doi:10.3772/j.issn.2095-915x.2021.05.003

基于 Transformer-CRF 的文言文断句方法研究——以唐代墓志铭为例

开放科学（资源服务）标识码（OSID）

韩旭

中国科学技术信息研究所　北京　100038

摘要：[目的 / 意义] 本文探索了文言文的断句规则，并以唐代墓志铭为例进行基于语义的句子边界识别，降低了文言文因缺少标点造成的阅读障碍，减少了人工标注标点的工作量，为中国古文的资料分析整理提供技术支撑。[方法 / 过程] 本文首先使用一种基于汉字偏旁的字表示方法，提取汉字本身隐含的语义信息进行表达。将基于偏旁的字表示输入 Transformer-CRF 模型，并对墓志铭中的缺失字进行了滑动窗口填补操作，降低缺失字对整体模型的影响。该模型在提高并行计算效率的基础上对输出结果进行关联，提高了准确率。[结果 / 结论] 实验表明，使用基于偏旁的字向量表示方式结合 Transformer-CRF 能提升唐代墓志铭的断句准确率，且对于缺失字附近的断句准确率有所提升，该方法对数字人文中信息收集和整理工作起到了一定的辅助支撑作用。

关键词：Transformer-CRF；繁体字向量；句子边界识别；古籍信息处理

中图分类号：TP391；G35

Research on Sentence Segmentation of Classical Chinese Based on Transformer-CRF: Taking Epitaph of Tang Dynasty as An Example

HAN Xu

Institute of Scientific and Technical Information of China, Beijing 100038, China

Abstract: [Objective/ Significance] In this paper, we explore the rules of sentence segmentation in classical Chinese, and take the

基金项目：中国科学技术信息研究所创新研究基金面上项目（MS2021-04），中国科学技术信息研究所重点工作（ZD2021-09）。
作者简介　韩旭（1991-），博士，助理研究员，主要研究方向为自然语言处理，E-mail：hanx@istic.ac.cn。

引用格式　韩旭 . 基于 Transformer-CRF 的文言文断句方法研究 : 以唐代墓志铭为例 [J]. 情报工程 , 2021, 7(5): 30-39.

epitaph of Tang Dynasty as an example to identify the sentence boundary based on semantics. This method reduces the reading obstacles caused by the lack of punctuation in classical Chinese, reduces the workload of manual punctuation, and provides technical support for the collation and analysis of ancient Chinese information. [Methods/Process] Firstly, this study uses a character representation method based on Chinese character radicals to extract the implied semantic information of Chinese characters. The word representation based on radical is input into Transformer-CRF model, which improves the efficiency of parallel computing and correlates the output results to improve the accuracy. In addition, the missing words in the epitaph are filled by the sliding window to reduce the impact of missing words on the overall model. [Results/Conclusions] The experimental results show that character representation based on radicals combined with Transformer-CRF model can improve the accuracy of sentence segmentation of Tang Dynasty Epitaphs, and improve the ability of sentence segmentation near missing characters. This method plays a supporting role in information collection and collation in Digital Humanities.

Key words: Transformer-CRF; traditional Chinese character vector; sentence boundary recognition; Ancient Chinese information processing

引言

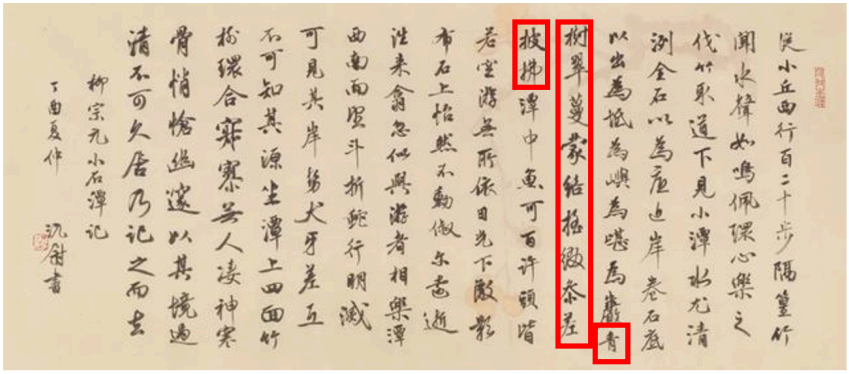
汉字作为世界上最古老的文字之一，距今已有六千多年的历史，是承载中华传统文化的重要工具，而文言文则是中国古代使用汉字书写的一种书面语言。对文言文中所述内容的深入研究，能够更为详尽地解读历史，建立对应的中国古籍索引，为中国古代历史研究提供更多的事实依据。

目前基于中文的预训练模型大多针对现代汉语，但文言文和白话文在书写方式、表达方式等方面都存在较大差异，难以使用相同语言模型对其进行深入解读。文言文和现代白话文的区别主要表现在以下几点：（1）相对于白话文而言，文言文其最突出的特色是行文简练、言文分离。古代文言文中每个汉字代表的语义都较为丰富，语言表述高度凝练，难以进行分词操作；而白话文则更接近日常交流的口语，讲求表意清晰，在文本理解和语义分割上都更加简单。（2）从书写格式而言，中国古代人在写作时，一般是从上至下从右至左纵向书写，且不使用标点符号、空格、分段等方式进行文字分割，主要靠知识积累及语感进行断句；而白话文则需在有语义分割的地方进行标点的标注，做到文字通俗易懂。（3）从文字的书写形式而言，中国古代大多使用繁体字进行书写，因此留存的文言文古籍均为繁体字表述；而白话文一般使用汉字文字改革之后的简体字进行记录。相对于繁体字而言，简体汉字从笔画和汉字结构上都有较大的改变，将文字变得简单易学，但也因此损失了一些笔画中的语义信息。文言文和白话文之间的具体差异如图 1 所示。

墓志铭是一种悼念性文体，是指存放在墓中载有死者传记的石刻。对古代墓志铭的深入分析能够刻画过世之人的生平，对建立古代人物数据库并对人物之间亲属关联的分析具有较为突出的作用，对建立中文古籍索引提供重要

doi:10.3772/j.issn.2095-915x.2021.05.003

的事实依据，因此墓志铭语料在古籍知识组织领域是一种较为珍贵的史料数据。除了具备文言文基本特征外，古代墓志铭还有另一个突出特点，即墓志铭数据大多来源于墓碑的摘抄或拓印，墓碑由于材质及存放位置等原因，经年累月风雨侵蚀，部分文字已经无法辨认，这些无法辨认的部分也会为研究人员带来一定的读译困难。



文言文	青树翠蔓	蒙络摇缀	参差披拂
白话文	青葱的树木翠绿的藤蔓，	遮掩缠绕摇动下垂，	参差不齐随风飘动。

图 1 文言文和白话文从书写、表达等方面的差异（唐·柳宗元《小石潭记》）

对文言文的分析首先要进行必要的语义分割，对文言文进行句子边界识别有助于后续的深入分析。针对文言文的文字特点，本文提出一种基于繁体字向量的 Transformer-CRF 模型，为唐代墓志铭文本进行断句。本文的主要贡献有以下几点：1. 使用了繁体字的字形特点生成特有的汉语繁体字向量，该向量在古代汉语中有较好的表现；2. 将墓志铭数据集中缺失的文字进行滑动窗口的数据填补，解决墓志铭数据缺失的问题；3. 主要针对唐代墓志铭进行数据标注，并同时生成唐代的文言文断句模型，帮助人们克服阅读障碍，并为后续有关古籍的知识组织工作提供有力的技术支撑。

1 相关工作

识别句子边界是一个重要的自然语言处理任务，目前该类任务普遍应用于语音转换文字的场景中。现代汉语在写作的过程中已经形成了标点标注的规则，因此无需进行句子边界识别步骤。根据 Li 等 [1] 的研究及在实践过程中的经验显示，标点符号的存在对文本分割和语义理解有着重要的作用，因此对于原始的文言文数据，应首先对其进行句子边界识别和标点标注。

目前，已有相关学者聚焦到文言文语境下的句子边界识别问题。早在 2007 年，陈天莹等 [2] 就提出古文句子边界识别的重要性，并使用 N-Gram 方法，以《论语》为数据集进行句子边界识别的训练。2008 年，黄建年等 [3] 将文言文的关注范围缩小到农业这一领域，提出可以使用正则表达式等基于统计与规则的方法对古籍进行断句位置的标注。张合等 [4] 使用了条件随机场解决文言文句子边界识

别问题，并取得了优于 N-gram 模型的效果。Huang 等 [5] 同样使用条件随机场作为初始模型，并加入古汉语的音节音韵信息以达到提高效果的目的。上述文章均使用传统的统计学习方法对文言文进行断句操作，并增加了一定的人工干预，这类工作需要有文言文阅读能力的专业研究者进行人工标注，难以实现大规模标注集。

2015 年，百度团队 [6] 提出了双向长短时记忆神经网络和条件随机场模型（Bi-LSTM-CRF）来解决自然语言处理中的序列标注问题，该算法在当时迅速成为最主流的解决方案。高甦等 [7] 使用深度学习对古汉语中的中医典籍领域进行了命名实体识别研究，取得了较好的成果。丁龙等 [8] 提出使用 Bert 模型对特殊的领域做实体识别，这也为文言文领域提出了新的研究思路。2016 年，Wang 等 [9] 提出了使用循环神经网络建立古汉语断句模型的方法，将深度学习方法引入古汉语断句问题中。王博立等 [10] 提出使用 GRU 模型对文言文进行断句，并引入了句长惩罚，以提高断句准确率。俞敬松等 [11] 提出使用 Bert 方法来进行文言文断句并针对具体问题进行微调适配，达到较好的效果。另外，还有学者考虑将文言文看作翻译问题进行模型构建 [12]，但由于文言文的复杂性，翻译效果还存在一定改进空间。上述研究均基于神经网络进行模型训练，没有深入考虑汉字本身可能带来的一些语义信息。

在自然语言处理的研究领域，目前较为通用的方法是使用词表示作为模型的输入。部分学者在基本词表示的基础上，对文字本身在演化过程中存在的一些语义信息进行了深入剖析。在英文语境中，Hovy 等 [13] 引入了卷积神经网络，对英文进行字母级别的划分，学习了英文中词根词缀的语义内涵，有效提高了英文词性标注、命名实体识别等任务的准确率。在中文语境下，由于单个汉字即是计算机接收的最小单元，因此难以使用卷积神经网络对其进行进一步拆分，韩旭等 [14] 提出将繁体字的偏旁作为汉字的一个特征进行训练，以此来增强繁体字的语义表示能力。这些基于文字本身的研究能在一定程度上提升后续任务的准确率。

目前大部分使用神经网络模型的文言文断句工作都聚焦于全朝代分析，较少针对某一历史时期或某种文体建立对应的语料分析模型。而针对唐代（公元 618 年 -907 年）的文言文研究，尤其是对特殊文体例如墓志铭的研究更为稀少。本文针对以上的相关研究工作，建立针对唐代墓志铭文体的一种基于 Transformer-CRF 的断句模型，并引入繁体字向量，加强模型的语义理解能力，提高模型的准确率。本文工作为数字人文领域古代人物数据库的建立提供了数据基础，为古籍知识组织的后续分析工作提供了技术支撑。

2 模型方法

2.1 基于偏旁的字向量表示

文言文表述言简意赅，通常每个字都能代表白话文中一个词语甚至多个词语的含义，且目前没有公认权威的能够代表文言文分词规则的相关词典，因此本文不考虑将文言文进行分词，在模型生成中选择以字为单元进行向量表

doi:10.3772/j.issn.2095-915x.2021.05.003

示。本文的字向量参考了本团队之前研究成果[14]中的方法，将汉字繁体字进行基于繁体字偏旁的字向量表示。汉字中的偏旁和英文中的词根词缀类似，不需要上下文作参考，偏旁本身即可一定程度上表示字的语义信息。如偏旁三点水“氵”表示和水有关的事物，“江”、“河”、“湖”、“海”都以“氵”作为偏旁，其代表的含义均和“水”有关。这种表述方式主要是因为汉字的起源是象形文字，由于繁体字形态更接近于甲骨文，保留了象形文字的某些特征，这种现象在繁体字上表现更为明显。因此本文首先将全部的文字转换为繁体字，再进行后续操作。

汉字本身即是计算机接收信息的最小单元，目前没有符号能系统的表示汉字的偏旁信息。本团队的前序研究中，主要根据汉字的 Unicode 编码进行偏旁分类。分析汉字的 Unicode 编码可知，相同偏旁的 Unicode 编码相邻，因此借助 Unicode 编码加人工校对的方式可以将每个汉字的偏旁信息进行标注。目前根据新华字典的检索列表，汉字共 214 个偏旁。将偏旁信息作为字向量初始值的一部分，使用 CBOW 连续词袋法共同训练，可以得到基于偏旁信息的繁体字向量。其具体的结构如图 2 所示。

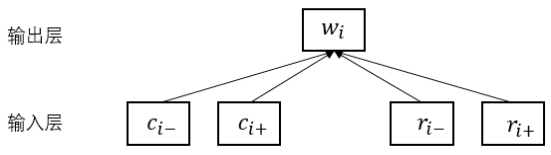


图 2 基于偏旁的繁体字向量的输入及输出

其中 c_i 表示字向量，r_i 表示偏旁向量。c- 表示目标字符前 n 个字符，c+ 表示目标字符后 n 个字符，r- 表示目标字符前 n 个字符的偏旁信息，r+ 表示目标字符后 n 个字符的偏旁信息。

计算的损失函数是分别求字符和偏旁两部分的上下文似然函数，条件概率使用 softmax 函数，如公式（1）、（2）。

$$L(w_i)=\sum_{k=1}^{2}\log P(w_i\mid h_{ik}) \quad (1)$$

$$P(w_i\mid h_{ik})=\frac{\exp(h_{ik}^{n}\hat{v}_{wi})}{\sum_{j=1}^{N}\exp(h_{ik}^{n}\hat{v}_{wi})},k=1,2 \quad (2)$$

这里 i 表示文本中第 i 个字符，h_1 和 h_2 分别代表字符上下文和偏旁上下文，h_{i1} 和 h_{i2} 分别代表上下文输入向量的加权平均值，$\hat{v}_{wi}$表示汉字 w_i 的输出向量。使用该模型输出的字向量作为字向量预训练模型，参与后文的序列标注训练。该模型的优势是在基础的字向量基础上，获取到汉字中相同的偏旁信息，增强字表示的语义性。

2.2 Transformer模型

早在 2014 年 Bengio 等[15]就提出在翻译模型中引入注意力机制，其核心的内容是在一个基本的编码器—解码器模型中，加入一个隐藏层的加权和，用来表示在当前状态时刻每个词对状态的影响概率。Google 团队在 2017 年发表了著名论文中[16]提到了自注意力机制。这种方式脱离了编码器—解码器的限制，通过文本本身的注意力训练，得到句子内部词语之间的依存关系，这也是 Transformer 模型的核心思想。和 RNN 模型相比，Transformer 模型解决了传统序列标注模型中无法并行计算的问题，极大地提高了运算速度，能够在维持模型效果的同时提高模型运行效率。

Transformer 的编码器部分，每一层都包含多头注意力机制及一个全链接前馈神经网络。并且将两个部分添加了残差连接及归一化操作，编码器模型的基本结构如图 3 所示，所有的编码器在结构上相同，但不共享参数以便学习更多的特征。

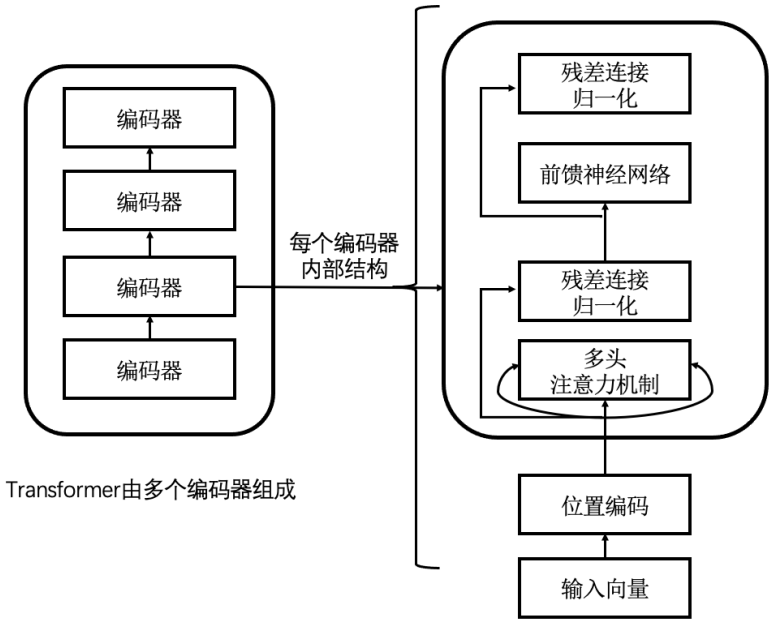


图 **3** 标准 **Transformer** 编码器结构

将前文生成的字向量作为 Transformer 的输入向量，模型首先需进行位置编码（positional encoding）。Transformer 模型抛弃了传统 RNN 模型的时间序列计算方式，因此需要对上下文的位置进行学习以保证模型的位置感知能力。位置编码计算方式如公式（3）、（4）。

$$PE(pos,2i)=\sin\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (3)$$

$$PE(pos,2i+1)=\cos\left(\frac{pos}{10000^{\frac{2i}{d}}}\right) \quad (4)$$

其中 pos 代表单词在句子中的位置，d 表示维度，i 表示字符位置，奇数维度用 cos，偶数维度用 sin 计算。字向量和位置向量相加，共同作为输入向量。

经过位置编码后的字向量进入到编码器进行多头注意力操作。多头注意力使用 concat 函数将不同的头连接起来形成高维度向量，获取不同层次的语义表示信息。注意力机制首先会将输入向量对应生成查询向量 Q、键向量 K 和值向量 V。并将 Q 和 K 与 V 的键值对映射到输出上。主要的计算公式如公式（5）（6）（7），上述计算均以矩阵形式完成。

$$f(Q,K_i)=Q^TK_i \quad (5)$$

$$a_i=soft\max(f(Q,K_i))=\frac{\exp(f(Q,K_i))}{\sum_i\exp(f(Q,K_i))} \quad (6)$$

$$Attention(Q,K,V)=\sum_i a_iV_i \quad (7)$$

Transformer 模型使用自注意力机制将输入传递到前馈神经网络中，并进行对应的残差连接和归一化操作。通过 Transformer 模型能够将输入序列进行文本特征提取，多头注意力能建

doi:10.3772/j.issn.2095-915x.2021.05.003

立不同的子空间并扩大学习参数，提升模型的准确度。且模型相比于传统 RNN 而言，由于使用位置编码信息，能够进行并行计算，极大地提升了模型效率。

2.3 CRF模型

Transformer 模型在输出层部分，序列之间没有相关关联，导致输出的结果可能会存在不合逻辑等问题，而 CRF 模型核心思想则是寻找一条最优路径，能够使得输出层的上下文标注存在约束性的规则连接，得到非独立的最优化标签。将 Transformer 的最顶层与 CRF 模型链接，能够保证输出标签的合理性及准确率。CRF 模型的目标是找到一个最大得分路径并输出。其路径分数的计算公式如（8）。

$$s(x,y)=\sum_{i=0}^{n}A_{yi,yi+1}+\sum_{i=0}^{n}p_{i,yi} \quad (8)$$

其中 A 是转移矩阵，$p_{i,yi}$ 表示第 i 个词语被标记为 y_i 的概率。将 Transformer 的结果输入 CRF 层，相当于两个分类器的效果相叠加，能够得到更优的模型效果。

2.4 文字标注方法

对于断句这一特殊的序列标注问题，本研究借鉴了命名实体识别的标注形式，将句子的首尾进行对应的标记。将句子的第一个字标记为 B（beginning），标点前的一个字标记为 E（end），其余字标记为 O（otherwise）。在标点选择上，仅选择带有语义停顿含义的标点进行边界识别：全停顿标点｛。？！｝，半停顿标点｛，；：｝共 6 种标点组成集合，在训练过程中 6 类标点不做具体区分，只标注断句位置。具体的标注样例如图 4 所示。

B	O	O	E	B	O	O	E	B	O	O	E
青	树	翠	蔓	蒙	络	摇	缀	参	差	批	拂

图 4 标注方式

2.5 整体模型框架

本文的整体模型框架如图 5 所示。

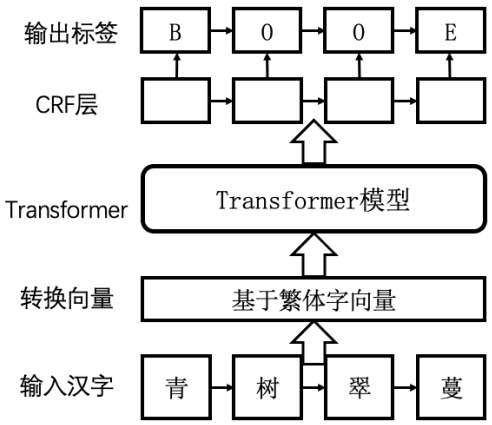


图 5 整体模型框架

首先使用本团队前序研究的字向量生成方法 [14]，将古文进行基于偏旁的繁体字向量训练。将训练好的基于偏旁的繁体字向量和位置向量进行连接，作为 Transformer 模型的输入向量进行训练。最后将 Transformer 的结果输入 CRF 层，并由 CRF 层训练最终的标签结果。

3 实验及结果分析

3.1 数据集

文言文作为中国古代通用的一种文字表述方式，其使用时间长达几千年，因此不同朝代不同文体之间，文言文的写作风格和形式都有所不同。将文言文内容进行朝代和文体的分类，更有利于断句的最终效果。墓志铭作为一类特

殊文体，目前还存在大量的未被电子化及标点化的数据，因此本文重点针对唐代墓志铭进行自动化句子边界识别，以减轻古文知识组织相关工作者的标注负担。

由于墓志铭的数据较难获取，且数据量不大，仅使用墓志铭数据进行深度学习建模，无法达到预期效果，本文使用同朝代相类似的文本类型进行数据集的扩充。本文主要使用的语料是唐代墓志汇编，因此扩展数据集的朝代同样选取唐代。在文体方面，通过对不同文体的调研及统计发现，人物传记类文体的写作风格和墓志较为相似，因此本文主要选取唐代人物传记类文本作为扩充数据集。本文所用的数据集具体情况如表 1 所示。

表 1 数据集数据细节

朝代	总体字符数	字表大小
唐·传记	6 160 233	7478
唐·墓志铭	1 116 850	5197

由于模型的运行效率和文本段落的长短强相关，因此对数据集进行分段切割，选取 60 字符为一个单元。为保证分割后每个单元内部的句子完整性，第 n 个单元和第 n+1 个单元选择重合 10 个字符，即第 n 个单元的最后 10 个字，是第 n+1 个单元的前 10 个字，并只取每个单元前 50 个字进行结果输出，这种选取方法可以尽可能避免单元分割带来的句子不完整问题。分好单元的文本需打乱顺序进行训练。本文按照训练集、验证集、测试集比例分别为 60%、20%，20% 来进行数据分配，其中训练集和验证集数据是唐传记数据和唐墓志铭数据混合而成，测试数据全部为墓志铭数据。

3.2 缺失字的填补方法

首先对墓志铭文本中的缺失文字进行调研。经过初步统计，墓志铭数据集中单个文字缺失的概率约在 1.7%，本文尝试将缺失字进行字填补。在字向量模型生成后，使用滑动窗口方法，将缺失字的前 n 个字和后 n 个字的字向量做加权平均，作为缺失字的字向量输入后续模型（实验部分最终选取 n=3）。本文在生成繁体字向量后，即进行缺失字填补操作。并将填补后的字向量和位置向量连接，作为 Transformer 模型的输入向量。另外，若缺失字连续出现 5 个以上，认为关键信息缺失过多，在后续断句任务中不予考虑。

在实验中，同时也考虑了仅使用统一的特殊字符对空缺文字进行填补，该方法和滑动窗口法的效果相近，但在缺失字位置附近的断句效果略差于滑动窗口法。使用统一字符方法将所有缺失字作为相同字标注，这种方法在连续缺失字情况下无法依靠上下文进行断句，因此效果较差。后续实验统一使用滑动窗口进行缺失文字的填补。

3.3 实验结果

本文模型的学习率设置为 0.005，迭代次数设置为 30 次，dropout 值为 0.4。评估函数采用基础的精确率、召回率、和 F1 进行评估，评估标准为是否在正确的位置进行句子分割。在实验过程中，使用 word2vec 和 LSTM 做基线模型，加入几种改进的模型进行对比，最终对比结果如表 2 所示。

$$P=\frac{\text{正确标注的断句位置数}}{\text{识别的所有断句位置总数}}\times 100\% \quad (9)$$

doi:10.3772/j.issn.2095-915x.2021.05.003

$$R=\frac{\text{正确标注的断句位置数}}{\text{所有断句位置总数}}\times 100\% \quad (10)$$

$$F1=\frac{2\times P\times R}{P+R}\times 100\% \quad (11)$$

表 2　唐代墓志铭数据评估结果

唐代墓志铭	精确率%	召回率%	F1-Measure%
Word2vec+LSTM	76.52	76.34	76.43
繁体字向量+LSTM	77.76	78.08	77.92
繁体字向量+Bi-LSTM-CRF	80.93	81.24	81.08
繁体字向量+Transformer	84.71	84.87	84.79
繁体字向量+Transformer-CRF	**85.10**	**85.43**	**85.26**

根据实验数据分析可知：首先，对比 word-2vec 和繁体字向量，可以看出在文言文语境下，繁体字向量的模型效果要好于 word2vec，这也证明了本团队前序工作的有效性。其次，同样使用繁体字向量的情况下，Bi-LSTM-CRF 模型的效果和 Transformer 模型从精准率、召回率和 F1 值上均存在较大差距，证明目前 Transformer 在序列标注任务上的表现能力已经远远优于改进后的 LSTM 模型。最后可以看出，本文在 Transformer 模型的基础上引入 CRF 之后的效果略优于单一 Transformer 模型，分析这一现象的原因在于本文在句子标注时，同时标注了句子的首字和尾字，因此序列标注本身存在一定的逻辑，即 E 和 B 应按顺序交替出现，而 CRF 的优势既是对序列的输出有一定的学习策略，因此在本文的应用场景下，Transformer-CRF 模型在三个指标下均表现出了最佳效果。另外，对于墓志铭数据中的缺失字而言，本文也能够进行对应的标注，证明了模型具备一定的泛化能力。选取一个带有缺失字的示例段落进行展示，具体的结果如下。

原文：

父□，齐殿内将军、开府、祭酒、冀州别驾。若乃将军逸气□□□而成□别□□才蹑□沂而化咏。跃马之荣未极，幼驷之贶攸归。君□蕴白蛇，出白荆峰之下；光飞紫□，腾於渥水之□。□□德□成珍，已超双□；不称力而为贵，自轻千里。

断句结果：（以 / 表示断句位置）

*父□ / 齐殿内将军 / 开府祭酒 / 冀州别驾 / **若乃将军逸气□ / □□而成□ / 别□□才蹑 / □沂而化咏** / 跃马之荣未极 / 幼驷之贶攸归 / 君□蕴白蛇 / 出白荆峰之下 / 光飞紫□ / 腾於渥水之□ / □□德□成珍 / 已超双□ / 不称力而为贵 / 自轻千里 /*

本段来源于《唐代墓志铭汇编续集》—垂拱。从断句结果可以看出，本模型能够做到基本的语义分割，在全停顿部分均预测正确，在半停顿部分更倾向于保留句子的完整性，断句位置均能给出合理解释。本文选取的段落缺失字较多且存在连续缺失，本模型在该种情况下仍然能进行断句切割，且能在连续缺失字之间进行语义分割预测。分析存在该现象的原因是，在缺失字填补过程中，使用已生成的繁字体向量进行滑动窗口缺失字填补，对缺失字部分的语义有一定的填充作用。另外，在缺失字连续且较多的情况下，除语义信息外模型还捕捉到了文本的结构信息进行断句。例如模型断句6,7,8三个位置（加粗表示），倾向于分为 5 个字一组的结构，且第三个字为“而”，这种断句位置的标注能对语义信息进行合理猜测提供了一定的辅助作用。该模型对于分析古籍知识发现尤其是带有缺失字的古籍文献具有重大意义，这为文言文整理工作提供新的思路，这也是本

模型区别于其他相似模型的一个显著特点。

4 总结

本文从古籍知识组织的实际应用需求出发，通过构建繁体字向量表示及序列标注模型来解决文言文标点缺失的问题。针对文言文的语言特点，以及墓志铭这种特殊文体存在的问题，引入了 Transformer-CRF 模型，针对唐代文体进行句子边界识别，并使用本团队前序工作中的基于偏旁的繁体字向量作为字表示。在实验中，以 word2vec 字表示以及 LSTM 模型为基线模型作对比，本文所用模型在三个主要评价参数上均获得了更优的表现，提高了文言文唐代墓志铭断句的准确率。在数据处理中通过滑动窗口加权平均的方法，对墓志铭中的缺失字进行填补，并能够在连续缺失字场景下对内部进行断句位置预测，为后续的分析工作提供了技术支撑。

目前对于墓志铭文体的相关电子资料和其他文体资料相比体量较小，数据量明显不足，因此在模型的表现上存在一定差距。但本文重点解决了缺失字环境下的断句问题，具有一定的研究意义。下一步的工作内容将对具体的标点进行预测，并进一步对古汉语中的实体构建知识库并识别。

参 考 文 献

[1] Li Z, Sun M. Punctuation as implicit annotations for Chinese word segmentation[J]. Computational Linguistics, 2009, 35(4):505–512.

[2] 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+196.

[3] 黄建年 , 侯汉清 . 农业古籍断句标点模式研究 . 中文信息学报 , 2008, 22(4):3 1—38

[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] Huang H H, Sun C T, Chen H H. Classical Chinese sentence segmentation[C]. Proceedings of CIPS-SIGHAN Joint Conference on Chinese Language Processing, Beijing. 2010:1-8.

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

[7] 高甦 , 金佩 , 张德政 . 基于深度学习的中医典籍命名实体识别研究 [J]. 情报工程 ,2019,5(1):113-123

[8] 丁龙 , 文雯 , 林强 . 基于预训练 BERT 字嵌入模型的领域实体识别 [J]. 情报工程 , 2019, 5(6):65-74.

[9] Wang B, Shi X, Tan Z, et al. A Sentence Segmentation Method for Ancient Chinese Texts Based on NNLM[C]. The Workshop on Chinese Lexical Semantics. Springer International Publishing. 2016:387-396.

[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(2):255-261.

[11] 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.

[12] 王雅松 , 刘明童 , 马彬彬 , 等 . 基于多翻译引擎的汉语复述平行语料构建方法 [J]. 情报工程 , 2020, 6(5):27-40。

[13] Ma X, Hovy E. End-to-end sequence labeling via bi-directional LSTM-CNNS-CRF[J]. arXiv preprint arXiv:1603.01354, 2016.

[14] 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):5-12.

[15] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv:1409.0473.

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