文章编号：1003-0077(2019)11-0057-07

基于 BERT 的古文断句研究与应用

俞敬松[1]，魏 一[1]，张永伟[2]

（1. 北京大学 软件与微电子学院，北京 100871；
2. 中国社会科学院 语言研究所，北京 100732）

摘 要：古汉语与现代汉语在句法、用词等方面存在巨大的差异。古文句与句之间通常缺少分隔和标点符号，现代读者难以理解。人工断句有助于缓解上述困境，但需要丰富的专业知识，耗时耗力。计算机自动断句有助于加速对古文的准确理解，从而促进古籍研究以及中华文化的弘扬。除自动断句，该文还尝试了自动标点任务。该方案自行预训练古汉语 BERT(Bidirectional Encoder Representations from Transformers)模型，并针对具体任务进行微调适配。实验表明，该方案优于目前深度学习中的主流序列切割 BiLSTM+CRF 模型，在单一文本类别和复合文本类别测试集上的 F_1 值分别达到 89.97%和 91.67%。更重要的是，模型表现出了很强的泛化能力，未参与任何训练的《道藏》测试集上的 F_1 值依然可达到 88.76%。自动标点任务仅使用少量较为粗糙的带标点文本训练集时 F_1 值为 70.40%，较 BiLSTM+CRF 模型提升 12.15%。两任务结果均达到当前最佳，相关代码和模型已经开源发布。

关键词：自动断句；自动标点；BERT；微调

中图分类号：TP391　　**文献标识码**：A

Automatic Ancient Chinese Texts Segmentation Based on BERT

YU Jingsong[1], WEI Yi[1], ZHANG Yongwei[2]

(1. School of Software and Microelectronics, Peking University, Beijing 100871, China;
2. Institute of Linguistics, Chinese Academy of Social Sciences, Beijing 100732, China)

Abstract: Ancient Chinese differs from modern Chinese in words and grammar. Since there are no explicit marks among sentences in ancient Chinese texts, today's readers find it's hard to understand them. It is also difficult and requires expertise in a variety of fields to segment ancient text. We investigate to perform automatic texts segmentation and punctuation based on recent deep learning technologies. By pre-training a BERT (Bidirectional Encoder Representations from Transformers) model for ancient Chinese texts ourselves, we get the current state-of-the-art results on both tasks via fine-tuning. Comparing to traditional statistical methods and current BiLSTM+CRF solution, our approach significantly outperforms them by achieving F_1-scores of 89.97% and 91.67% on small-scaled single category corpus and large-scaled multi-categories corpus, respectively. Especially, our approach shows its good generalization ability by achieving F_1-score of 88.76% on a fully new Taoist corpus. On the punctuation task, our method F_1 score reached 70.40%, which exceeded the baseline BiLSTM+CRF model by 12.15%.

Keywords: automatic texts segmentation; automatic punctuation adding; BERT; fine-tuning

0 引言

中华文字源远流长，战国竹简中就曾发现标点符号的使用，但古人习惯于不加标点。直到近代，大量古文篇章依然由连续的字符序列组成，缺少标点[1]。古人需要根据自己的知识背景和学术流派在阅读时自行断句。古文自动断句及标点是指根据古代汉语句子特点，结合现代汉语的标点符号用法，让计算机自动切割、断开连续的文本字符序列为句，然后添加标点[2]。中国历史久远，文本类型繁杂，高适应性的自动断句非常困难。标点时，由于借用现代汉语标点体系，处理古籍时难免会不适应。古文断句是古文整理与研究中最基本的任务，也可以说是

收稿日期：2019-05-15　　**定稿日期**：2019-06-18

最重要任务之一。

随着机器学习的兴起，人们开始尝试将统计学习方法应用于古文自动断句。文献[3]指出，这些研究大都以所获取的标点过的古籍的80%作为训练集，20%作为测试集的方法进行研究。在实际应用中，需要处理的文本领域和范围却可能是未曾见过的，根据我们的测试，之前论文所报告的方法在处理其他新类型文本时的表现远低于其在论文测试集上报告的水平。本文将新的深度学习方法用于古文自动断句及标点任务，除期望在测试集上有更好的结果外，还期望模型有更好的鲁棒性和泛化能力。

1 相关文献

早期的古文句读研究主要基于规则与统计方法。黄建年等通过规则方法对农业古籍进行断句[4]。由于规则的覆盖度和泛化能力太差，统计方法得到了更多的关注。陈天莹等[5]通过n-gram提取上下文信息以预测断句位置。张开旭等[6]在条件随机场(conditional random fields, CRF)模型的基础上，引入互信息与t-测试差，在《史记》和《论语》语料上进行训练和测试，最终F_1值接近80%。Huang等[7]使用音韵特征，利用条件随机场模型进行断句。许京奕[8]根据邻接搭配强度间的关系特征，在《史记》语料测试集上F_1值达到82.16%。这些传统统计学习方法需要人工设计特征，过多依赖于先验知识，因此导致模型规模、表达力都非常有限。

深度学习方法中循环神经网络(recurrent neural network，RNN)为时序性网络结构，通常是序列标注问题的首选。Huang等[9]使用双向长短时记忆网络(bidirectional long short term memory，BiLSTM)+CRF模型，在词性标注、命名实体识别等序列标注任务上取得了当时的最优结果。王博立等[10]将RNN+CRF模型引入古文断句任务，通过使用基于GRU[11](gated recurrent unit)的双向RNN，利用大量语料进行端到端训练，将F_1值提升到74.75%。

为解决RNN的网络结构不能并行计算以及可能发生的梯度消失与梯度爆炸问题，Vaswani等[12]提出Transformer模型。Transformer不但可以实现并行计算，其在机器翻译等许多任务上取得了优异的结果。2018年Devlin等[13]基于Transformer提出BERT模型，利用大量文本，通过类似于完形填空的Cloze任务[14]以及上下句判断任务，获得预训练词向量与句向量，随后根据不同的任务进行微调，在包括命名实体识别、问答等多项自然语言处理任务中取得当时的最优结果。

本文将BERT方法引入古文领域的研究，自行训练中国古文字级别的BERT模型，获得高质量向量表示，随后同样通过微调进行古文断句。我们以双层BiLSTM+CRF作为基线方法，对比BERT预训练+微调方法。本文在算法研究上完成了断句和标点两项任务。为了让本文成果在北京大学国家重大学术文化项目《儒藏》工程①中落地使用，本文提出真实语料环境下的长篇章级别自动断句算法并进行了测试。前人文献中还未曾对此有过报道。

2 模型设计

2.1 基线模型和本文的改进

2.1.1 基线模型框架

本文选择双层BiLSTM+CRF模型作为基线模型。首先将输入的文字转为向量表示，随后通过双层BiLSTM模型提取特征信息，最后将BiLSTM提取的特征通过CRF层解码，得到全局最优标注。较之王博立等的模型[10]，本文将门控单元由GRU换为LSTM[15]，再加深神经网络的层级，以获得更优的文字特征表示。考虑到Zaremba等[16]已经通过实验证明，将dropout[17]机制引入循环神经网络，可在语言建模等任务上提升LSTM表现，本文还为基线模型添加了dropout机制，以抑制过拟合问题。

2.1.2 BiLSTM模块

单向门控RNN通过对于输入序列的时序处理，虽可以提取更好的特征，但无法关注到下文信息。双向RNN通过对输入序列进行正反两个方向的处理，最终获得某一时刻输入的上下文信息。通过多层RNN叠加可使特征提取进一步优化。LSTM在RNN的基础上引入了门控单元，使其可以更好地存储上下文信息，在一定程度上解决了由于长序列输入带来的梯度消失与梯度爆炸问题。

2.1.3 CRF模块

CRF[18]是无向概率图模型，与有向概率图模型相比，CRF通过使用全局归一代替局部归一，使其可以获得全局最优的结果，避免了序列标注任务中的独立性假设以及可能出现的标记偏执问题。

---

① http://www.ruzang.com/，北京大学《儒藏》工程主页，教育部哲学社会科学研究重大课题攻关项目。

对于给定输入的标记概率可形式化表示如式(1)所示。

$$P(Y \mid X) = \frac{\prod_i \psi_i(x, y)}{Z} \quad (1)$$

其中，Z 为归一化因子。

这种对输入序列先通过门控 RNN 单元提取特征，随后用 CRF 进行解码的模型框架已经在基于序列标注的词性标注、命名实体识别等任务中取得了优异的结果[9]。

2.2　BERT＋微调模型

古文断句以字为分割。比如判断句式“……者……也”中的“者”“也”后面就是明显的断句位置。对于自动化模型来说，如果可通过大量“阅读”古汉语文章，发现文字和文字序列的蕴含特征，就可以通过有指导的训练过程实现自动断句，以及在断句位置添加标点符号。因此通过预训练获得高质量的字向量挖掘古文字的隐含信息是优化古文断句任务的前提。在此基础上通过对目标文本设计训练集，添加并优化分类模型就可以解决特定目标任务。

本文所提出的模型框架共分为两部分：预训练模块与微调模块。首先参考 BERT 架构在大量未标记文本上通过文字位置、上下文文字预测任务和上下句连贯性判断任务获得字向量。然后使用带有断句标记以及标点符号的文本进行训练，建构自己的断句和标点模型。

2.2.1　BERT 模块

本文采用 Google 开源的 BERT 模型实现①构造古文字向量。BERT 的设计基于 Transformer 网络结构。Transformer 对当前的输入，分别计算出 Key，Query 和 Value 向量，并基于上述向量对每个输入使用注意力机制，以获得当前输入与周围语境的关系和自身所包含的信息。通过多层累加和多头注意力机制，不断获取当前输入更为合适的向量表示。由于在输入层加入了位置信息，Transformer 可以做到并行高速计算。BERT 基于 Transformer 引入了 Input Mask 和片段标记，将大量无标注文本转化为有指导的学习问题。Input Mask 用于指导模型根据上下文判断当前词是否恰当，片段标记用于判断上下句是否连贯。例如，

输入：【CLS】北【MASK】有 鱼【SEP】其 名【MASK】鲲【SEP】

CLS 连贯性标记：IsNext

输入：【CLS】北 冥【MASK】鱼【SEP】不【MASK】其几千里也【SEP】

CLS 连贯性标记：NotNext

2.2.2　微调优化模块

考虑到 BERT 模型训练的工作量巨大[13]，本文先在约 1 千万字的小规模语料上进行实验，并与多种既有方法进行初步比对。我们尝试了 BERT＋直接分类输出，BERT＋CRF 以及 BERT＋BiLSTM＋CRF 三种架构，发现更复杂的模型并没有获得更好结果。三者的结果相差不到 0.3%，BERT＋BiLSTM＋CRF 与 BERT＋直接二分类的结果相差不足 0.05%。我们遵循最简原则，选择 BERT＋直接二分类方案，断句模型如图 1 所示，标点模型如图 2 所示。

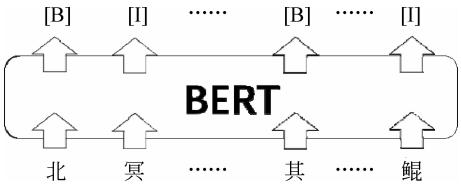

图 1　BERT＋直接分类输出

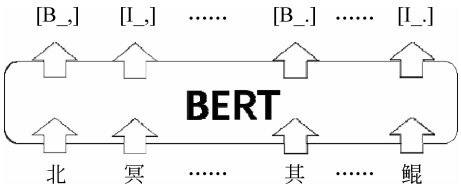

图 2　BERT＋标点输出

2.2.2　微调模块输出

王博立等[10]与张合等[19]选择了考虑文字在句中位置的六元标记作为断句标志。例如，句子“尧让天下于许由，曰：‘日月出矣，而爝火不息；’”使用六元标记的结果为：[B] [M] [M] [M] [E3] [E2] [E] [S] [B] [E3] [E2] [E] [B] [M] [E3] [E2] [E]。本文对此也进行了相关实验，发现六元标记方法不能提升 BERT 整体模型的表现，其 F_1 值与使用下文提到的二元标记的模型的 F_1 值相差不到 1%。

断句和标点任务重点在于对每个字进行分类。对于断句任务，我们通过上述实验，最终选择最简洁的二元标签组，T＝{[B]，[I]}，只对句子的开头和句子中的内容进行标注。上一段提到的句子将被标记为：[B] [I] [I] [I] [I] [I] [I] [B] [B] [I] [I] [I] [B] [I] [I] [I] [I]。

对于添加标点任务，我们选择的标点集为 punc＝

① https://github.com/google-research/bert

{(,),(。),(?),(!),(、),(:) ,(;)},其对应的分类标记为 Tpunc={[B_,],[I_,],[B_。],[I_。],[B_?],[I_?],[B_!],[I_!],[B_、],[I_、],[B_:],[I_:],[B_;],[I_;]}。比如,句子"天之苍苍,其正色邪?",其标点分类标记为:[B_,][I_,][I_,][I_,][B_?][I_?][I_?][I_?]。

3 实验结果

3.1 语料介绍

因为无法获得现成语料,本文使用从网上获取的古文语料进行实验。文本类别包括史集、诗集、儒藏、集藏、子藏和道藏。根据实验目的不同,对语料按类和文本数量进行了归并。验证小规模语料情形的《史藏》训练集只包括单一文体类别;验证大规模语料情形的混合训练集使用《史藏》《诗藏》《儒藏》《集藏》和《子藏》五种类别,但将《道藏》排除在外。任何测试集中的文本都不包括在任何一种训练集中,也不会参与任何预训练任务。测试集类别与训练集对应。最后,我们以独立的《道藏》文本验证本文工作泛化能力。为获得更为可靠的结果,我们选用的语料规模,无论是训练还是测试均远大于包括王博立等[10]在内的前人使用的数量。例如,我们的测试集合在千万字量级,而之前的工作测试集合仅数十万字而已。详情如表 1 所示。

标点任务由于获取的带标点古文文本较为杂乱,多种文本类别并存,标点使用非常不一致,有只有逗号、句号的,也有使用了所有现代标点符号的。我们最终仅使用相对标点质量较高的《道藏》文本进行标点任务的训练和测试。《道藏》标点语料的训练集与测试集规模如表 2 所示。

表 1 断句训练与测试集规模

数据集	总字数/千万	字表大小	内容来源
史藏训练集	15	16 172	史集
史藏测试集	0.4	8 481	史集
混合训练集	37	18 656	史集、诗集、儒藏、集藏、子藏
混合测试集	1.2	11 430	史集、诗集、儒藏、集藏、子藏
道藏测试集	2.4	10 300	道藏

表 2 标点训练与测试集规模

数据集	总字数/千万	字表大小	内容来源
道藏标点训练集	2.6	10 886	道藏
道藏标点测试集	0.3	7 032	道藏

我们从多个互联网来源获取古文语料,包括殆知阁[①]等。虽然语料规模大,但绝大部分都是繁简体混合的。古文繁简体转换非常困难,目前尚无法高准确率自动化进行。相较于纯粹的繁体或简体文本,繁简混合实际上给工作带来了更大的难度,所以我们保留文本原始形态,未作任何处理。

对于断句任务,基于不同的训练集划分与标签集我们总计训练测试了 6 个模型:①使用混合训练集的 BERT+微调模型(以下简称 BERT_A);②使用混合训练集的基线模型(以下简称 R-CRF_A);③使用混合训练集和六元标记的基线模型(R-CRF_A6);④仅使用《史藏》训练集的 BERT+微调模型(以下简称 BERT_H);⑤仅使用《史藏》的基线模型(以下简称 R-CRF_H);⑥仅使用《史藏》数据和六元标记的基线模型(R-CRF_H6)。

对于标点任务,我们只对比了 BERT+微调模型与 BiLSTM+CRF 模型。

3.2 测评标准

对于模型的表现,我们使用根据准确率(以下简称 P)和召回率(以下简称 R)计算的 F_1 值作为评价标准:即

$$P=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FP}} \tag{2}$$

$$R=\frac{\mathrm{TP}}{\mathrm{TP}+\mathrm{FN}} \tag{3}$$

$$F_1=\frac{2\times P\times R}{P+R} \tag{4}$$

其中,TP 为预测正确的断句标志数量,FP 为预测错误的断句标志数量,FN 为未预测出的断句标志数量。

断句任务中的 BERT_A 模型只采用 B、I 两种标签。上述公式中 TP 统计的是句子开始标记,即 B 标签。使用六元标记的各基线模型 TP 统计的是[B]和[S](单字成句标记)。

标点任务中使用微平均作为衡量指标,计算公式如式(5)~式(7)所示。

① http://www.daizhige.org/

$$P_{\text{micro}} = \frac{\sum_{i=1}^{k} \text{TP}_i}{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{FP}_i} \tag{5}$$

$$R_{\text{micro}} = \frac{\sum_{i=1}^{k} \text{TP}_i}{\sum_{i=1}^{k} \text{TP}_i + \sum_{i=1}^{k} \text{FN}_i} \tag{6}$$

$$F_{1_{\text{micro}}} = \frac{2 \times P_{\text{micro}} \times R_{\text{micro}}}{P_{\text{micro}} + R_{\text{micro}}} \tag{7}$$

因为对每种标点使用二元标注集合，所以上述公式中 TP 统计的是各种标点的开始符号，比如逗号的开始符号[B_,]。

3.3　实验结果与分析

3.3.1　自动断句任务

真正的古文正文由长串文字序列组成，无法直接用来测试断句。测试数据的构造方法文献[10]并未提及。考虑到门控 RNN 虽在一定程度上解决了长距离依赖问题，但如果序列长度过长，计算依然是不可能的。为不失一般性，我们将连续的几句或十几句合并为不超过 64 个字的小段落作为测试对象（参数选择理由参见本文第 4 节）。不同模型的具体结果如表 3～表 5 所示。

在断句任务的任何一种测试方法中，BERT_A 是所有模型中表现最优者，无论准确率、召回率还是 F_1 值均有更好的表现。文本类别对结果几乎没有影响。随着不同类别的文本加入训练，实验难度增加，模型能力是否随之增强需要实验予以证明。我们分别设置了小规模单一类型文本测试集与大规模复合类别文本测试集分别进行实验。与 BERT_H 相比，可以看出，随着训练语料类别的增多、语料规模的增加，训练得到的模型效果也越好。在从未参与任何训练的《道藏》测试集上，BERT 模型依然显示了较好的泛化能力。

表 3　单一文本类别小规模语料《史藏》测试结果

模型	准确率/%	召回率/%	F_1/%
R-CRF_H	81.70	76.51	79.02
R-CRF_A	83.55	77.43	80.37
R-CRF_H6	82.37	80.79	81.57
R-CRF_A6	81.83	83.44	82.62
BERT_H	88.49	87.51	87.99
BERT_A	**89.52**	**90.42**	**89.97**

表 4　复合文本类别大规模语料混合测试结果

模型	准确率/%	召回率/%	F_1/%
R-CRF_H	78.62	68.63	73.29
R-CRF_A	86.79	78.48	82.42
R-CRF_H6	78.90	75.26	77.04
R-CRF_A6	85.83	84.62	85.22
BERT_H	87.34	82.08	84.63
BERT_A	**92.32**	**91.07**	**91.67**

表 5　模型泛化能力《道藏》测试集结果

模型	准确率/%	召回率/%	F_1/%
R-CRF_H	78.17	67.79	72.61
R-CRF_A	84.44	74.04	78.90
R-CRF_H6	77.48	73.73	75.56
R-CRF_A6	81.12	79.68	80.39
BERT_H	85.68	80.33	82.92
BERT_A	**89.54**	**87.99**	**88.76**

与二元标注的 BiLSTM＋CRF 基线模型相比，在大规模复合类别文本测试集上，使用小规模语料的《史藏》训练集的 BERT_H 比 R-CRF_H 提升近 10%，即使 R-CRF_A 训练使用混合训练集，也依然与之无法比较。由此可见，本文提出的方法在断句任务上可以获得更优结果。即使训练语料较小也可以学到更强的泛化能力。

实验表明，在六元标记模式下，分类任务由二分类变为难度更大的多分类任务，然而对基线模型的正确率提升有限。特别是混合训练集情形，由于文体丰富、词汇量较大，六元标记模式效果还更差一些。随着更多先验知识的引入，召回率却提升明显。但总体上依然与 BERT＋微调方案存在差距。

3.3.2　自动标点任务

如前所述，限于条件，只使用了标点质量相对较高的《道藏》进行了训练和测试。测试集同样由字符总数不超过 64 的若干连续句子组成。本文模型与基线模型的结果对比如表 6 所示。

表 6　《道藏》标点测试集结果

模型	准确率/%	召回率/%	F_1/%
BiLSTM＋CRF	60.39	56.26	58.25
BERT＋微调	**70.92**	**69.88**	**70.40**

本文模型结果较 BiLSTM+CRF 模型结果，F_1 值提升了 12.15%。自动标点任务结果与断句任务相比还存在一定差距。一方面，训练语料规模过小，另一方面语料中标点质量较差，规则不统一、随意性也较强。

4 方法的实际应用

4.1 篇章断句

在最近的古文断句研究论文中，都将其作为序列标注问题进行处理。多数序列标注问题，如分词或命名实体识别等任务，输入句子的长度一般不会过长。由于古文既没有断句，亦没有段的概念。若自动断句技术投入使用，就需要处理古籍中动辄数千字的超长段落，而且输入长度很难固定。BERT 模型随着序列长度的增长，所需要训练时长成二次方增长[13]，硬件要求难以满足。

通过对大规模训练语料的统计，我们发现在这么大的古文训练集中，超过 99%的句子长度不会超过 21 个字，具体分布详情见图 3。基于此，我们最终决定将一个文本片段（并非真正意义上的段落）的最大长度限制在 64 个字以内。这就意味着一次处理长度为 64 的序列，最少也可以切分为 2～3 句。在此基础上，我们提出滑动窗口式古文断句方法。每次模型处理长度不超过 64 字的片段，但只取前一或两个断句标记前的部分输出，剩余的部分归并到后续语料中再切割出 64 个字组成处理片段，以保证最后一个或几个断句标记不受不完整输入的影响。反复执行这一动作，直到全文文本处理完毕。解码时不回溯尝试其他可能的选择，这样既可以保证断句质量，也确保了预测速度。

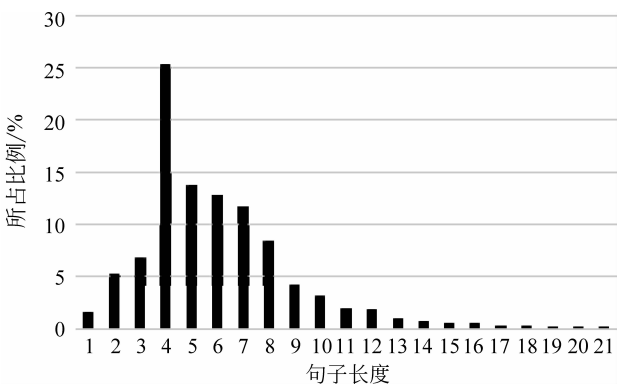

图 3 不同句长所占比例

4.2 应用示例

以下示例段落选自《道藏—吕祖三尼医世说述管窥》，本文不包括在任何训练语料中。原文及输出结果如下（缺失和错误的标记以脚注的形式指出）。

【原文及标点】 要知天地无心，非块然两大也，言其自然之道耳！作善而百祥随之，作不善而百殃随之，皆自然之道也。而致殃致祥之柄？乃自人操之而天随之，是可见人有转移造化之力矣。其故何哉？

【断句结果，断句位置显示句号】 要知天地无心。非块然两大也。言其自然之道耳。作善而百祥随之。作不善而百殃随之。皆自然之道也。而致殃致祥之柄。乃自人操之而天随之。是可见人有转移造化之力矣。其故何哉。

【标点结果】 要知天地无心，非块然两大也，言其自然之道耳。① 作善而百祥随之，作不善而百殃随之，皆自然之道也。而致殃致祥之柄，②乃自人操之，③而天随之。④ 是可见人有转移造化之力矣。其故何哉？

5 结论

本文使用 BERT+微调模型解决古文断句与标点问题，在三类测试集上与 BiLSTM+CRF 基线模型进行对比，文本模型均获得了更佳表现。主要结论如下：

（1）BERT+微调模型基本不受古文训练语料类别的影响。但更多类别、更大数量的训练集依然可以提升模型表现。

（2）BERT+微调模型在训练集受限的情况下依然有良好表现，而且泛化能力出色。只要保证基本的语料规模，模型就可以处理全新文本。

从这两点分析以及断句结果看，BERT 模型应该是学到了中国古文句读的精髓。本文方法与基线模型对比，还进一步减少了人工特征设计工作。依靠滑动窗口式的长篇章段落处理方法，本文的工作已经可以正式投入古籍研究工作，正式成为古籍研究者的人工智能助手。

本文基于 Tensorflow[20] 的相关代码和模型已经开源发布在 Github 上⑤。从本文实验结果看，在

① 原为“！”，“。”应该也可以

② 原为“？”，但另一专家标注版本同“，”

③ 多加“，”，但似乎并不为错

④ 此处应为“，”，“。”不正确

⑤ https://github.com/ToolsForAncientChineseText/Text-segmentation

多数情况下，直接使用我们提供的模型即可获得较好结果。

参考文献

[1] 赵巧丽. 略谈古人对古书句读的认知机制[J]. 今日南国(理论创新版)，2008(3)：187-188.

[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] 顾磊，赵阳. 古籍智能整理研究现状及存在的问题[J]. 图书馆学研究，2016(9)：54-58.

[4] 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.

[5] 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(03)：192-193，196.

[6] 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.

[7] Hen Hsen Huang，Chuen Tsai Sun，Hsin Hsi Chen. Classical Chinese sentence segmentation[C]//Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing，2010.

[8] 许京奕. 古汉语文本自动句读研究[D]. 北京：北京大学博士学位论文，2011.

[9] Huang Z，Xu W，Yu K，et al. Bidirectional LSTM-CRF models for sequence tagging[J]. arXiv preprint arXiv：1508.019901，2015.

[10] 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(02)：255-261.

[11] Chung J，Gulcehre C，Cho K，et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[J]. arXiv preprint arXiv：1412.3555，2014.

[12] Vaswani A，Shazeer N，Parmar N，et al. Attention is all you need[J]. Neural Information Processing Systems，2017：5998-6008.

[13] Devlin J，Chang M，Lee K，et al. BERT：Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv：1801.04805，2018.

[14] Taylor W L. Cloze Procedure：A new tool for measuring readability[J]. Journalism Bulletin，1953，30(30)：415-433.

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

[16] Zaremba W，Sutskever I，Vinyals O，et al. Recurrent Neural Network Regularization[J]. arXiv preprint arXiv：1409.2329，2014.

[17] Srivastava N，Hinton G E，Krizhevsky A，et al. Dropout：A simple way to prevent neural networks from overfitting[J]. Journal of Machine Learning Research，2014，15(1)：1929-1958.

[18] Lafferty J D，Mccallum A，Pereira F，et al. Probabilistic models for segmenting and labeling sequence data[C]//Proceedings of the 18th International Conference on Machine Learning，2002.

[19] 张合，王晓东，杨建宇，周卫东. 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(9)：3326-3329.

[20] Abadi M，Agarwal A，Barham P，et al. TensorFlow：Large-scale machine learning on heterogeneous distributed systems[J]. arXiv preprint arXiv：1603.04467，2016.

俞敬松(1971—)，博士，副教授，主要研究领域为自然语言处理、人工智能辅助语言学习。
E-mail：yjs@ss.pku.edu.cn

张永伟(1984—)，博士，副研究员，主要研究领域为自然语言处理、语料库语言学。
E-mail：zhangyw@cass.org.cn

魏一(1995—)，硕士研究生，主要研究领域为自然语言处理。
E-mail：weiyi9506@pku.edu.cn