简数量庞大,共计约20余万字,约占现存秦简数量的70%;这批简牍史料翔实,几乎包含了所有文书类秦简的内容,是研究秦代县级行政机构的主要文献。由于埋藏条件恶劣,大部分简牍出土时存在严重的破损残断现象,为简文的整理工作带来困难。此外,里耶秦简的释文校正过程,几乎全是人工作业,仅前2卷的校释工作就持续了10年之久,耗时较长。并且相较于传世文献,里耶秦简具有出土文献封闭性更强、同质性更低的语料特点,已有的古文分词与断句方法不能完全适用于里耶秦简。因此,将里耶秦简作为实验语料进行自动断句与分词研究,不仅可以提高里耶秦简后3卷校释工作的效率,也可以为里耶秦简语料库的深入加工奠定基础,同时可以探索适用于出土文献的自动断句与分词方法。

目前在古汉语断句与分词领域,主要的研究成果集中于基于统计与深度学习的方法。其中深度学习虽然具有较强的学习能力,但同时也对数据量提出更高的要求。在基于统计学习的方法中,条件随机场(conditional random fields,CRF)具有判别式模型对数据量要求小、准确率高的优势,在古文断句、分词等信息处理任务中能展现出良好的性能。同时,结合里耶秦简残断问题严重、语料总量相对较小等特点,通过条件随机场模型,着力开展面向里耶秦简的自动断句与分词研究,以期服务于里耶秦简后3卷的校释工作及语料库的深入加工建设。

1 古文断句与分词相关研究

已有研究者在古文断句与分词领域开展相关研究。早期主要采用基于规则的方法,但存在效果欠佳、泛化能力较差等问题。目前,这一方面的研究主要以基于统计与深度学习的方法为主。

在自动断句研究方面,陈天莹等[3]最早使用前后文N-gram模型的统计方法进行古汉语自动断句。张合等[4]提出一种层叠式CRF模型,通过6字位标记集有效地对古文进行断句。张开旭等[5]基于CRF模型,引入互信息与t-测试差为特征,在《论语》的自动断句任务中取得了76.2%的F_1值。王博立等[6]、俞敬松等[7]分别采用GRU(gated recurrent unit)与BERT(bidirectional encoder representations from transformers)模型的深度学习方法开展古文断句实验。

相较于古文断句,古汉语自动分词的研究成果更加丰硕。石民等[8]基于CRF模型,对《左传》设置了分词与词性标注一体化的对比实验,开放测试的F_1值达到94.6%。梁社会等[9]以《孟子》为实验语料,基于CRF模型进行分词的F_1值达到较高水平。王嘉灵[10]采用CRF模型对《汉书》进行自动分词的实验中,F_1值达到94.4%。黄水清等[11]、严顺[12]、王晓玉等[13]、杨世超等[14]分别通过条件随机场模型对《左传》《论语》《孟子》等古籍开展自动分词研究,并取得了较好的实验结果。刘昱彤等[15]提出AP-LSTM-CRF算法,可有效地从大规模古汉语语料中发现新词。俞敬松等[16]将非参数贝叶斯模型和BERT模型相结合,在古汉语分词实验中表现出较好的泛化能力。程宁等[17]为避免错误扩散问题,基于BiLSTM-CRF模型设计了古汉语断句与词法分析一体化的标注体系。

基于深度学习的模型虽然可以通过训练语料自动学习特征,并在古文断句、分词等任务中展现出较高的性能,但同时也对训练语料的规模提出更高的需求。目前在已公布的里耶秦简中,实际可用的简文字数不足7万,并且由于残断破损情况较为严重,大部分的简文内容并不完整。同时,里耶秦简作为秦朝洞庭郡迁陵县遗留的文书档案,相对于传世文献,在语料资源上呈现出更加匮乏的特点。因此,深度学习模型并不完全适用于里耶秦简。

CRF模型虽然需要针对特定的语料人工制定特征模板,但是对所用语料的适应性更强,且对数据集规模的要求相对较小。湖南省文物考古研究所计划将里耶秦简分为5卷出版,目前仅出版前2卷。简文的整理首先需要对简牍上的文字进行释

读,然后对简文内容开展校释工作以便于后续研究使用,这一过程便与断句、分词任务息息相关。因此,基于里耶秦简的语料规模与简文特点,更加适用CRF模型。此外,通过制定特定的特征模板对里耶秦简的语料进行训练,可以使模型更加适用于后续3卷简文的校释工作

综合上述分析,将CRF模型应用于里耶秦简的自动断句与分词研究中。结合里耶秦简残断现象普遍的实际特点,设计了对称与不对称特征模板,并开展了断句、分词,以及断句分词一体化对比实验。同时加入了深度学习、预训练模型的对照实验,以寻找适合里耶秦简断句、分词的最佳方案,为里耶秦简后3卷的校释工作与语料库的深入加工提供可行的思路与方法。

2　CRF模型介绍

CRF模型可以综合考虑上下文信息,并通过全局归一化在分词、词性标注等任务中展现出良好的性能[18]。CRF在中文信息处理中的使用原理是将断句、分词等任务转化为字的序列标注问题。给定随机变量$X=\{x_1, x_2, \cdots, x_T\}$为观测序列,随机变量$Y=\{y_1, y_2, \cdots, y_T\}$是与之对应的标记序列。在给定$X$的条件下,$Y$的链式条件概率分布为

$$p(y|x)=\frac{1}{z(x)}\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x\right)+\sum_{i,l}\mu_l g_l\left(y_i,x\right)\right) \tag{1}$$

其中,

$$Z(x)=\sum_y\exp\left(\sum_{i,k}\lambda_k t_k\left(y_{i-1},y_i,x\right)+\sum_{i,l}\mu_l g_l\left(y_i,x\right)\right) \tag{2}$$

上式中,$Z(x)$是归一化因子,在所有可能的输出序列上进行求和,使得状态概率之和为1。t_k是依赖于当前和前一个位置的转移特征,g_l是仅依赖于当前位置的状态特征,λ_k与μ_l分别是上述2个特征函数的对应权重,通过训练数据获得[19]。CRF完全由特征函数t_k、g_l及其相应的权重λ_k与μ_l确定。

以分词任务为例,若定义标记集为$T=\{B,M,E,S\}$,文本"倉守陽敢言之"分词标注后的观测序列为$X=\{倉,守,陽,敢,言,之\}$,相应的标记序列为$Y=\{B,E,S,B,M,E\}$。CRF模型在训练时的结构如图1所示。

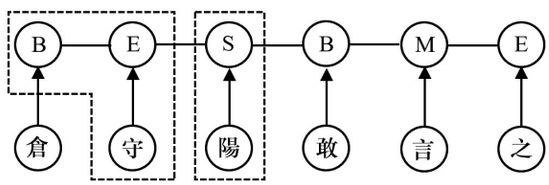

图1　CRF训练结构

其中,转移特征函数t_k通过训练数据学习标记之间的约束关系;状态特征g_l学习输入的文本序列与其对应标记的发射概率。CRF通过上述训练数据学到的内容如表1所示。最后,CRF通过学到的特征,对测试语料中每个字的标记标签进行预测从而达到自动分词的目的。在一体化实验中,通过设置2层标记集来同时完成自动断句与分词任务。例如,倉B-N,B表示该字位于词首,N则表示其后不应断句。

表1　特征函数学习内容

转移特征函数t_k	值	状态特征函数g_l	值
B, E	1	B, 倉	1
E, S	1	E, 守	1
S, B	1	S, 陽	1
B, M	1	B, 敢	1
M, E	1	M, 言	1
		E, 之	1

注:转移特征函数通过相邻两字的标记进行学习。因此,由文本"倉守陽敢言之"6字内容共学到5组转移特征内容。

此外,为进行对比实验,增加BiLSTM-CRF、BERT-BiLSTM-CRF模型。其中,BiLSTM-CRF模型是双向LSTM(long short-term memory)模型(bi-directional LSTM、biLSTM)与CRF模型的叠加。其中BiLSTM通过将前向LSTM与后向LSTM结合,以获得文本的历史信息和未来信息,然后将2个方向的信息组合作为BiLSTM的输出结果。BiLSTM-CRF模型将BiLSTM层的输出作为CRF层的输入,CRF通过学习标记序列之间的约束条件可以计算出标签间的最优序列[20]。

BERT模型是基于自注意力机制的预训练语言模型,通过结合上下文语境信息使得生成的字向量可以动态表示[21]。BERT-BiLSTM-CRF模型通过在BiLSTM-CRF模型上叠加BERT模型,能够结合上下文语境信息,捕捉句子中各个词之间的依赖关系,实现词的多义性表示。

3 实验语料与分词原则

3.1 语料简介

里耶秦简是2002年在湖南省龙山县里耶古城出土的秦代简牍,共约37000余枚,总计约20余万字。湖南省文物考古研究所按工作进度,已出版前2卷。其中第1卷包括1号井第5、6、8层出土的简牍,总计2627枚;第2卷录入第9层出土的3423枚简牍,2卷共计6050枚简牍。

武汉大学历史学院陈伟教授的研究团队,对应2卷《里耶秦简》分别同期出版了《里耶秦简牍校释》第1卷与第2卷(下简称校释),其中包含了大量的校订工作。本文使用的基础语料是湖南省文物考古研究所出版的2卷《里耶秦简》,同时为保证语料质量及完善程度,参照陈伟团队推出的校释,对简文内容进行校对,并采用其中新的校释及改释结果。

3.2 语料预处理

由于埋藏条件比较恶劣,甚至曾被进行过焚烧销毁处理,致使里耶秦简出土时存在严重的残断问题,如图2所示。在对简牍进行整理时,整理者为尽可能还原简文的实际面貌,在释文中做了诸多标注。但这些标注本身并不属于简文的内容,因此需要进行相应的处理才能作为最终的实验语料。根据实际情况,处理过程主要包括以下2个方面。

图2 简牍8-461正

1)简文内容处理。为确保释文与原简文的一致性,校释小组对根据残画拟释的字,用【】表示;为更明确地体现简文在简牍中的位置,在简文书写的转行处用罗马数字I、II、III等表示行号;针对简牍中的文字空白处,释文一律空出2个字符位置。为减少这些符号对实验结果的影响,将符号进行了删除处理。

简文中的集外字,为避免直接删除造成内容缺失,对简文中的126个集外字暂以希腊、拉丁等字母进行一对一替换。

2)语料清洗。部分简牍由于破损严重,其释文没有实际的文字信息,根据具体情况可分为如下3种类型。第1种类型的释文,内容只有☒、□、……符号,如简牍8-345(8-345是简号,表示第8层编号为345的简牍。):☒……□□□□□☒。其中"☒"表示残断符,"□"表示简文中无法辨认的字,一字一"□",若未释字数不确定,则用"……"表示。第2种类型的简牍,内容只有图案,在释文中只以"(圖案)"字样标出。如8-328的释文:(圖案)。第3种类型,因为原简已找到对应的缀连简,其内容录入到简号较小的简文中。如9-487的内容:【説明】與二九五綴合。对上述3种情况的简文

进行删除,共计269条,将剩余的5781条简文作为实验语料。

3.3　分词原则与词表

分词原则是分词工作的基础,现代汉语已经形成相对具有科学性、通用性较高的分词规范。但古汉语分词尚未形成统一的标准,而且古汉语与现代汉语差别较大,不能直接套用现代汉语的分词规范[22]。为避免训练数据的处理过程出现前后文分词不一致的情况,以及确保评估结果的准确性,将《信息处理用现代汉语分词规范》[23]《资讯处理用中文分词规范》[24]《北京大学现代汉语语料库基本加工规范》[25-26]作为参考,结合里耶秦简的实际情况,制定如下5条分词原则。

(1)分词单位主要是词,也包括里耶秦简中结合紧密、使用稳定的词组。

(2)分词时按照从长到短的顺序逐层切分。

(3)分词时遵循从宽原则,介于词与短语之间的词组,在不影响语义理解的情况下,不做切分处理。

(4)凡是收入《里耶秦简词表》的词条一般都视为一个切分单位,不再进行切分。

(5)由于简牍残断导致简文过于简短,致使内容难以理解,同时校释未给出相应的解释说明。结合里耶秦简作为上古汉语具有以单字词为主的特点,将这类简文按单字词进行切分处理。

《里耶秦简词表》的提取思路主要参考南京师范大学陈小荷团队对传世文献及其注疏的相关工作。注疏文献是相关学者人工分析的积累,可将其用于词汇收集,进而用于古文分词等信息处理工作[27]。校释是陈伟教授的团队在释文的基础上进一步校订、断读的成果,其中汇集了大量的字词校释成果。因此,将校释中的注释内容进行提取并整理成《里耶秦简词表》,作为开展分词工作的重要基础。

根据上述原则对训练语料进行人工处理,可最大程度上避免前后文分词不一致的问题,为模型提供质量较高的训练语料,同时可为模型评估提供可靠的标准。

4　实验设计与过程

4.1　实验设计

4.1.1　特征模板设计

利用CRF模型进行中文信息处理时,特征的选择以及特征模板的设计在很大程度上决定了模型的应用效果。CRF通过特征模板设置滑动窗口,以及添加二元同现等特征,对上下文进行表示[19]。

滑动窗口设置过小可能会导致信息丢失,若过大则可能会引起过拟合。为充分利用上下文信息,将滑动窗口设置为左右1~3个字,并引入二元同现特征进行对比实验。李成华等[28]在利用CRF模型对维吾尔语进行分词研究时,突破只使用对称模板的传统,采用了对称与非对称特征相结合的方法来设计特征模板。考虑到大部分简文内容不完整的特殊情况,也增设了采用对称与不对称特征模板的对比实验。

根据上述分析,共设计了10个特征模板,如表2所示,前6个是对称模板,后4个为不对称模板。针对不同的处理任务,通过具体实验选取效果最好的特征模板。

4.1.2　标记集设计

CRF在中文信息处理中的使用原理是将断句、分词等任务转化为字的序列标注问题。针对断句、分词等不同任务,设计了不同的标记集合。

表2　特征模板集合

模板1	模板2	模板3	模板4	模板5	模板6	模板7	模板8	模板9	模板10
1C	1C+2	2C	2C+2	3C	3C+2	1C+*2	1C+2*	2C+*2	2C+2*

注:C表示字符,nC表示滑动窗口为$\pm n$个字符,+2表示二元字符同现。以模板9为例,其表征含义为:滑动窗口为前后2个字符,当前窗口前面为二元字符同现,后面为一元同现的不对称模板。

基于CRF模型进行自动断句,主要是判断文本序列中的某个字符是否位于断句处,若是就标记为Y,否则标为N。对5-35的简文进行断句标记后的结果为:遷N陵N洞N庭Y以N郵N行Y。

在分词任务中,因为里耶秦简词表中词汇的平均长度为1.80,并且存在3个字及3个字以上的词,因此本文选用4词位的标注集合:T={B,M,E,S},4个标记分别表示词首、词中、词尾字及单字词。对5-35的简文标注词位后的格式为:遷B陵E洞B庭E以B郵M行E。

执行断句分词一体化任务时,需对词位与是否断句同时进行标注,因此需要3层标记符号。对5-35进行2层标注后的格式为:遷B-N陵E-N洞B-N庭E-Y以B-N郵M-N行E-Y。

通过设置不同的标记集,可以便于利用CRF模型处理不同的任务。一体化实验的原理是同时面向断句与分词的多分类任务,将两层标记结合到一起,能够为模型同时执行2个任务提供便利。

4.1.3 语料划分与评估指标

为避免偶然性,使得到的实验结果更加可靠,在所有实验中均采用5-折交叉验证的方法。将语料平均分成5份,轮流取其中的4份为训练语料,剩余1份用于测试。将5次评估结果的均值作为判定模型效果的依据。

参考中文信息处理中评估模型性能的常用指标,用人工校对后的断句、分词结果作为标准,将准确率P(precision)、召回率R(Recall)、F_1值作为衡量模型效果的指标:

$$P = \frac{S}{S_1} \tag{3}$$

$$R = \frac{S}{S_2} \tag{4}$$

$$F_1 = \frac{2 \times P \times R}{P + R} \tag{5}$$

其中,S表示模型正确断句或者分词的数量;S_1表示模型断句或者分词的总量;S_2表示语料中的断句或分词总量。P和R从不同的方面对模型效果进行评估,F_1则是基于P和R的调和平均值,本文将F_1值作为主要的评估指标对实验结果进行分析。

4.2 实验过程

4.2.1 自动断句实验

通过CRF模型将字面信息作为主要特征,分别利用上述10个特征模板进行实验,以筛选断句效果最好的特征模板,实验结果如表3所示。

表3 断句实验结果

模板	P/%	R/%	F_1/%
1	73.85	64.92	69.09
2	78.15	69.85	73.76
3	75.56	67.82	71.48
4	80.63	71.14	75.59
5	75.86	68.31	71.88
6	80.70	70.98	75.53
7	75.83	67.00	71.13
8	77.95	68.71	73.03
9	77.55	69.33	73.21
10	79.24	70.67	74.70

通过对比模型在不同特征模板上的实验结果,可以得到如下3个结论。

1)滑动窗口对实验结果的影响。通过对比模板1、3、5的F_1值,发现随着滑动窗口的增大,自动断句的F_1值呈现增大的趋势。滑动窗口设置为2比设置为1时的F_1值提升了2.39个百分点,而滑动窗口设置为3相对于设置为2时的F_1值却仅提升了0.4个百分点。说明滑动窗口越大,模型学到的特征越丰富,测试结果的准确率、召回率以及F_1值均有所提升;但随着滑动窗口不断增大,其对于模型表现能力的提升作用呈现下降趋势。

2)添加二元特征的影响。通过对比模板1与模板2、模板3与模板4、模板5与模板6这3组F_1值,发现增加二元同现特征后各测试集的F_1值分别提升了4.67、4.11、3.65个百分点。说明增加二元特征可以有效提升模型在断句任务上的标注能力。在滑动窗口为2、二元字符同现的情况下(模板4),断句的效果最佳,F_1值达到75.59%。

3)模板是否对称的影响。通过模板2、7、8与模板4、9、10这2组实验,发现模型采用对称模板的性能均优于采用不对称模板的表现。在使用不对称模板的情况下,在当前窗口后面比在前面增加二

元同现特征的效果相对好一些。

4.2.2　自动分词实验

里耶秦简的内容属于上古汉语,上古汉语的特点是以单字词为主,因此将语料按照单字词切分作为参照基线(baseline)。基于字面特征结合10个特征模板分别进行分词实验,结果见表4。

表4　分词实验结果

模板	*P*/%	*R*/%	F_1/%
1	91.31	91.50	91.40
2	93.97	94.90	94.43
3	91.00	91.26	91.13
4	93.26	94.41	93.83
5	90.65	91.02	90.83
6	92.82	94.05	93.43
7	93.60	94.21	93.91
8	93.50	94.09	93.80
9	92.93	93.75	93.34
10	92.93	93.69	93.31
Baseline	58.14	75.89	65.84

对表4中的实验数据进行分析,得到以下结论。

1)滑动窗口对分词结果的影响。不同于自动断句,随着滑动窗口的扩大,分词的F_1值反而呈现略微下降的趋势。句子本来就比词包含更长的文本序列,模型执行断句任务时,需要观察更多的上下文信息以学习更丰富的特征。滑动窗口对2个任务的不同影响正是说明了断句与分词在这一点上的不同。

2)二元特征的影响。当滑动窗口为1、2、3时,增加二元特征分别使F_1值提升了3.03、2.7、2.6个百分点,依然是提升分词性能的有效方案。其中,模板2的分词效果最好,F_1值达到94.43%,比参照基数高出28.59个百分点。

3)模板是否对称的影响。采用对称模板依然比不对称模板呈现出更好的分词性能,但是与断句任务不同的是,在当前窗口前面比在后面增加二元同现特征的效果相对好一些。

4)参照基数中的准确率为58.14%,召回率却达到75.89%,与先秦汉语以单字词为主的特点相契合。

4.2.3　参数*c*对模型性能的影响

使用CRF进行模型训练时,需要确定超参数*f*与*c*。*f*表示特征的最低频次,低于该阈值的特征将被舍弃。由于大部分里耶秦简残断问题突出,且语料总量不是很大,将*f*设为默认值1。参数*c*可以调整欠拟合和过拟合之间的平衡度,数值越大对训练数据的拟合程度越高。基于上述实验选定的最优模板,即模板4与模板2,分别针对2个任务调整参数*c*,以考察不同参数值对断句以及分词标注结果的影响程度。实验结果如表5、表6所示。

表5　断句实验参数*c*调整

参数	*P*/%	*R*/%	F_1/%
1.0	80.63	71.14	75.59
1.2	80.46	71.37	75.64
1.4	80.57	71.36	75.68
1.6	80.42	71.45	75.67
1.8	80.40	71.46	75.67
2.0	80.32	71.48	75.64

表6　分词实验参数*c*调整

参数	*P*/%	*R*/%	F_1/%
1.0	93.97	94.90	94.43
1.2	93.99	94.89	94.44
1.4	93.98	94.84	94.41
1.6	94.01	94.83	94.42
1.8	94.02	94.84	94.43
2.0	94.01	94.83	94.42

通过对比实验结果可以发现,在断句任务上,当参数*c*取值为1.4时,模型的性能改进最大,但也仅仅提升了0.09个百分点;在分词任务上,通过调整参数值,F_1值最高也只提升了0.01个百分点。说明在自动断句与分词任务上,参数*c*对模型标注效果的影响非常微弱,基本没有差别。

4.2.4　断句分词一体化实验

进行古文分词时,一般是在断句的基础上,再进行自动分词处理,两步走的方法通常会存在错误扩散问题[17]。因此,设置了断句、分词一体化实验,用于选取适合里耶秦简断句与分词任务的最佳方案。将10个特征模板分别用于一体化实验,并与两步走的实验作对比,结果如表7所示。

表7 一体化对照实验

模板	一体化断句 F_1值	仅断句 F_1值	一体化分词 F_1值	仅分词 F_1值
1	71.04	69.09	91.44	91.40
2	74.36	73.76	94.44	94.43
3	73.12	71.48	91.20	91.13
4	75.79	75.59	93.85	93.83
5	73.11	71.88	90.96	90.83
6	75.65	75.53	93.48	93.43
7	71.83	71.13	93.76	93.91
8	74.24	73.03	93.84	93.80
9	74.13	73.21	93.31	93.34
10	75.24	74.70	93.32	93.31

通过对比实验结果,发现采用一体化方式的整体效果优于分步处理的情况。通过10个特征模板的F_1值,对于自动断句任务,一体化方式比仅断句的F_1值平均高出0.91个百分点;对于自动分词任务,一体化比仅分词的F_1值平均高出0.02个百分点。因此,应用CRF模型对里耶秦简进行断句与分词,一体化的方式整体表现出更高的性能。

对一体化方式没有大幅度提升模型性能的原因进行分析,应是由于一体化实验需要对语料进行2层标注,更加细化的标注集使字符标记类别的划分更加复杂。同时简文具有大多不完整且部分过于简短的实际特点,2层标注在丰富语料特征的同时也增加了模型进行序列标注的难度。

虽然一体化方式对各任务性能的提升能力有限,但是这一方法可以最大程度避免分步处理的错误扩散问题。而且实验结果也表明该方法确实可以有效地应用于里耶秦简的自动断句与分词,同时可以提高对里耶秦简进行信息处理的效率。

4.3 对照实验

通过上述实验可知,CRF模型在里耶秦简的断句与分词任务中具有较好的性能表现,为更加确定CRF模型对于里耶秦简的适用性,加入BiLSTM-CRF、BERT-BiLSTM-CRF模型进行对照试验;同时,将“云梦睡虎地秦简”的内容融入原有语料,以探究语料规模、题材的变化对模型性能的影响。实验结果如表8所示。

表8 4种模型的对照实验结果

模型	自动断句			自动分词		
	P/%	R/%	F_1/%	P/%	R/%	F_1/%
CRF	80.61	71.52	75.79	93.99	94.89	94.44
BiLSTM-CRF	79.13	72.67	75.76	94.01	94.95	94.48
BERT-BiLSTM-CRF	77.74	72.96	75.27	94.60	95.10	94.85
CRF+睡简	62.81	54.25	58.22	90.58	91.10	90.84

由实验结果可知,在里耶秦简的自动断句与分词任务上,BiLSTM-CRF与BERT-BiLSTM-CRF模型并没有显著提升模型的性能。2个模型在自动断句任务上的F_1值分别下降了0.03与0.52个百分点;在分词任务上的F_1值分别提升了0.04与0.41个百分点,提升作用微弱,但2个模型的训练时长最高却达到CRF的300多倍。对加入云梦睡虎地秦简后的语料,利用CRF模型开展自动断句与分词任务,2个任务的模型性能反而出现明显降低的趋势,F_1值分别下降了17.46与3.6个百分点。云梦睡虎地秦简虽然也属于秦代简牍,但其题材主要涉及法律制度、医学等方面,与里耶秦简的公务文书题材差异较大,再次说明了简牍文献之间同质性较低、封闭性较强的特点。

5 结论

以已公布的里耶秦简1卷和2卷为语料,对面向出土文献的古文自动断句与分词进行了探究。根据简文残断情况严重、语料同质性较低、封闭性较强等特点,结合CRF模型对数据量要求较小、准确率较高的优势,开展了面向里耶秦简的自动断句与分词研究。研究结果表明,一体化的标注方案整体表现出更高的性能,自动断句、分词的F_1值分别

达到75.79%与94.44%。相较于BiLSTM-CRF、BERT-BiLSTM-CRF等深度学习模型,CRF模型在里耶秦简自动断句与分词任务中不仅性能上表现良好,而模型训练速度更快。研究成果可以有效地辅助后续3卷简文的校释工作与简文内容的深加工处理。

CRF模型相较于深度学习,虽然存在需要人工制定特征模板的局限,却恰好适用于出土文献封闭性强、同质性低的语料特点。同时CRF模型对数据量要求较小的特点,也更加匹配里耶秦简语料规模相对较小的实际情况,并在自动断句与分词任务中呈现出时间成本低、准确率高的优势。因此,针对特定的领域知识需要综合考虑模型的量级、训练成本与效率性能,以面向特定的任务需求选择更加适用的模型。

致谢:曲安京教授在本研究过程中提出建议与指导。

参考文献(References)

[1] 欧阳剑. 面向数字人文研究的大规模古籍文本可视化分析与挖掘[J]. 中国图书馆学报, 2016, 42(2): 66-80.
[2] 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, 61(12): 43-49.
[3] 陈天莹, 陈蓉, 潘璐璐, 等. 基于前后文n-gram模型的古汉语句子切分[J]. 计算机工程, 2007, 33(3): 192-193.
[4] 张合, 王晓东, 杨建宇, 等. 一种基于层叠CRF的古文断句与句读标记方法[J]. 计算机应用研究, 2009, 26(9): 3326-3329.
[5] 张开旭, 夏云庆, 宇航. 基于条件随机场的古汉语自动断句与标点方法[J]. 清华大学学报(自然科学版), 2009, 49(10): 1733-1736.
[6] 王博立, 史晓东, 苏劲松. 一种基于循环神经网络的古文断句方法[J]. 北京大学学报(自然科学版), 2017, 53(2): 255-261.
[7] 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, 33(11): 57-63.
[8] 石民, 李斌, 陈小荷. 基于CRF的先秦汉语分词标注一体化研究[J]. 中文信息学报, 2010, 24(2): 39-45.
[9] 梁社会, 陈小荷. 先秦文献《孟子》自动分词方法研究[J]. 南京师范大学文学院学报, 2013(3): 175-182.
[10] 王嘉灵. 以《汉书》为例的中古汉语自动分词[D]. 南京: 南京师范大学文学院, 2014.
[11] 黄水清, 王东波, 何琳. 以《汉学引得丛刊》为领域词表的先秦典籍自动分词探讨[J]. 图书情报工作, 2015, 59(11): 127-133.
[12] 严顺. 基于CRF的古汉语分词标注模型研究[J]. 江苏科技信息, 2016(8): 10-12.
[13] 王晓玉, 李斌. 基于CRFs和词典信息的中古汉语自动分词[J]. 数据分析与知识发现, 2017, 1(5): 62-70.
[14] 杨世超, 纪月, 赵立鹏. 基于条件随机场的古汉语分词研究[J]. 电脑知识与技术, 2017, 13(22): 183-184.
[15] 刘昱彤, 吴斌, 谢韬, 等. 基于古汉语语料的新词发现方法[J]. 中文信息学报, 2019, 33(1): 46-55.
[16] 俞敬松, 魏一, 张永伟, 等. 基于非参数贝叶斯模型和深度学习的古文分词研究[J]. 中文信息学报, 2020, 34(6): 1-8.
[17] 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, 34(4): 1-9.
[18] Sutton C, McCallum A. An introduction to conditional random fields[J]. Foundations and Trends in Machine Learning, 2011, 4(4): 267-373.
[19] Lafferty J, McCallum A, Pereira F C N. Conditional random fields: Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning. San Francisco: Morgan Kaufmann Publishers. 2001: 282-289.
[20] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. Computer Science, arXiv preprint arXiv: 1508.01991, 2015.
[21] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv: 1810.04805, 2018.
[22] 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, 40(3): 136-142.
[23] GB/T 13715-92, 信息处理用现代汉语分词规范[S]. 北京:中国标准出版社, 1993.
[24] 黄居仁, 陈克健, 陈凤仪, 等.《资讯处理用中文分词规范》设计理念及规范内容[J]. 语言文字应用, 1997(1): 94-102.
[25] 俞士汶, 段慧明, 朱学锋, 等. 北京大学现代汉语语料库基本加工规范[J]. 中文信息学报, 2002, 16(5): 49-64.
[26] 俞士汶, 朱学锋, 段慧明. 大规模现代汉语标注语料库的加工规范[J]. 中文信息学报, 2002, 16(6): 58-64.
[27] 陈小荷. 先秦文献信息处理[M]. 北京: 世界图书出版社公司北京公司, 2013: 13-69.
[28] 李成华, 孙雅婧, 张世娟, 等. 基于CRF模型的维吾尔语分词研究[J]. 中南民族大学学报(自然科学版), 2019, 38(4): 596-604.

Automatic sentence segmentation and word segmentation for Liye Qin Bamboo manuscripts based on CRF model

FENG Huimin[1,2], GUO Shuaishuai[2], LIU Ming[2]

1. Department of Basic Courses, Shandong Agricultural Engineering University, Jinan 250100, China
2. Institute for Advanced Study in History of Science, Northwest University, Xi′an 710127, China

Abstract Information processing of ancient Chinese seldom uses unearthed documents as corpus to carry out relevant research. The number of Liye Qin bamboo manuscripts reached ten times that of all the Qin slips unearthed before, which can fill many gaps in the historical records of the Qin Dynasty. In this paper , we used them as experimental corpus and explored the automatic sentence segmentation and word segmentation of unearthed documents based on the CRF model. We combined the actual characteristics of the corpus and set up different feature templates to verify the generalization ability of model sequence labeling on different tasks. We set up a joint approach to sentence segmentation and word segmentation as comparative experiment to select a better performance processing plan. At the same time, a comparative experiment was designed between deep learning methods and pretrained models. The results proved that the overall performance of the joint approach in each task was improved and that the F1-score of automatic sentence segmentation and word segmentation reached 75.79% and 94.44%, respectively. Since it′s faster and takes less time, this approach is more suitable for the Liye Qin bamboo slips. The research results can serve the proofreading work of the last three volumes of Liye Qin bamboo slips and the in-depth processing and construction of the corpus.

Keywords CRF model; Liye Qin bamboo manuscripts; automatic sentence segmentation; automatic word segmentation

（责任编辑 王微）