

文章编号: 1003-0077(2021)04-0008-08

## 基于深层语言模型的古汉语知识表示及自动断句研究

胡韧奋<sup>1,2</sup>, 李 绅<sup>1</sup>, 诸雨辰<sup>3</sup>

(1. 北京师范大学 中文信息处理研究所, 北京 100875;

2. 北京师范大学 汉语文化学院, 北京 100875;

3. 北京师范大学 文学院, 北京 100875)

**摘 要:** 古文句读不仅需要考虑到当前文本的语义和语境信息, 还需要综合历史文化常识, 对专家知识有较高要求。该文提出了一种基于深层语言模型(BERT)的古汉语知识表示方法, 并在此基础上通过条件随机场和卷积神经网络实现了高精度的自动断句模型。在诗、词和古文三种文体上, 模型断句  $F_1$  值分别达到 99%、95% 和 92% 以上。在表达较为灵活的词和古文文体上, 模型较之传统双向循环神经网络方法的  $F_1$  值提升幅度达到 10% 以上。实验数据显示, 模型能较好地捕捉诗词表达的节奏感和韵律感, 也能充分利用上下文信息, 实现语序、语法、语义、语境等信息的编码。在进一步的案例应用中, 该方法在已出版古籍的断句疑难误例上也取得了较好的效果。

**关键词:** 古汉语; 自动断句; 深层语言模型

**中图分类号:** H087

**文献标识码:** A

## Knowledge Representation and Sentence Segmentation of Ancient Chinese Based on Deep Language Models

HU Renfen<sup>1,2</sup>, LI Shen<sup>1</sup>, ZHU Yuchen<sup>3</sup>

(1. Institution of Chinese Information Processing, Beijing Normal University, Beijing 100875, China;

2. College of Chinese Language and Culture, Beijing Normal University, Beijing 100875, China;

3. School of Chinese Language and Literature, Beijing Normal University, Beijing 100875, China)

**Abstract:** Sentence segmentation of ancient Chinese texts is a very difficult task even for experts in this area, since it not only relies on the sentence meaning and the contextual information, but also requires historical and cultural knowledge. This paper proposes to build knowledge representation of ancient Chinese with BERT, a deep language model, and then construct the sentence segmentation model with Conditional Random Field and Convolutional Neural Networks. Our model achieves significant improvements in all of the three ancient text styles. It achieves 99%, 95% and 92%  $F_1$  scores for poems, lyrics and prose texts, respectively, outperforming Bi-GRU by 10% in lyrics and proses which are more difficult to segment. In further case studies, the method achieves good results in the difficult cases in published ancient books.

**Keywords:** ancient Chinese; automatic sentence segmentation; deep language model

## 0 引言

汉语典籍记载和文献编纂有着悠久的历史, 涵盖政治、历史、哲学、文学等各领域。中国人也尤其注重古籍的整理与利用, 《永乐大典》《四库全书》都

是历史上重要的文献整理工程。然而, 古典文献的一个重要特点是不使用标点符号, 这与古人因声求气、涵咏情性的文化有关, 却给现代读者带来了困难。因而古文句读便成为当代古籍整理中一项非常重要的工作。

然而, 古文句读却对专家知识有极高要求, 因为

收稿日期: 2019-09-09 定稿日期: 2019-10-19

基金项目: 国家自然科学基金(62006021); 教育部人文社会科学研究青年基金(18YJC751073); 国家社会科学基金(18ZDA238)

句读不仅需要考虑当前文本的意义和语境信息,还需要综合历史文化常识。宋代大儒朱熹读韩愈文章,便有“然不知此句当如何读”<sup>[1]</sup>之惑。近代经学大师黄侃在致陆宗达的信中也表示“侃所点书,句读颇有误处,望随时改正。”<sup>[2]</sup>

在现有的古籍数据中,大部分尚未实现句读。据本文统计,殆知阁古代文献藏书 2.0 版语料库规模约 33 亿字,其中仅 25% 左右数据包含标点,可见古籍整理是一项浩大的工程,自动句读技术有强烈的现实需求。

自然语言处理技术的发展使得自动断句成为可能。张开旭等人<sup>[3]</sup>提出一种基于条件随机场的古文自动断句方法,对《论语》和《史记》的文本进行实验,其《论语》断句的  $F$  值达到 76% 左右,而《史记》断句的  $F$  值则在 68% 左右。王博立等人<sup>[4]</sup>提出一种基于循环神经网络的古文断句方法,采用基于 GRU 的双向循环神经网络进行古文断句,该模型对古文断句的  $F$  值达到 74%~75%。由于现有模型对文本意义和语境信息理解并不充分,断句效果距离实用尚有距离,还需要进一步提升。

近年来,ELMO、BERT 等预训练语言模型极大地提升了语言信息表示的效果,并在文本分类、语言推断、文本生成、阅读理解等一系列自然语言处理任务中取得了突出的成绩提升<sup>[5-6]</sup>。然而,现有的语言模型多基于大规模百科或新闻语料训练,缺乏古汉语语言知识编码。为了改进现有古文断句模型,促进古汉语信息处理技术的发展,本文在 33 亿字古汉语语料库上训练深层语言模型,实现了古汉语知识的高效表示,并在此基础上利用条件随机场和卷积神经网络学习句读模型。系统在诗、词和古文三种文体上开展了测试,其  $F_1$  值分别达到 99%、95% 和 92% 以上,在断句难度较高的词和古文文体上,本文方法较之王博立等人的双向 GRU 模型界值提升幅度达到 12% 以上。

与前人工作相比,本文的贡献体现在以下几个方面:首先,通过深层语言模型实现了高质量的古汉语知识表示,使模型在“理解”的基础上句读;第二,根据断句任务中语言特征和标签信息之间的关系,设计了深层语言模型+条件随机场(BERT+CRF)、深层语言模型+卷积神经网络(BERT+CNN)两种序列标注方法,较之传统神经网络方法取得了显著的性能提升。从评测效果看,本文提出的断句方法在多种类型语料中均取得了实用级效果,并能有效检测出已出版古籍中的断句错误。此

外,为了更好地服务古籍整理和文献研究,我们构建了在线古诗文断句工具<sup>①</sup>。

## 1 基于深层语言模型的古汉语知识表示

### 1.1 神经词向量表示方法

语言知识表示是自然语言处理技术的重要基础,在现有的模型中,通常以词语为单位进行语言特征表示。为了将词义信息编码到词语表示中,Mikolov 等人<sup>[7]</sup>提出了一种神经词向量(neural word embeddings)表示方法,并发布了训练词向量的工具包 Word2Vec。其模型基于语言学家 Harris<sup>[8]</sup>提出的词义分布假说:上下文相似的词,其意义也相近。具体来说,词向量的训练基于大规模语料库,依次取中心词和它左右两边的上下文词,通过神经网络模型构建两种预测方式:利用上下文词语预测中心词(CBOW 模型),利用中心词预测上下文词语(Skip-gram 模型)。通过训练,可以得到定长的稠密实数词向量,其维度通常为 50~300,每一维均由一个实数表示。与 Word2Vec 类似,利用词语和上下文的共现信息,Pennington 等人提出了 GloVe 模型<sup>[9]</sup>,Levy 和 Goldberg 提出了正值逐点互信息模型(PPMI)和奇异值分解模型(SVD)<sup>[10]</sup>。

神经词向量表示能够较好地捕捉词语的语法和语义性质,例如,“麦克风”和“话筒”向量的 cosine 相似度极高,甚至还可以实现“国王”-“男人”+“女人” $\approx$ “王后”、“天”-“天天”+“人人” $\approx$ “人”这样的词法、词义推理<sup>[11]</sup>。这种词向量表示方法被广泛运用到文本分类、机器翻译、语义搜索、自动问答等各种自然语言处理任务中,大大提升了自然语言理解和生成的效果。

在古汉语特征表示领域,Li 等人<sup>[11]</sup>基于《四库全书》训练了古汉语字义表示,图 1 给出了其向量空间经 PCA 降维后一个局部区域的汉字分布情况。由图可见,该区域语义表示与数字、时间等概念密切相关,形成了较为明显的数字词簇和时间词簇,如单位“千、百、万”,数词“一”到“九”,序数词“甲乙丙丁戊己庚辛”,时间词“日夜月夕”“年岁”等。

然而,传统的词向量表示方法仍然面临一个突出的问题:即仅能为每个词获取一个词向量,无法区分同形词和多义词的不同义项。在古代文言文表

① <https://seg.shenshen.wiki/>



图1 古汉语神经词向量示例

达中,往往单字成词,每个单字词可承载的意义极为丰富,其同形词和一词多义现象比现代汉语更为突出,这不仅为现代人理解文言文含义造成困难,也为计算机表示古汉语带来了挑战。为了有效解决这个问题,本文引入了基于深层语言模型的古汉语知识表示方法。

## 1.2 深层语言模型表示方法

### 1.2.1 BERT 模型

本文参考 Devlin 等人提出的 BERT 模型学习古汉语知识表示,并将自动断句作为下游任务对整个网络进行微调(fine-tuning)。BERT 模型可以基于大规模语言数据学习上下文敏感的词语和句子表示(contextual embeddings)。与 Word2Vec 同一词形仅能生成一个词向量不同,预训练的 BERT 模型可以联系上下文“理解”词义,为词语“订制”独一无二的语境向量表示,从而很好地解决同形词和一词多义问题。BERT 模型的学习主要涉及两个核心模块:编码器和目标任务,本节将分别对两者进行介绍。

在编码器的选择上,BERT 采用 12 层或 24 层 Transformer 模型进行特征学习。如图 2 所示,Transformer 模型的输入为字符向量、片段向量和位置向量之和。模型内每一层由两部分组成:多头自注意力(multi-head self attention)和全连接神经网络(fully connected neural networks),每个网络的输出均经过层归一化操作(layer normalization)。其中,多头自注意力网络中每个隐单元的输入均由上一层隐单元输出加权平均得到,使得每个隐单元均能和上一层所有隐单元直接关联,这样一来,每个隐单元都可以较好地编码全局语义信息。

在目标任务上,BERT 模型采用了完形填空和句子预测这两项任务。在完型填空任务中,15%的

单词会被选中,其中,80%被替换为[MASK],10%被替换为一个随机词,10%保持不变,模型需要据此预测被选中的词。在句子预测任务中,模型需要判断句子 A 和 B 是否相邻。通过两个目标任务,语言模型能够同时捕捉词语和句子级别的语言知识。

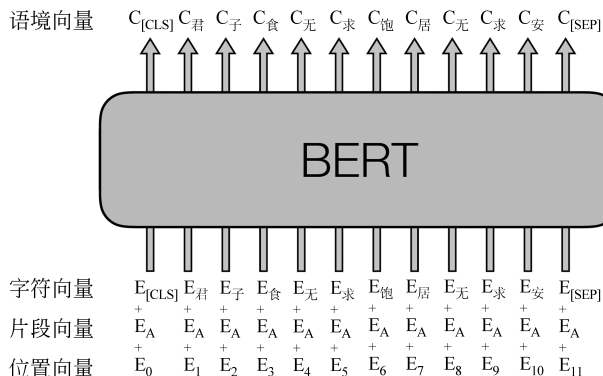


图2 古汉语知识表示模型的输入和输出示例

### 1.2.2 古汉语知识表示

在训练 BERT 模型时,Devlin 等人采用字级别中文维基百科语料库训练了中文语言模型,其编码以句子为单位。本文在此基础上引入海量古汉语语料库进行增量训练,考虑到古汉语句子长度较短,且大量训练数据不含断句和标点信息,本文将段落作为输入单位。

如图 2 所示,训练模型时,输入字符串  $S$ , Transformer 模型首先将其转换成字符序列,在开始和结束位置处添加[CLS]和[SEP]标签,并给出其位置和片段信息,将三者向量求和作为模型输入。输入向量经过预训练模型编码,在每个位置都可以得到对应的输出  $C_{token}$ ,每个输出均为一个 768 维的语境向量。其中,[CLS]对应位置的语境向量可视为编码了整个片段的语义信息,常作为下游文本分类任务的输入。

与 Word2Vec 训练产生的词向量相比,BERT 模型输出的语境向量能够编码细粒度的词义信息,表 1 以“安”为例,给出了两种模型的最近邻信息。计算 BERT 模型最近邻时,我们从《论语》中选取了四条“安”含义不同的语料,经预训练模型编码,获取了“安”在不同上下文中的语境向量,随后以《史记》语料为查找对象,找到与其最相近的语境向量表示。由表中内容可见,基于 Word2Vec 模型的最近邻词语聚焦在表示“安宁”“平安”“使安定”意义的古汉语词汇上,而 BERT 模型可以针对句中词语根据当前上下文给出语境向量表示,因而能够捕捉细粒度的词义信息。

表 1 “安”的最近邻示例

模型	原词	最近邻(cosine 相似度)
Word2Vec	安	宁、永、平、保、乐、镇、……
BERT	君子食无求饱,居无求安。 (安定、安稳)	① 高而不坏,地得为安。 ② 夫轻万乘之重不以为安,而乐出於万有一危之涂以为娱,臣窃为陛下不取也。 ③ 客寝甚安,殆非就国者也。
	修己以安百姓,尧舜其犹病诸。 (使……安定或稳定)	① 在今尔安百姓,何择非其人。 ② 所谓贤人者,必能安天下而治万民,今身且不能利,将恶能治天下哉! ③ 行无穷之欲,甘心快意,结怨於匈奴,非所以安边也。
	安见方六七十,如五六十,而非邦也者。 (表反问,哪里、怎么)	① 安得长者之语而称之! ② 由,譬使仁者而必信,安有伯夷、叔齐? ③ 安有说人主不能出其金玉锦绣,取卿相之尊者乎?
	子温而厉,威而不猛,恭而安。 (感到满足)	① 若关雎“乐而不淫,哀而不伤”,则是有庄敬而安者也。 ② 礼得其报则乐,乐得其反则安。 ③ 而安而色,曰予所好德,女则锡之福。

## 2 古汉语自动断句模型

预训练语言模型不仅可以实现高效的古汉语词义表示,还可通过微调(fine-tuning)机制接入下游任务,如文本分类、序列标注、语义推理等。在微调过程中,伴随下游任务的训练,整个语言模型的参数也随之迭代更新。

自动断句模型可以被视为一个典型的序列标注任务,即输入字符串,针对每个字符预测在该位置是否断句,例如,输入“君子食无求饱居无求安”,模型应预测“OOOOOSOOOO”,其中,“O”表示该位置后不应断句,“S”表示“饱”后应断句。Devlin 等人在 BERT 模型的基础上提出了基于全连接神经网络分类器的序列标注方法,并在 CoNLL-2003 命名实体识别任务上取得了最优效果,其模型结构如图 3(a)所示。这种序列标注方法虽然能够利用 BERT 模型输出的高效语义表示做标签预测,但存在收敛速度慢、未考虑标签之间的依赖关系等问题。为了改进序列标注方法,本文基于深层预训练语言模型引入条件随机场(CRF)和卷积神经网络(CNN)模型,实现了更为高效、准确的中文断句方法。

条件随机场是一种经典的序列标注模型,在中文分词、词性标注、命名实体识别等自然语言处理任务中均有着广泛应用<sup>[12]</sup>。伴随神经网络模型的兴起,Huang 等人<sup>[13]</sup>在双向长短期记忆网络(Bi-LSTM)上添加了 CRF 层,用于求解概率最优的标签路径,在一系列序列标注任务上取得了明显的效果提升。

受前人工作启发,本文将条件随机场模型接入

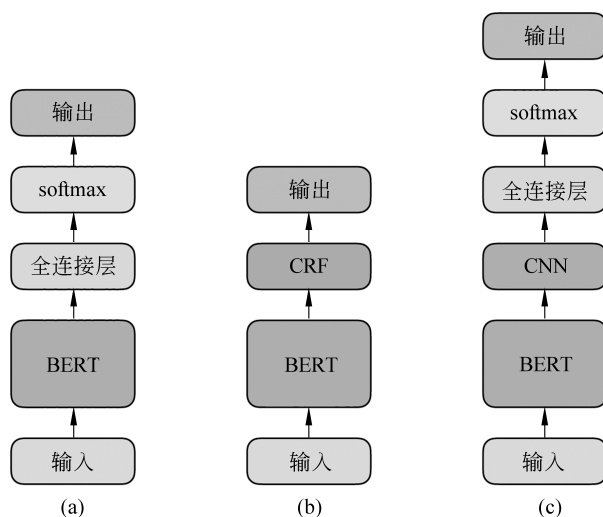


图 3 古汉语自动断句模型结构图

深层语言模型,其结构如图 3(b)所示。通过观察 BERT+CRF 模型的预测结果,我们发现由于 CRF 模型预测时仅能考虑当前位置及之前位置的特征,未能充分地利用上下文信息进行断句,造成了一些断句错误,例如:

**BERT+CRF:** 行未三四十里 ○ 忽乌刺赤者急下马拜跪 ● 伏其言 ● 侏离莫能晓 ○ 而其意则甚哀窘

该例中的句读重点有二:一是“拜”“跪”“伏”为连续的动作,二是“其言”与后文的“其意”呼应,均应作主语。

为了进一步提升模型对上下文语言特征的编码能力,本文在 BERT 模型基础上引入卷积神经网络做特征抽取,并基于其编码结果利用一层全连接神经网络实现断句标记分类,其结构如图 3(c)所示。由于 CNN 模型能够通过卷积对两侧上下文信息进



行编码,综合决策后作出了正确的断句决策,结果如下所示:

**BERT+CNN:** 行未三四十里 ○ 忽乌刺赤者急下马拜跪伏 ○ 其言侏离莫能晓 ○ 而其意则甚哀窘

### 3 实验及评估

#### 3.1 数据集

古汉语深层语言模型训练基于殆知阁古代文献藏书 2.0 版语料库<sup>①</sup>,共计 33 亿字,由于数据中繁简体字混合出现,考虑到繁转简准确率更高,在预处理阶段采用 zhconv 工具<sup>②</sup>将文本统一转成简体。

在自动断句任务中,我们从 Github 中华古诗词数据库<sup>③</sup>中获取了带标点的古诗词数据,其中诗 311 691 首,词 20 643 首,从殆知阁古典文献藏书 2.0 语料库中获取带标点的文言文语料 8 163 988 条(以段落为单位)。由于诗词具有较为明显的格律特征,如大部分古诗为四、五、七言,而词牌名可以提示断句规则,为了帮助模型更好地学习语义和韵律信息,在预处理数据时保留了古诗题目,并去除词牌名。针对数量较少的词数据,取 10% 作为测试集,针对数量较多的古诗和文言文数据,各取 5 000 条作为测试集,其余诗、词、文言文数据合为训练集,并从训练集中随机抽取 10 000 句作为验证集。

#### 3.2 模型及参数设置

古汉语 BERT 模型训练采用 12 层 Transformer 模型,hidden size 为 768,自注意力机制的 head 数量为 12,总参数量为 1.1 亿,采用 4 块 1 080Ti 型号的 GPU 并行训练 100 万步得到语言模型。

在断句模型上,本文将王博立等人<sup>[4]</sup>提出的双向 GRU 模型(Bi-GRU)作为基线(baseline)模型,实验中将 GRU 模型的 hidden size 设为 256,考虑到本文训练数据规模远大于文献[4]中的数据集,我们另增加了一组 hidden size 为 2 048 的实验。此外,将 Devlin 等人提出的 BERT+全连接层(fully connected layer)序列标注模型应用到断句任务中(简称为 BERT+FCL),并构建了 BERT+CRF 与 BERT+CNN 模型。其中,CRF 层采用 Tensorflow 默认设置,CNN 层使用了 100 个宽度为 3 的卷积核,用于抽取特征。所有模型均训练到验证集收敛为止。

#### 3.3 实验结果

五组模型的断句实验结果如表 2 所示。从测试数据类型的角度看,无论是双向 GRU 模型,还是融入深层预训练语言模型的方法,均呈现出古诗断句效果最优、词次之、古文再次之的特点,这与文体表达的规律性和韵律性密切相关,也折射了不同文体断句难度的差异。

表 2 断句模型实验结果

模型	诗			词			文		
	$P/\%$	$R/\%$	$F_1/\%$	$P/\%$	$R/\%$	$F_1/\%$	$P/\%$	$R/\%$	$F_1/\%$
Bi-GRU <sub>256</sub>	95.17	94.67	94.92	82.85	74.26	78.32	81.60	76.34	78.89
Bi-GRU <sub>2048</sub>	96.73	96.55	96.64	86.80	83.38	85.06	83.95	79.19	81.50
BERT+FCL	98.98	99.33	99.16	95.33	94.90	95.11	91.65	92.31	91.98
BERT+CRF	<b>99.10</b>	99.29	<b>99.19</b>	95.50	94.77	95.13	91.66	<b>92.41</b>	<b>92.03</b>
BERT+CNN	99.04	<b>99.35</b>	<b>99.19</b>	<b>95.77</b>	<b>95.45</b>	<b>95.61</b>	<b>91.77</b>	92.26	92.01

从模型表现的角度看,集成 BERT 深层预训练模型后,与基线模型相比,三种模型在三类文体上的断句效果都得到了巨幅提升。其中,古诗断句  $F_1$  值接近 100%。在语言表达较为灵活、多样的词和古文测试集上,综合表现最优的 BERT+CNN 模型,比之 Bi-GRU<sub>2048</sub> 模型提升幅度达到 10% 以上,词断句  $F_1$  值达到 95% 以上,古文断句  $F_1$  值达到 92% 以上。

此外,通过观察基线模型 Bi-GRU 的实验结果,

不难发现,其断句召回率( $R$ )大大低于精确率( $P$ ),即大量断句标记未被识别,这一特点在难度较高的文体(词、古文)上表现尤为突出。融入深层语言模型后,断句召回率与精确率基本持平,均达到了较高的水平。在集成预训练模型的三种方法中,BERT

① <http://www.daizhige.org/>

② <https://pypi.org/project/zhconv/>

③ <https://github.com/chinese-poetry/chinese-poetry>

+CRF 和 BERT+CNN 与 BERT+FCL 相比均有小幅提升。

通过分析测试数据(表 3),我们发现,由于深层语言模型可从海量数据中学习语言知识表示,在古汉语领域,其优势具体体现在以下两个方面:

第一,能够较好地捕捉古诗文表达的节奏感和韵律感,例如,表 3 第 1、2 句。其中,句 1 为五言诗,句 2 为长短句交错的词,该二例断句与节奏、韵律关联紧密,而 Bi-GRU 模型未能捕捉这种语言表达性质,因而出现了应断未断(“日后”“瑶台月”“心事”处)与不应断而断(“攀”处)错误,三种集成预训练语言模型的方法均能正确识别。

第二,对上下文信息的利用较为充分,如表 3 例句 3、4 所示。其中,句 3 需联系前文理解“行修”为夫君,其所娶妻子“贞懿贤淑”(主谓搭配)，“行修”对其十分尊敬(主谓搭配)。句 4 中,“齐人”“宋人”“邾娄人”为并举,模型需联系上下文在三者之间句读,此外,“伐郑”与“救郑”前后呼应,二者之后均应断句。“书者……”是对“伐郑一救郑”事件的点评,意为《春秋》记载这件事,是在称赞中原国家可以互救。

综上所述,深层语言模型所编码的古汉语知识在一定程度上涵盖了语序、语法、语义、语境等多层次的语言信息,对于后续的自然语言处理任务有重要的贡献。

表 3 模型断句示例

编号	Bi-GRU <sub>2048</sub>	BERT+FCL/CRF/CNN
1	应接有不暇 ○ 幽忧能顿除 ○ 明当增短日后且怆离居	应接有不暇 ○ 幽忧能顿除 ○ 明当增短日 ○ 后且怆离居
2	梅花发 ○ 寒梢挂著瑶台月瑶台月和羹心事履霜时节 ○ 野桥流水声呜咽 ○ 行人立马空愁绝 ○ 空愁绝 ○ 为谁凝伫 ○ 为谁攀 ● 折	梅花发 ○ 寒梢挂著瑶台月 ○ 瑶台月 ○ 和羹心事 ○ 履霜时节 ○ 野桥流水声呜咽 ○ 行人立马空愁绝 ○ 空愁绝 ○ 为谁凝伫 ○ 为谁攀折
3	李十一郎行修 ○ 初娶江西廉史王仲舒女贞懿贤淑 ○ 行修 ● 敬之如宾 ○ 王女有幼妹尝挈以自随 ○ 行修亦深所鞠爱	李十一郎行修 ○ 初娶江西廉史王仲舒女 ○ 贞懿贤淑 ○ 行修敬之如宾 ○ 王女有幼妹 ○ 尝挈以自随 ○ 行修亦深所鞠爱
4	秋 ○ 荆伐郑 ○ 公会齐人 ○ 宋人邾娄人救郑书者 ○ 善中国能相救	秋 ○ 荆伐郑 ○ 公会齐人 ○ 宋人 ○ 邾娄人救郑 ○ 书者 ○ 善中国能相救

#### 4 案例应用分析

为了进一步验证 BERT 模型在处理断句任务中的应用效果,我们根据司马朝军<sup>[14]</sup>、颜春峰和汪少华<sup>[15]</sup>等学者的研究,搜集了已出版古籍文本中 65 则与断句相关的错误案例,并排除了在训练集中出现过的 5 则语料,得到 60 则测试数据。其中,11 则来自中华书局 1997 年版《钦定四库全书总目》,49 则来自中华书局 1987 年版《周礼正义》。这两本古籍均由该领域专家完成整理和句读标点,并经多次校对,其中的误例可谓句读任务的难点所在。

《钦定四库全书总目》由李学勤作序,是今人重要的古籍整理成果。我们从司马朝军的研究中找出了 11 则与断句相关的标点错误,其分别在《春秋后传》《春秋献义》《数学九章》《姑溪词》等条中,覆盖了经部、子部、集部三类典型文献。我们将这 11 例去除标点后作为输入,由模型进行断句,其中,8 则模型完全断句正确,3 则断句不完全正确。试举正误例各一如下:

**例 1** 柏何人,斯敢奋笔而进退孔子哉? (《诗经》第 216 页)

作者按:“斯”字上属。“何人斯”为上古习语<sup>①</sup>。

当作:柏何人斯,敢奋笔而进退孔子哉?

模型:柏何人斯 ○ 敢奋笔而进退孔子哉 (模型断句正确)

**例 2** 其中如“大衍”类著卦发微,欲以新术改《周易》揲著之法,殊乖古义。古历会稽题数既误,且为设问,以明大衍之理。(《数学九章》第 1406 页)

作者按:此段标点有破句。

当作:其中如“大衍”类著卦发微,欲以新术改《周易》揲著之法,殊乖古义、古历。会稽题数既误,且为设问,以明大衍之理

模型:其中如大衍类著卦发微 ○ 欲以新术改 ○ 周易 ○ 揲著之法 ○ 殊乖古义 ○ 古历会稽题数既误 ○ 且为设问 ○ 以明大衍之理 (模型断句存在错误)

考虑到上古语言与中古语言的差异,为了验证

① 《诗经·小雅·何人斯》:彼何人斯? 其心孔艰。

断句模型在处理上古语言时的效果,我们又选择王文锦、陈玉霞点校的《周礼正义》一书,将颜春峰和汪少华整理的 49 则断句误例送入模型测试。其中,模型能正确断句 27 则,断句不完全正确的有 22 则。

《周礼正义》的模型断句误例中,较为集中的是对字义的考证,尤其是引《说文》时的错误,比如“服,牝服,车之材”误断作“服牝,服车之材”。“服”作为《说文》中的字头,其用法与其他古文表达有较大区别。此外,因盟誓、考课、葬礼等礼仪制度不明而致误亦有数例。

从经典古籍中的断句疑难案例可以看出,本文提出的自动断句方法在处理古籍一般句式表达时有明显优势。而在处理《说文》、古代制度等专业性较强的数据时尚存在问题,这与该类型数据相对较少有关。总的来说,本文方法在已出版古籍的断句疑难误例上取得了很好的效果,测试共计 60 例(均为专家标点错误,并经多次校对未查出),而模型能完全正确断句 35 例,达到了较为实用的水平。

## 5 总结与展望

古汉语信息处理技术在古籍整理和古代文献、文学研究中扮演着重要的角色。为了实现高效的古汉语知识表示,本文基于 33 亿字古汉语语料库学习深层语言模型,并在此基础上实现了高精度的断句模型,在诗、词和古文三种文体上,模型断句  $F_1$  值分别达到 99%、95% 和 92% 以上。通过分析实验数据,我们发现模型能较好地捕捉诗词表达的节奏感和韵律感,也能充分利用上下文信息,实现语序、语法、语义、语境等信息的编码。在进一步的案例应用中,本文方法在已出版古籍的断句疑难误例上也取得了较好的效果。

从应用角度看,本文提出的断句方法既可以用于大规模古籍整理中预断句工作,大大减轻专家负担,也可用于校对环节,帮助检测人工断句或标点的错误。在后续工作中,我们希望将基于深层语言模型的古汉语知识表示方法应用到古文翻译、古诗文创作等其他古汉语信息处理任务中去。

## 参考文献

[1] 朱熹. 韩文考异. 影印文渊阁四库全书(第 1073 册)

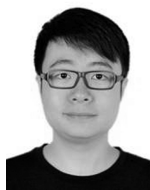
[M]. 中国台湾: 台湾商务印书馆, 1986: 566.

- [2] 黄侃. 黄侃手批白文十三经[M]. 上海: 上海古籍出版社, 1983: 5.
- [3] 张开旭, 夏云庆, 宇航. 基于条件随机场的古文自动断句与标点方法[J]. 清华大学学报(自然科学版)网络. 预览, 2009, 49(10): 163-166.
- [4] 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.
- [5] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations [C]//Proceedings of NAACL. New Orleans, USA: Association for Computational Linguistics, 2018: 2227-2237.
- [6] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of NAACL. Minneapolis, USA: Association for Computational Linguistics, 2019: 4171-4186.
- [7] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of NIPS. Lake Tahoe, USA: Neural Information Processing Systems Foundation, 2013: 3111-3119.
- [8] Zellig Harris. Distributional structure[J]. Word, 1954, 10(23): 146-162.
- [9] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation[C]//Proceedings of EMNLP. Doha, Qatar: Association for Computational Linguistics, 2014: 1532-1543.
- [10] Levy O, Goldberg Y. Neural word embedding as implicit matrix factorization[C]//Proceedings of NIPS. Montreal, Canada: Neural Information Processing Systems Foundation, 2014: 2177-2185.
- [11] Li S, Zhao Z, Hu R, et al. Analogical reasoning on Chinese morphological and semantic relations[C]//Proceedings of the ACL. Melbourne, Australia: Association for Computational Linguistics, 2018: 138-143.
- [12] Zheng X, Chen J, Shang G. Deep neural network-based Chinese semantic role labeling [J/OL]. ZTE Communications, 2018: 1-12. <http://kns.cnki.net/kcms/detail/34.1294.TN.20180102.1045.002.html>. [2018-01-02].
- [13] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging[J/OL]. arXiv preprint arXiv: 1508.01991, 2015.
- [14] 司马朝军. 中华书局《钦定四库全书总目》整理本校记[J]. 人文论丛, 2013, 1: 357-394.
- [15] 颜春峰, 汪少华. 从《周礼正义》点校本谈避免破句的方法[J]. 古汉语研究, 2014, 2: 47-55, 95.



胡韧奋(1988—), 博士, 讲师, 主要研究领域为计算语言学。

E-mail: irishu@mail.bnu.edu.cn



李绅(1992—), 硕士, 主要研究领域为自然语言处理。

E-mail: shen@mail.bnu.edu.cn



诸雨辰(1988—), 博士, 讲师, 主要研究领域为中国古典文献学。

E-mail: zhuyuchen@bnu.edu.cn

(上接第 7 页)

- [22] 胡峰松, 张璇. 基于梅尔频率倒谱系数与翻转梅尔频率倒谱系数的说话人识别方法[J]. 计算机应用, 2012, 32(09): 2542-2544.
- [23] 潘凌云, 孙达传, 吴美朝. 语音识别中基于语谱图的语音音素分割方法[J]. 杭州大学学报(自然科学版), 1995(01): 42-46.
- [24] Badshah A M, Ahmad J, Rahim N, et al. Speech emotion recognition from spectrograms with deep convolutional neural Network[C]//Proceedings of the International Conference on Platform Technology & Service. 2017.
- [25] 马义德, 袁敏, 齐春亮, 等. 基于 PCNN 的语谱图特征提取在说话人识别中的应用[J]. 计算机工程与应用, 2005(20): 81-84.
- [26] 中国社会科学院语言研究所. 863 语音识别语音语料库 RASC863—四大方言普通话语音库[A]. 中国中文信息学会, 2003: 4.
- [27] 殷志刚. 语音语料库的建设和作用[N]. 中国社会科学院院报, 2007-7-23(03).
- [28] 陈小莹, 陈晨, 华侃, 等. 语音语料库的设计研究[J]. 科技信息, 2008(36): 5-6.
- [29] 杨鸿武, 梁青青, 郭威彤, 等. 一个面向言语工程的兰州方言语料库[J]. 西北师范大学学报(自然科学版), 2009, 45(06): 54-59.
- [30] 邹法欣. 语音语料库的设计与实现[D]. 桂林: 广西师范大学硕士学位论文, 2012.
- [31] 高原, 顾明亮, 孙平, 等. 多用途汉语方言语音数据库的设计[J]. 计算机工程与应用, 2012, 48(05): 118-120.
- [32] 陈昌仪. 赣方言概要[M]. 南昌: 江西教育出版社, 1991.
- [33] Fan Xu, Jian Luo, Mingwen Wang, et al. Speech-driven end-to-end language discrimination towards Chinese dialects[J]. ACM Transactions on Asian and Low-Resource Language Information Processing, 2020, 19(5): 1-24.
- [34] Fan Xu, Mingwen Wang, Maoxi Li. Building parallel monolingual Gan Chinese dialects corpus[C]//Proceedings of the 11th Conference of the Language Resources and Evaluation Conference (LREC), 2018: 244-249.



颜为之(1985—), 博士研究生, 讲师, 主要研究领域为自然语言处理, 计算语言学。

E-mail: 124901714@qq.com



王明文(1964—), 通信作者, 博士, 教授, 博士生导师, 主要研究领域为自然语言处理、信息检索、数据挖掘、机器学习。

E-mail: mwwang@jxnu.edu.cn



徐凡(1979—) 博士, 副教授, 主要研究领域为自然语言处理。

E-mail: xufan@jxnu.edu.cn