文章编号：1003-0077(2020)04-0001-09

基于 BiLSTM-CRF 的古汉语自动断句与词法分析一体化研究

程 宁[1]，李 斌[1,2]，葛四嘉[1]，郝星月[1]，冯敏萱[1]

(1. 南京师范大学 文学院，江苏 南京 210097；
2. 哈佛大学 计量社会科学研究所，美国 剑桥 02138)

摘 要：古汉语信息处理的基础任务包括自动断句、自动分词、词性标注、专名识别等。大量的古汉语文本未经标点断句，所以词法分析等任务首先需要建立在断句基础之上。然而，分步处理容易造成错误的多级扩散，该文设计实现了古汉语断句与词法分析一体化的标注方法，基于 BiLSTM-CRF 神经网络模型在四种跨时代的测试集上验证了不同标注层次下模型对断句、词法分析的效果以及对不同时代文本标注的泛化能力。研究表明，一体化的标注方法对古汉语的断句、分词及词性标注任务的 F_1 值均有提升。综合各测试集的实验结果，断句任务 F_1 值达到 78.95%，平均提升了 3.5%；分词任务 F_1 值达到 85.73%，平均提升了 0.18%；词性标注任务 F_1 值达到 72.65%，平均提升了 0.35%。

关键词：古文断句；分词；词性标注；BiLSTM-CRF；古汉语信息处理

中图分类号：TP391 **文献标识码**：A

A Joint Model of Automatic Sentence Segmentation and Lexical Analysis for Ancient Chinese Based on BiLSTM-CRF Model

CHENG Ning[1], LI Bin[1,2], GE Sijia[1], HAO Xingyue[1], FENG Minxuan[1]

(1. School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097, China;
2. Institute for Quantitative Social Science, Harvard University, Cambridge, MA 02138, USA)

Abstract: The basic tasks of ancient Chinese information processing include automatic sentence segmentation, word segmentation, part-of-speech tagging and named entity recognition. To avoid the error accumulation in the pipeline processing, this paper proposes a joint approach to sentence segmentation and lexical analysis. The BiLSTM-CRF neural network model is used to verify the generalization ability and the effect of sentence segmentation and lexical analysis on different label levels on four cross-age test sets. Experiments show that the joint model achieves improvements on the F_1-score of sentence segmentation, word segmentation and part-of-speech tagging: yielding 78.95% for sentence segmentation (with an average increase of 3.5%), 85.73% for word segmentation (with an average increase of 0.18%), and 72.65% for part-of-speech tagging (with an average increase of 0.35%).

Keywords: sentence segmentation of ancient Chinese; word segmentation; part-of-speech tagging; BiLSTM-CRF; ancient Chinese information processing

0 引言

词法分析是中文信息处理最基础的任务，包括自动分词、词性标注、命名实体识别等。古汉语信息处理的基础任务除了上述任务之外，还包括自动断句。中国古籍文本浩如烟海，大部分文本都是没有断句标示的，这为读者阅读和研究古籍带来很大的困难。利用先进的自然语言处理技术对古汉语进行自动断句、自动词法分析，不仅可以方便读者阅读古籍，而且对古籍整理、古汉语学科发展以及古汉语智能应用具有重要意义。古汉语信息处理大部分的研

收稿日期：2019-09-16 **定稿日期**：2019-10-16

基金项目：国家自然科学基金(71673143)；国家语委科研项目(WT135-24，YB135-61)；江苏省高校哲学社会科学优秀创新团队建设项目(2017STD006)

究都是针对某个特定的子任务来进行，如古汉语的自动断句、自动分词、词性标注、专名识别等，而且大部分学者所采用的研究方法及研究手段不尽相同，要完成整个古汉语信息处理的基础任务，需要依次完成各项子任务，这在很大程度上影响了机器处理效率，而且用机器分好的句子再进行分词以及词性标注等工作很容易造成标注错误的多级扩散，影响整体标注的准确率。本文设计实现了古汉语自动断句与词法分析一体化的标注体系，利用神经网络模型(BiLSTM-CRF)对断句、分词、词性信息进行联合学习。由于带标注的古汉语语料库相对匮乏，前人多根据某部专书来进行研究，语料规模相对较小，训练的模型不能很好地应用在其他类型的古文上。本文在已有资源的基础上，补充构建了源自四种成书于不同时代的古籍带标注语料，并利用神经网络模型在不同测试集上验证了一体化标注的效果。

1 相关研究

目前古文自动断句以及词级别的自动分析所采用的方法大体上可以分为三种：基于规则的方法、基于传统概率统计机器学习的方法以及基于深度学习的方法。

黄建年等[1]采用正则表达式抽取出古文断句规则，构建了农业古籍断句及标点的模式识别库，利用规则库对古籍进行断句和标点的实验，断句的正确率达到48%，标点的平均正确率达到35%。

基于规则的方法有两点不足之处：一是需要投入大量的精力去构建规则库；二是只能针对某一种特定的古文类型进行处理，针对其他古文类型时规则库就需要进行相应的扩充或删改，方法缺乏很好的泛化能力，不适用于大规模、跨时代、多体裁的古文处理。由于上述原因，且伴随着机器学习技术的发展，目前纯粹的规则方法在古汉语词法分析中使用得越来越少，结合隐马尔可夫模型(hidden Markov model，HMM)、最大熵模型(maximum entropy model，ME)、条件随机场模型(conditional random field，CRF)等传统的机器学习模型以及现在比较流行的深度学习技术来构建古文自动分析模型，已逐步成为主流方法。

张开旭等[2]将断句问题看作是序列标注问题，将条件随机场模型应用到古文自动断句任务处理上，在《史记》和《论语》语料上进行实验，F_1值接近80%，是一项比较有代表性的基于传统机器学习的古文断句方法研究。黄瀚萱[3]采用字标注的形式利用CRF模型在古文献(《论语》《孟子》《史记》等)上进行断句实验，同时采用HMM模型与之进行对比，得到CRF模型整体优于HMM模型的结果，而且文中做了大量训练集的对比实验，还探讨了模型跨时代的泛化能力。在词汇级自动分析任务方面，Hwee Tou Ng等[4]探讨了分词、词性标注一体化与分步走的优劣，得出基于字标注的一体化方法是最佳实验方案的结论。于江德等[5]将分词、词性标注、命名实体识别的标注标签纳入到统一的标签体系当中，用ME模型在Bakeoff2007的PKU语料上进行了封闭测试，验证了三位一体字标注的方法性能更优。石民等[6]采用CRF模型对《左传》的自动分词和词性标注进行探索，进行了分词词性标注一体化的对比实验。实验结果表明，一体化的标注方法使得分词和词性标注效果都有明显的提高。王晓玉等[7]选取字符分类和词典标记作为CRF模型的分词特征，有效提高了中古汉语分词的精度，分词结果的F_1值在开放测试中达到89%～95%。

基于传统机器学习的模型取得了一定的成果，但应用传统的机器学习方法进行古文自动分析存在两个问题，一是针对特定类型的古文，需要人工定制特征模板，耗时耗力；二是实验所采用的数据集规模相对较小，所定制的特征模板往往不能适应不同时代和体裁的古籍文本，模型的泛化能力有待进一步探究。采用深度学习的方法可以根据训练语料自动学习断句特征，从而避免复杂的特征工程。

Zheng等[8]提出利用神经网络模型来进行中文分词。金宸等[9]使用长短时记忆网络(LSTM)来解决中文分词问题，克服了传统神经网络难以处理长距离依赖的问题，取得了较好的分词效果。Yao等[10]采用双向长短时记忆网络(BiLSTM)进行汉语的分词任务，在PKU语料上取得了96.5%的F_1值，比传统分词方式得到的结果更优。冯蕴天等[11]采用深度信念网络(DBN)对汉语的命名实体进行抽取，在《人民日报》语料上取得了总体上优于条件随机场的精度。Wang等[12]提出基于神经网络模型的古文断句方法，采用循环神经网络(RNN)对大规模数据集进行断句实验，得到与CRF模型相媲美的性能。HAN等[13]提出了一种基于字根嵌入(radical embedding)的BiLSTM-CRF模型研究古文的自动断句问题，该研究对古文字符和对应的字根分别进行字向量的预训练，然后拼接为一个长向

量作为输入参数。实验在唐代墓志铭测试语料上得到81.34%的 F_1 值，比不加入字根进行训练有明显提升。

综合上述研究，本文将古文的断句、分词、词性及实体标签进行融合，形成一体化的标签体系，并基于 BiLSTM-CRF 神经网络模型实现了古汉语的一体化自动标注，且通过变更标注层次在不同时代的语料上对比了断句、分词、词性的实验结果。

2　模型介绍

2.1　BiLSTM-CRF

深度学习的兴起，尤其是适用于序列标注的 RNN[14] 模型及其变体极大地变革了整个自然语言处理学界的研究方法。RNN 可以看作是相同网络的多重叠加结构，其针对序列中的每一个元素都执行相同的操作，每一个操作都依赖于之前的计算结果。理论上，RNN 可以利用任意长的序列信息，但实际操作中只能回顾之前的几步。LSTM[15] 神经网络是一种特殊的 RNN，在原有的 RNN 模型基础上增加了输入门、遗忘门、输出门，神经元会选择性地遗忘对于当前输出无用的信息。其继承了 RNN 能够保留前序信息的优势，又克服了 RNN 无法真正捕捉文本中长距离依赖的问题。LSTM 这一模型结构在基于深度学习方法的自然语言处理任务中得到广泛应用。

BiLSTM 模型由 Schuster 等[16] 在 1997 年提出，目的是解决单向 LSTM 无法保留后文信息的问题，主要思想是在训练序列前向和后向分别设置两个 LSTM 结构，通过拼接两个方向的 LSTM 来捕捉前序和后序的信息，最大限度地保留了整个训练序列的信息。

本文使用的 BiLSTM-CRF 模型结构最早由 Huang 等[17] 提出。BiLSTM 层的输出是一个概率矩阵，这个概率矩阵是 BiLSTM 基于每个时刻上的最优结果得到的，这样输出的标签并没有考虑前一时刻标签对当前时刻标签的影响。例如，图 1 输入序列中出现了“孟子”一词，其中“孟”为词首，“子”为词尾，模型有可能将“孟”和“子”都预测为词首，这种情况在古汉语词法分析任务中是要避免的。CRF 是一个无向图模型的框架，它能够被用来定义在给定一组需要标记的观察序列的条件下，一个标签序列的联合概率分布。假定 X 是将要被标注的数据序列的随机变量，Y 是相应的标签序列的随机变量。例如，X 是自然语言的句子集合，Y 是标注这些句子的词性集合。随机变量 X 和 Y 是联合分布的，根据观测序列和标签序列对，构建了一个条件模型 $P(Y|X)$。在 BiLSTM 的输出层拼接 CRF 层，这样 BiLSTM 的输出序列就变成了 CRF 的观测序列，然后 CRF 计算整个序列在概率上的最优解，考虑到了序列元素标签之间的相互影响。

2.2　古汉语一体化自动分析模型架构

古汉语一体化自动分析模型架构主要包括三层：embedding 层、双向 LSTM 层和 CRF 层。架构如图 1 所示。

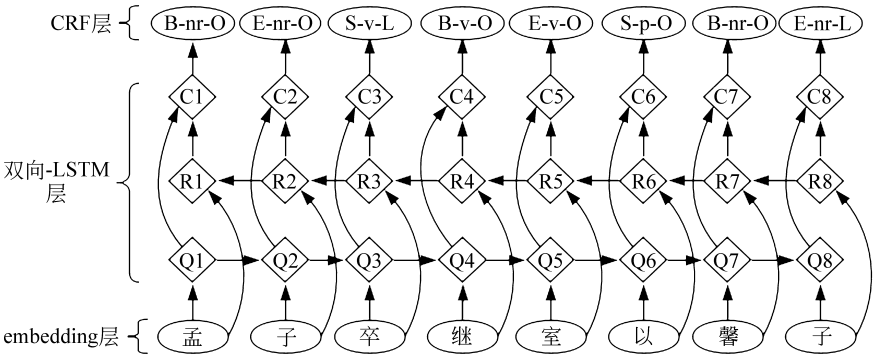

图 1　BiLSTM-CRF 模型架构

将训练好的字向量作为 embedding 层加入 BiLSTM，Q 代表当前字符上文，R 代表当前字符下文，C 代表当前时刻的上下文特征。CRF 层根据 BiLSTM 输出的概率矩阵给出预测结果标签。断句与词法分析一体化模型的每个预测结果标签由字在词中的位置、词性以及断句标示符号组成，例如，卒 S-v-L，S 表示单字词，v 表示词性为动词，L 表示该字处于句末。

(1) embedding 层

字向量是对文本中各个字符的分布式表示,即把语料中的字符映射成低维空间上的稠密向量(dense vector),即将每一个字都映射成一个 K 维向量。训练语料中的字符序列只需在 $N\times K$ 维的字向量矩阵中查找各字符对应的 K 维向量,然后将特征向量作为模型的输入即可。本文利用 word2vec[18] 模型在大规模古汉语生语料(包括四库全书及其他古汉语语料)上预训练字向量。

(2) 双向 LSTM 层

该层包含前向 LSTM 层和后向 LSTM 层,前向层(图 1 中的 Q 层)从前往后编码字符的上文信息,后向层(图 1 中的 R 层)从后往前编码字符的下文信息,然后将两层结合起来(图 1 中的 C 层),模型就可以表示出每个字符的上下文信息。LSTM 记忆单元具体工作流程用式(1)~式(6)描述如下:

遗忘门: $f_t=\sigma(\boldsymbol{W}_f\cdot[h_{t-1},x_t]+\boldsymbol{b}_f)$ (1)

输入门: $i_t=\sigma(\boldsymbol{W}_i\cdot[h_{t-1},x_t]+\boldsymbol{b}_i)$ (2)

$$\widetilde{C}_t=\tanh(\boldsymbol{W}_C\cdot[h_{t-1},x_t]+\boldsymbol{b}_C) \quad (3)$$

状态更新: $C_t=f_t*C_{t-1}+i_t*\widetilde{C}_t$ (4)

输出门: $o_t=\sigma(\boldsymbol{W}_o[h_{t-1},x_t]+\boldsymbol{b}_o)$ (5)

$$h_t=o_t*\tanh(C_t) \quad (6)$$

其中,f_t 表示 t 时刻遗忘门,i_t 表示 t 时刻输入门,C_t 表示忘记无用信息,增强有用信息后的细胞状态,o_t 表示输出门,x_t 和 h_t 表示 t 时刻的输入层向量和隐藏层向量。σ 和 tanh 是两种激活函数,$\boldsymbol{W}$ 表示权重矩阵,$\boldsymbol{b}$ 表示偏置向量。

(3) CRF 层

设双向 LSTM 层的输出序列为 $x=\{x_1,x_2,x_3\cdots,x_n\}$,CRF 层将此序列作为观测序列进行建模,给每一个字选择一个基于整句最优的字标注标签,最后利用维特比(Viterbi)算法获得整个句子最优的标签序列 $y=\{y_1,y_2,y_3\cdots,y_n\}$。在双向 LSTM 层加入 CRF 层,有效利用 BiLSTM 的特征学习能力和 CRF 对观测序列的建模能力,解决了 BiLSTM 无法给出整句最优标签的问题,且省去了 CRF 手工构建特征的工作。

3 语料构建和实验

3.1 语料库构建

根据不同的历史分期来选择古籍文本,分别从繁体版《左传》(汉)、《梦溪笔谈》(宋)、《阅微草堂笔记》(清,偏口语)、《清史稿》(民国)抽取相同大小的语料作为本文的实验数据集。分时代构建语料库是为了探讨模型基于各时代混合语料进行训练后其对不同时代文本标注的泛化能力。数据集在机器辅助分词和词性标注的基础上进行人工校对,使用 Kappa 进行标注一致性检验,Kappa 值高于 0.8,标注一致程度较高。词性标记规范参考 LDC 发布的《左传》语料库①,共计 21 个词性标签。实验数据集按照 8∶1∶1 的比例分为训练集、验证集和测试集,其中训练集是由《左传》《梦溪笔谈》《阅微草堂笔记》《清史稿》80%的语料组合而成的混合语料。基于这个混合语料来探讨模型对各时代文本的标注能力。实验语料将“:,。;!?”六种标点作为断句符,每两个断句符所分割的文本序列视为一个句子,其他标点忽略。表 1 是实验数据集的整体概况。

表 1 实验数据集

数据集	训练集			验证集			测试集		
	字数/万	词数/万	句数/万	字数	词数	句数	字数	词数	句数
左传(汉)	7.5	6.5	1.5	9 136	7 755	1 917	9 280	7 738	2 046
梦溪笔谈(宋)	7.8	6.3	1.3	9 483	8 384	1 662	9 825	8 378	1 643
阅微草堂笔记(清)	7.8	6.9	1.4	9 722	8 699	1 745	9 789	8 680	1 784
清史稿(民国)	8.0	5.7	1.2	10 248	8 851	1 651	9 991	8 159	1 432
合计	31.1	25.4	5.4	3.9 万	3.4 万	6 975	3.9 万	3.3 万	6 905

3.2 一体化的词位标记设计

Xue[19] 最早提出了一种基于字的序列标注学习方法,在其研究工作中使用 LL(词首)、LR(单字词)、MM(词中)、RR(词尾)四种标记来表示字的切分标注信息,从而首次将分词任务形式化地表达成

① LDC 左传语料库 https://catalog.ldc.upenn.edu/LDC2017T14

了序列化标注任务。本文借鉴这种字标注的方法构建了古汉语一体化分析的标注体系。对于模型来说，面对的问题实际上是一个标签多分类的问题，即将每一个字分到特定的标签类型下。

分词层（word segmentation layer，WS）：采用 B、I、E、S 四种标记，其中 B 代表当前字占据一个多字词的词首，I 代表当前字占据一个多字词的词中，E 代表当前字占据一个多字词的词尾，S 代表当前字是一个单字词，这种字标注序列经过转换就可以得到句子的分词结果。例如，

字标注：九 B 月 E，S 晉 B 惠 I 公 E 卒 S。S 懷 B 公 E 立 S，S

转换后：九月，晉惠公 卒。懷公 立，

词性层（POS tagging layer，POS）：标注每个字所属词的词性，同时将实体标签（人名 nr、地名 ns）融合到 POS 中。在 WS 基础上加入 POS，使得每个字对应其在词中的位置以及其所代表的词的词性或者实体信息。

九 B-t 月 E-t，S-w 晉 B-nr 惠 I-nr 公 E-nr 卒 S-v。S-w 懷 B-nr 公 E-nr 立 S-v，S-w

每个字分别标注分词标记和词性标记，中间用“-”连接，如上句“晉 B-nr 惠 I-nr 公 E-nr”，说明“晉”是一个人名的第一个字符，“惠”是一个人名的中间一个字符，“公”是一个人名的最后一个字符，这样就将“晉惠公”切分出来且识别为人名，人名的实体标签用“nr”来表示。

断句层（sentence segmentation layer，SS）：标注某个字符是否处于句末。在 WS 和 POS 的基础上加上 SS 层，使每个字符对应分词、词性、断句三层标签。

九 B-t-O 月 E-t-L 晉 B-nr-O 惠 I-nr-O 公 E-nr-O 卒 S-v-L 懷 B-nr-O 公 E-nr-O 立 S-v-L

如果语料中的某个字符处于断句处，如上句中的“月”“卒”“立”，则在其词性标记后面打上标签“L”。若某个字并不是处在断句处，则在其词性标记后面打上标签“O”。

在语料预处理过程中，三层标签类别（WS，POS，SS）可以有不同的处理方式：

三层标签有 WS＋POS＋SS（如，卒 S-v-L）。在此标注层次下可以计算各子任务的标注效果，如断句（SS）的效果。

两层标签有 WS＋POS（如，卒 S-v）。在此标注层次下可以计算分词（WS）的效果及词性标注的效果（WS＋POS）。

一层标签有 WS（如，卒 S）、SS（如，卒 L）。可以计算断句的效果或分词的效果。

3.3　评价指标

实验训练集用于模型的特征学习和训练，测试集用于验证模型自动标注的结果。对于自动标注结果的评价，在序列标注问题中，使用最为常用的评价指标 F_1 值（调和平均值）来衡量模型的效果。F_1 值由 P（准确率）和 R（召回率）算出，如式（7）所示。

$$F_1=\frac{2*P*R}{P+R} \tag{7}$$

其中，P 值的计算如式（8）所示。

$$P=\frac{\text{正确的标记数量}}{\text{机器的标记数量}} \tag{8}$$

R 值的计算如式（9）所示。

$$R=\frac{\text{正确的标记数量}}{\text{语料所有的标记数量}} \tag{9}$$

本文基于上述评价指标计算断句、分词、词性的标注结果。断句计算方式是基于句子而不是基于字，即根据标签“L”进行计算，机器和人工标注结果都为“L”则正确。分词、词性的计算方式是基于词而不是基于字。以词性标注为例，假设“孟子”一词预测为“孟 S-nr 子 S-nr”，尽管模型基于字标对了词性，但是分错了词，正确答案应该是“孟 B-nr 子 E-nr”。判断某字所属的词性是否正确，要先判断该字是否分对了词，即在分词正确的基础上进行计算。

3.4　实验设计与结果分析

实验一　验证古汉语一体化分析加入字向量的必要性，并考察不同维度的字向量对一体化标注结果的影响。一般来讲，字向量的维度越高，其中蕴含的语义特征就越丰富，但二者并不绝对呈正相关关系。本文基于近 15 亿字繁体版古汉语生语料（来源：四库全书及其他古汉语语料）进行字向量的预训练，工具选择 word2vec，模型选择 CBOW（continuous bag-of-words model），将字向量维度设置为 50 维、100 维、128 维和 200 维，测试语料选择《左传》测试集，标注层次采用“WS＋POS＋SS”，即断句与词法分析一体化的标注方法。通过在验证集上进行人工调参，最终所采用的超参数（hyper-parameter）如表 2 所示。

表2 实验超参设置

参数	参数取值
字向量维度	50/100/128/200
隐藏层数	1
隐藏单元数	200
最小样本数	64
dropout率	0.5
优化器	Adam
学习率	0.001

在BiLSTM-CRF结构中，通过在验证集上进行实验发现BiLSTM的层数对精度影响微弱，因此将模型的隐藏层数即BiLSTM的层数设为1。在序列标注任务中，隐藏节点数通常取200～600，这里取200作为参数。最小样本数设为64，每个样本大小控制在50到60之间。模型的优化采用在序列标注问题中效果较好的“Adam”算法。采用Dropout方法来降低过拟合。在BiLSTM层和全连接层之间加入参数为0.5的Dropout，这样可以弱化各个特征之间由于数据量太小导致的过多相互作用，从而使得模型的泛化能力最优、过拟合程度最低。实验结果如表3所示。

表3 断句与词法分析一体化 F_1 值(%)

字向量维度	断句效果	分词效果	词性标注效果
不加入字向量	82.16	88.23	78.36
50维	83.07	89.39	79.53
100维	83.89	90.19	80.59
128维	**84.11**	**90.24**	**80.88**
200维	83.58	89.83	80.42

从表3可以看出，加入字向量对古汉语的断句与词法分析任务都是有必要的，尤其是词性标注任务，提升了2.52个百分点。在字向量维度设定上，实验得出128维对古汉语一体化自动标注的效果最好。为了验证该维度下字向量训练效果，用余弦相似度来计算两个字向量之间的语义相关度：设字向量 $\boldsymbol{A}=(A_1,A_2,\cdots,A_n)$，$\boldsymbol{B}=(B_1,B_2,\cdots,B_n)$，余弦相似度的计算如式(10)所示。

$$\cos\theta=\frac{\sum_{i=1}^{n}(A_i\times B_i)}{\sqrt{\sum_{i=1}^{n}(A_i)^2}\times\sqrt{\sum_{i=1}^{n}(B_i)^2}} \tag{10}$$

其中，i 表示向量的维数，A_i 表示 $\boldsymbol{A}$ 所代表的这个字第 i 维的具体数值。以“也”和“曰”二字为例，计算结果如表4所示。

表4 语义相关度计算结果

与“也”语义最相关	与“曰”语义最相关
矣 0.662	云 0.696
之 0.659	謂 0.584
乎 0.658	也 0.514
謂 0.652	言 0.500
非 0.593	問 0.465
歟 0.584	耶 0.434
耶 0.571	荅 0.415
哉 0.563	答 0.413
以 0.525	為 0.412

实验二 为了测试BiLSTM-CRF模型标注古文的性能，本文采用了迭代膨胀卷积网络IDCNN[20](iterated dilated convolutions)和不拼接CRF层的BiLSTM模型与之对比。膨胀卷积网络(DCNN)最早由Yu等[21]提出并应用在图像语义分类问题上，IDCNN模型结构是在此基础上产生的，该模型结构借鉴CNN和RNN的优势，兼顾了并行化处理以及上下文特征提取的广度，在序列标注任务中同样得到广泛应用。本实验选择《左传》测试集，采用断句与词法分析一体化的标注方法，其他实验变量保持一致(如训练语料、字向量维度等)，考察不同模型在一体化的标注层次下分词任务的标注效果。实验结果如表5所示。

表5 不同模型对《左传》的分词效果

神经网络模型	《左传》测试集/%		
	P	R	F_1
IDCNN	88.25	89.28	88.76
BiLSTM	**89.39**	90.05	89.71
BiLSTM-CRF	89.37	**91.13**	**90.24**

通过对比实验结果发现，BiLSTM-CRF模型在古汉语分词任务上准确率低于BiLSTM模型0.02%，基本没有差别，在召回率上比不拼接CRF层的BiLSTM模型高出1.08%，在 F_1 值上比IDCNN高出1.48%，比BiLSTM高出0.53%。因此，对于古汉语的分词任务，BiLSTM-CRF模型的

性能在整体上高于 IDCNN 模型和 BiLSTM 模型。

实验三　针对《左传》《梦溪笔谈》《阅微草堂笔记》《清史稿》四种文本进行专门实验。本实验所采用的训练语料和测试语料来自同一种文本，以《左传》为例，训练语料和测试语料均来自《左传》。实验目的是探讨断句、词法分析一体化的模型对各类文本的建模能力，同时也为了与实验四基于混合语料的实验进行对比。实验标注层次采用“WS＋POS＋SS”，即断句与词法分析一体化的标注方法，实验参数与之前保持一致。实验结果如表 6 所示。

表 6　BiLSTM-CRF 模型在“WS＋POS＋SS”层次下对各语料的实验结果（%）

标注层次		左传			梦溪笔谈			阅微草堂笔记			清史稿		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
断句与词法分析一体化	断句效果	85.8	83.0	84.4	72.4	67.4	69.8	70.2	71.7	71.0	87.7	87.0	**87.4**
	分词效果	89.9	92.1	**90.9**	86.8	84.8	85.8	85.8	87.9	86.8	82.8	77.3	80.0
	词性效果	81.0	83.0	**82.0**	66.7	65.1	65.9	71.1	72.9	72.0	72.7	68.0	70.3

由于四种文本在成书年代和体裁上不同，模型针对各语料的实验结果差异较大。通过对比分词、词性、断句的 F_1 值，发现分词效果《左传》最好，《阅微草堂笔记》次之，《清史稿》效果最差；词性标注效果《左传》最好，《阅微草堂笔记》次之，《梦溪笔谈》效果最差；断句效果《清史稿》和《左传》相对较好，在精度上远超《阅微草堂笔记》和《梦溪笔谈》。针对模型标注错误进行分析发现：《梦溪笔谈》含有大量不重复的各学科领域的专业术语，例如，与音乐学科相关的“南呂調皆用七聲：下五、高凡、高工、尺、高一、”，其中“下五”“高凡”等词都属于专有名词，这些专有名词数据相对稀疏，模型很难学到相关特征，这是导致其词性标注效果较差的主要原因。

实验四　本实验从两个维度来进行设计：①横向上探讨基于混合语料的模型在相同标注层次下不同时代语料上的标注差异，并结合实验三的实验结果考察模型对不同时代文本的泛化能力；②纵向上对比同一测试语料不同标注层次的标注差异，验证分词、词性、断句一体化标注方法的有效性。实验模型采用 BiLSTM-CRF，训练语料采用混合语料，字向量维度选择 128 维。实验结果如表 7 所示。

表 7　基于混合语料的 BiLSTM-CRF 模型在不同标注层次下对各语料的实验结果（%）

标注层次		左传			梦溪笔谈			阅微草堂笔记			清史稿		
		P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
仅断句		83.6	79.5	81.5	69.0	64.4	66.6	68.1	68.7	68.4	86.8	83.9	85.3
仅分词		88.8	91.4	90.0	87.4	85.8	86.6	85.8	87.1	86.4	81.2	77.2	79.2
词性	分词效果	88.9	91.2	90.0	86.9	86.2	86.6	85.5	86.8	86.1	82.1	77.4	**79.7**
	词性效果	79.2	81.2	80.2	67.6	65.6	**66.6**	72.2	73.2	72.7	71.8	67.7	69.7
断句与词法分析一体化	断句效果	86.5	81.9	**84.1**	72.0	71.1	**71.5**	73.7	73.0	**73.3**	85.2	88.8	**86.9**
	分词效果	89.4	91.1	**90.2**	87.6	85.9	**86.8**	86.3	87.0	**86.6**	81.7	77.0	79.3
	词性效果	80.1	81.7	**80.9**	67.4	65.4	66.4	72.5	72.9	**72.7**	72.8	68.6	**70.6**

通过对比模型在不同标注层次下对各测试集的标注结果，有四个方面的结论。

（1）通过观察相同标注层次下各测试集的 F_1 值，发现基于混合语料作为训练集的模型对各测试集的标注结果并不均衡，与实验三的实验结果相类似。通过将本实验在断句与词法分析一体化标注层次下的实验结果与实验三进行对比发现：《梦溪笔谈》在断句、分词与词性标注任务上分别提升了 1.7、1.0 和 0.5 个百分点；《阅微草堂笔记》在断句、词性标注任务上分别提升了 2.3、0.7 个百分点；《清史稿》在词性标注任务上提升了 0.3 个百分点；《左传》的各标注任务上略有下降。该实验结果表明，基于混合语料的一体化模型学到了各语料的一些同质性特征，提升了对某些测试集的标注性能，但与此同时

各语料的差异性也干扰了模型的综合判断，降低了对某些测试集的标注性能，模型针对各时代文本进行一体化标注的泛化能力还有待提升。

(2) 通过观察不同标注层次下分词任务 F_1 值在各测试集上的表现，断句与词法分析一体化的标注层次整体最优。仅分词的标注层次无论在哪种测试集上 F_1 值都低于一体化的模式，这说明断句、词法分析一体化的标注方法对古汉语的分词任务有所提升。

(3) 通过观察不同标注层次下断句任务 F_1 值在各测试集上的表现，发现断句与词法分析一体化的标注层次整体最优，这说明一体化的标注方法对古汉语的断句任务有所提升。以《左传》测试集为例，一体化标注层次下断句 F_1 值比单独断句提升2.6%。这些提升并不局限于《左传》测试集，在其他测试集上同样有类似的提升效果。这说明在古汉语自动断句任务上，断句、词法分析一体化的标注方法比分步走的标注方法性能更优。

(4) 将断句与词法分析一体化的标注层次与词性标注层次进行对比可以发现，一体化的标注层次在大部分测试集上的 F_1 值效果好于词性标注层次，以《左传》测试集为例，一体化标注层次下的分词效果比词性标注层次下的分词效果提升0.2%，词性标注效果提升0.7%。该实验结果验证了断句与词法分析一体化的分词、词性标注效果，在整体上要高于不加断句信息的分词、词性标注效果。

综合(2)～(4)的分析，发现一体化的标注体系对断句、分词、词性标注任务都有性能提升(F_1 值)，而且这种提升不局限于某一种测试集，具体提升效果如表8所示。

表8 断句与词法分析一体化标注体系对各子任务的提升 F_1 值(%)

标注任务	左传	梦溪笔谈	阅微草堂笔记	清史稿
断句	+2.6	+4.9	+4.9	+1.6
分词	+0.2	+0.2	+0.2	+0.1
词性标注	+0.7	−0.2	+0	+0.9

尽管一体化的标注方法对分词、断句、词性标注任务的性能提升效果有限，但实验验证了一体化标注方法的可行性。该标注方法可以避免单个任务标注错误的多级扩散，以《左传》为例，如果进行分步处理，要先对其进行断句，然后在断句的结果上进行分词、词性标注，这样会造成错误的多层累加，整体性能不如一体化的标注方法；而且断句、词法分析一体化的标注方法可以大大提高古汉语在词、句级别信息处理的效率。

4 总结与展望

本文设计实现了古汉语断句与词法分析一体化的标注体系，基于 BiLSTM-CRF 神经网络模型在《左传》《梦溪笔谈》《阅微草堂笔记》《清史稿》四种跨时代的测试集上验证了不同标注层次下模型对断句、分词、词性标注的效果以及一体化标注模型对不同时代文本标注的泛化能力。研究表明一体化的标注方法在古汉语的断句、分词及词性标注任务上性能均有提升。综合各测试集的实验结果，断句任务 F_1 值达到78.95%，平均提升了3.5%；分词任务 F_1 值达到85.73%，平均提升了0.18%；词性标注任务 F_1 值达到72.65%，平均提升了0.35%。

未来的研究将扩大语料规模和改进模型，着眼于大规模跨时代语料环境下的深度学习模型设计，加入注意力机制和迁移学习方法，探索模型针对各时代文本的适应能力，以开发出性能更好的跨时代和文体的古汉语一体化分析系统。

参考文献

[1] 黄建年，侯汉清. 农业古籍断句标点模式研究[J]. 中文信息学报，2008，22(4)：31-38.

[2] 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 (10)：1733-1736.

[3] 黄瀚萱. 以序列标记方法解决古汉语断句问题[D]. 台湾：交通大学硕士学位论文，2008.

[4] Hwee Tou Ng, Jin Kiat Low. Chinese part-of-speech tagging: One-at-a-time or all-at-once? Word-based or character-based? [C]//Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing. 2004：277-284.

[5] 于江德，胡顺义，余正涛. 三位一体字标注的汉语词法分析[J]. 中文信息学报，2015，29(6)：1-7.

[6] 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)：39-46.

[7] 王晓玉，李斌. 基于 CRFs 和词典信息的中古汉语自动分词[J]. 现代图书情报技术，2017，1(5)：62-70.

[8] Zheng X, Chen H, Xu T. Deep learning for Chinese word segmentation and POS tagging[C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013：647-657.

[9] 金宸，李维华，姬晨，等.基于双向 LSTM 神经网络模型的中文分词[J].中文信息学报，2018，32(2)：29-37.

[10] Yao Y，Huang Z. Bi-directional LSTM recurrent neural network for Chinese word segmentation[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，2016：1197-1206.

[11] 冯蕴天，张宏军，郝文宁，等. 基于深度信念网络的命名实体识别[J]. 计算机科学，2016，43(4)：224-230.

[12] Wang B，Shi X，Tan Z，et al. A sentence segmentation method for ancient Chinese texts based on NNLM[C]//Proceedings of the CLSM. Singapore，2016：387-396.

[13] Han X，Wang H，Zhang S，et al. Sentence Segmentation for classical Chinese based on LSTM with radical embedding[J]. The Journal of China Universities of Posts and Telecommunications，2019，26(2)：1-8.

[14] Rumelhart D E，Hinton G E，Williams R J. Learning representations by back-propagating errors[J]. Nature，1986，323 (6088)：533-536.

[15] Hochreiter S，Schmidhuber，J. Long short-term memory[J]. Neural Computation，1997，9(8)：1735-1780.

[16] Schuster M，Paliwal K K. Bidirectional recurrent neural networks[J]. IEEE Transactions on Signal Processing，1997，45(11)：2673-2681.

[17] Huang Z，Xu W，Yu K. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv：1508.01991，2015.

[18] Mikolov T，Chen K，Corrado G，et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv：1301.3781，2013.

[19] Xue N. Chinese word segmentation as character tagging [J]. Computational Linguistics and Chinese Language Processing，2003，8(1)：29-48.

[20] Strubell E，Verga P，Belanger D，et al. Fast and accurate entity recognition with iterated dilated convolutions[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. 2017：2670-2680.

[21] Yu F，Koltun V. Multi-Scale Context Aggregation by Dilated Convolutions[J]. arXiv preprint arXiv：1511.07122，2015.

程宁(1993—)，硕士，主要研究领域为计算语言学。
E-mail：chengninmo@foxmail.com

葛四嘉(1994—)，硕士，主要研究领域为计算语言学。
E-mail：sijiage007@gmail.com

李斌(1981—)，通信作者，博士，副教授，主要研究领域为计算语言学。
E-mail：libin.njnu@gmail.com