计算机科学
COMPUTER SCIENCE

结合门控机制的卷积网络实体缺失检测方法

叶瀚，李欣，孙海春

引用本文

叶瀚，李欣，孙海春.结合门控机制的卷积网络实体缺失检测方法[J]．计算机科学, 2023, 50(5): 262-269.

YE Han, LI Xin, SUN Haichun. Convolutional Network Entity Missing Detection Method Combined with Gated Mechanism [J]. Computer Science, 2023, 50(5): 262-269.

相似文章推荐（请使用火狐或 IE 浏览器查看文章）

Similar articles recommended (Please use Firefox or IE to view the article)

基于卷积神经网络多源融合的网络安全态势感知模型

Multi-source Fusion Network Security Situation Awareness Model Based on Convolutional Neural Network

计算机科学, 2023, 50(5): 382-389. https://doi.org/10.11896/jsjkx.220400134

基于多模态生成对抗网络的多元时序数据异常检测

Multimodal Generative Adversarial Networks Based Multivariate Time Series Anomaly Detection

计算机科学, 2023, 50(5): 355-362. https://doi.org/10.11896/jsjkx.220400221

基于多级多尺度特征提取的CNN-BiLSTM模型的中文情感分析

Chinese Sentiment Analysis Based on CNN-BiLSTM Model of Multi-level and Multi-scale Feature Extraction

计算机科学, 2023, 50(5): 248-254. https://doi.org/10.11896/jsjkx.220400069

基于多事件语义增强的情感分析

Sentiment Analysis Based on Multi-event Semantic Enhancement

计算机科学, 2023, 50(5): 238-247. https://doi.org/10.11896/jsjkx.220400256

伪异常选择驱动学习的视频异常检测

Pseudo-abnormal Sample Selection for Video Anomaly Detection

计算机科学, 2023, 50(5): 146-154. https://doi.org/10.11896/jsjkx.220400227

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400126

结合门控机制的卷积网络实体缺失检测方法

叶 瀚 李 欣 孙海春
中国人民公安大学信息网络安全学院 北京 102623
(yehan@ppsuc.edu.cn)

摘 要 实体信息充足与否直接影响着有赖于文本实体信息的相关应用,而常规的实体识别模型仅能对已存在的实体进行识别。文中提出以序列标注任务定义实体缺失检测任务,并提出了相应的3种实体缺失检测模型的训练数据构造方法。根据实体缺失任务的识别特点,提出了融合门控机制的卷积神经网络与预训练语言模型相结合的实体缺失检测方法。通过实验发现,基于预训练语言模型与门控卷积网络的模型对人名类、组织类、地点类实体缺失识别的F1最高分别达80.45%,83.02%和86.75%,显著高于基于LSTM的实体识别模型。通过字频统计发现,运用不同标注方法的数据集所训练的模型的准确率与被标注字符字频存在相关性。

关键词:门控机制;异常检测;预训练语言模型;卷积神经网络
中图法分类号 TP391

Convolutional Network Entity Missing Detection Method Combined with Gated Mechanism

YE Han,LI Xin and SUN Haichun
School of Information and Cyber Security,People's Public Security University of China,Beijing 102623,China

Abstract The adequacy of the entity information directly affects the applications that depend on textual entity information,while conventional entity recognition models can only identify the existing entities. The task of the entity missing detection,defined as a sequence labeling task,aims at finding the location where the entity is missing. In order to construct training dataset,three corresponding methods are proposed. We introduce an entity missing detection method combining the convolutional neural network with the gated mechanism and the pre-trained language model. Experiments show that the F1 scores of this model are 80.45% for the PER entity,83.02% for the ORG entity,and 86.75% for the LOC entity. The model performance exceeds the other LSTM-based named entity recognition model. It is found that there is a correlation between the accuracy of the model and the word frequency of the annotated characters.

Keywords Gated mechanism,Abnormal detection,Pre-trained language model,Convolutional neural network

1 引言

文本中的实体信息是文本数据中的重要内容,如果文本缺乏有效的实体信息,则对后续的文本信息分析与处理有着较大影响。缺失实体信息的数据输入,难以获取高质量的分析结果。

本文中的实体信息与命名实体识别(Named Entity Recognition,NER)任务所指的实体含义一致。实体缺失检测任务需具体指出文本中缺失实体的位置,这是对文本内容的检测分析。

对文本内容的检验分析模型在商业、教育领域已有较为广泛的运用,具体应用任务有商品评价分析、文本评分分析等。Fan等[1]使用深度学习方法构建了端到端的商品评价"有用性"评估模型,对Amazon和Yelp网站上的真实评论数据进行评估。Yang等[2]从商品评价的"全面性"对文本评价内容进行评论质量评估。Alikaniotis等[3]基于长短时记忆网络和词级别的评分向量构建了自动化的文本评分模型。Tay等[4]利用组合神经网络输出机制,较好地解决了长短时记忆网络无法捕捉到长序列中的长距离信息的难题,在论文评分测试数据集ASAP上超越了同类型的深度学习模型。

这些文本评估检验方法集中在对文本的评分,缺乏类似于实体缺失检测的细粒度、字词级的具体评估,较为相似的研究问题是文本语法检查任务。Sun等[5]基于随机条件场和依存句法树对文本序列中出现的语法错误进行标注。Hao等[6]使用Transformer模型构造了语法错误标注模型。上述的评分与检查模型聚焦于文本的语法形式内容,对于文本中的语义信息检查能力较弱,而实体缺失检测任务侧重于语义内容的缺失和错误。

到稿日期:2022-04-12 返修日期:2022-09-08
基金项目:公安部技术研究计划项目(2020JSYJC22,2021JSZ09)
This work was supported by the Ministry of Public Security Technology Research Program(2020JSYJC22,2021JSZ09).
通信作者:李欣(lixin@ppsuc.edu.cn)

如果将实体缺失检测任务作为序列标注问题进行处理，则在形式上与命名实体识别任务较为相似。命名实体识别是一项基础的自然语言处理任务。对于非结构化文本分析，命名实体识别是诸多复杂应用的基础工作。不同领域的应用中，根据需求的不同，对实体有着不同的定义。例如医学电子病历文本中，疾病名称、理化检验结果等内容被定义为实体；在金融领域，公司名称、产品名称等内容被定义为实体。文本中的实体识别问题已有许多研究与解决方案，其中 Lample 等[7]使用的长短时记忆网络-条件随机场（Bidirectional Long Short-Term Memory and Conditional Random Field，BiLSTM-CRF）模型是在各领域中命名实体识别任务应用得较为广泛的模型。

命名实体识别任务与实体缺失检测任务作为序列标注任务时，虽然具备形式上的相似性，但实体缺失的标注依赖于上下文推断实体的存在。这是该任务的主要难点。

对于表单等结构化文本，其缺失检测较为简单，已有多种方式对缺失内容进行补充。Liu 等[8]尝试利用平均值补全、决策树、K 近邻法等机器学习方法对医学检测结果、诊断数据进行补全。Biessmann 等[9]使用多种编码器对包括非结构化文本数据在内的数据进行编码，以为数据补全提供信息。Li 等[10]使用结合规则的方法对企业注册地址与类别信息进行规范化补全，实现了大规模企业注册数据分布的可视化。但这些方法不能主动地预知文本中的信息缺失，以上的模型只能在已知的空缺中填入文本，对于非结构化文本的实体缺失标注问题尚没有较为成熟的解决方法。

本文提出使用门控卷积网络（Gated Convolution Network，GCNN），通过对上下文提取语义信息来预测实体缺失。实验证明，该方法能够有效检测文本中的实体缺失，并对实体缺失的具体位置进行预测。同时，本文提出 3 种实体缺失标注模式并探索了它们对实体缺失预测准确率的影响，并从被标注字符字频的角度寻找 3 种标注方法产生差异的相关因素。

本文的主要工作是：将实体缺失检测问题作为序列标注问题进行处理，提出了基于门控卷积网络的实体缺失预测模型，并提出了 3 种训练实体缺失的数据标注方法；通过实验对不同数据标注方法的性能差异进行了研究，并通过模型对比和统计探究了被标记字符字频与模型预测性能的关系。

2　相关工作

2.1　序列标注任务

序列标注（Sequence Tagging）是自然语言处理中的基础任务，其任务是对一个序列的文本输出相应长度的标注序列，属于该类型的任务包括词性标注、语义角色标注、命名实体识别等。

本文所提出的实体缺失任务与命名实体识别任务存在基本形式上的相似性，并在识别目标上存在一定的关联性。命名实体识别需要找到文本中的实体，而实体缺失检测需要在文本中找到缺失实体的文本位置。基于任务形式的相似性，本文考察了已有的命名实体识别架构。

基于深度学习方法的命名实体识别方法中，长短时记忆网络-条件随机场架构是被研究得较为充分的神经网络结构之一。Graves 等[11]首先在语音识别中应用 LSTM。Lample 等[7]使用双向 LSTM（Bidirectional Long-short Term Memory，BiLSTM）与 CRF，融合字符向量、词向量完成文本中的命名实体识别。Ma 等[12]则利用卷积神经网络完成字符向量的特征抽取，再使用 LSTM-CRF 结构完成实体标注。

预训练语言模型的出现，如 BERT[13]，为 LSTM-CRF 获取更高层次语义特征、提升实体标注性能提供了更高效的文本表示方法，有效提高了命名实体识别的预测性能，因此在多个领域都得到了应用。Li 等[14]在医疗记录数据上应用 BERT-BiLSTM-CRF 模型，而 Liu 等[15]在中国古代史料文本上将其与 BERT-CRF 等模型进行对比，实验结果均印证了 LSTM-CRF 模型与预训练语言模型的组合应用在不同语料条件下的命名实体识别任务中具有较好的适用性。

Fu 等[16]通过检验测试集中的实体是否在训练集中出现并计算其覆盖比率，在多个公开数据集上通过实验发现，如果测试集实体未在训练集中出现，则模型在预测这一类实体时较为困难，这意味着实体字符本身对模型的预测有一定的影响。Agarwal 等[17]在模型中分解实体与实体上下文信息，分别通过命名实体识别实验印证了这一现象。实验结果说明，LSTM-CRF 模型无论是使用 BERT 还是基于词向量的方法，仅依赖实体的上下文进行实体识别极为困难，在各个指标上效果都较差。实体字符本身对于实体的识别十分重要。

对于实体缺失识别任务来说，其任务形式与命名实体识别任务具有相似性，均需要根据语义对特定的字符标注特定的标签：在命名实体识别中，模型需要对实体字符进行标注，在实体缺失检测中对实体缺失位置进行标注。但是该任务的挑战在于：如果实体字符不存在，基于命名实体识别的实体缺失检测模型能否对缺失实体位置的上下文信息进行捕捉和记忆。该问题是本文的关注点。

考虑到实体缺失检测对模型的上下文语义学习能力要求较高，本文使用了预训练语言模型作为文本表示方法。

2.2　文本内容检测

在不同的应用场景下对文本内容检测的任务目标有所区别。目前研究较为充分的问题包括文本评分和文本异常检测。

文本评分是对于一段文本，按照一定标准给出确定范围的一个数值作为文本质量的评估结果。该类研究通常应用于教育领域。

文本异常检测则是在大量文本中找出异常文本，即寻找内容质量与其他大部分样本偏差较大，或不符合特定要求的文本记录。

在文本评分方向，He 等[18]结合人工评分和多种预定义文本特征对作文内容进行自动化评分。基于深度学习的方法突破了手工设定规则的系统限制。Taghipour 等[19]尝试了 CNN，RNN 等神经网络结构，以在不需要特征工程的条件下解决作文评分问题。其中长短时记忆网络结构性能的表现较好。Alikaniotis 等[3]提出了一种模型来学习特定词语对最终分数的贡献度，结合长短时记忆网络捕获句子的含义，其评分性能超过了相似的深度学习方法。

文本评分能够很好地在特定标准下对文本质量进行评价。但应当注意的是，深度学习本身的不可解释性使得这种评分无法帮助改进文本质量。而本文所关注的实体缺失探测旨在观测实体缺失的具体位置。

在文本异常检测方向，Kumaraswamy 等[20]提出使用关系梯度提升法来融合领域知识对特定领域文本进行异常检测。但是此方法需要人工定义模型所需的一阶逻辑断言，以融合领域知识。Cichosz[21]在网络博客数据上应用了基于词向量的文本表示方法，并使用了支持向量机和聚类离群检测两种策略。该方法将异常检测作为分类问题研究，有助于文本挖掘，但无监督的方法依赖于参数的人工控制。在无监督深度学习方法上，Ruff 等[22]提出了融合上下文内容与注意力机制的异常文本分类方法。Ruff 等[23]总结了应用深度学习方法的文本异常检测，有 3 种基本建模思路，分别是基于分类的方法、基于概率的方法和基于文本重建的方法。这些方法都在特定领域内的应用中取得了较好的效果，但无法指明或分析文本中具体的异常位置与类别，且其大多数基于无监督的方法并不适合特定领域文本中的错误检查，而较为适合完整文本中在句子整体语义层面的异常侦测。

3 结合门控卷积网络的实体缺失检测方法

3.1 实体缺失预测

文本的信息存档、结构化提取和内容分析等工作与其文本中关键实体信息是否充足存在相关性。文本中的实体信息相对于文本所描述的内容是否充足、完整是文本处理中需要解决的问题。

如果某一文本数据对象中缺乏有效、足量的实体信息，那么对该文本对象进行更高级的文本信息分析有可能是无效的。

以与实体缺失检测相关的命名实体识别任务为例。命名实体识别作为自然语言领域处理的基本任务，其结果对于后续的文本分类应用、关系抽取、图谱构建等任务有着重要影响。但是目前的命名实体识别模型往往假设文本中已有充分的实体文本信息。在实际的业务应用中，这一假设往往不能被满足，因此需要对实体的缺失进行准确的标注，以帮助对数据完整性负有责任的一方完善、更正数据。以警情信息处理流程为例，对于实体信息缺失的警情文本，需要反向追溯至记录文本的专职人员对相关内容进行补充。

由于各领域实际业务中对实体的定义与要求各不相同，本文接下来将对实体缺失识别预测任务进行定义，并在公开数据集上验证模型在实体缺失预测任务上的能力。

3.2 模型任务与结构

本文提出将实体缺失检测任务定义为序列标注任务，任务的定义如下：实体缺失检测模型的输入为一段文本 $A=\{a_1,\cdots,a_n\}$。其中 a_n 为构成文本的字符，输出为每个字符间隔标注序列 $B=\{b_1,\cdots,b_{n-1}\}$。对于任意 n，$b_n\in\{d,O\}$，d 代表两字符间应存在实体，即间隔处存在实体缺失，O 代表两字符间不存在实体缺失。

作为序列标注任务的实体缺失预测任务的输入和输出与命名实体识别任务相似。其区别在于命名实体识别任务要对每一个字符进行标注，本文所定义的实体缺失识别任务是对字符之间是否存在缺失实体进行标注。

该任务在任务目标上与命名实体识别任务有显著差别：命名实体识别任务需要标注实体的位置，而实体缺失检测需要标注实体缺失的位置。但是两项任务在形式上具有相似性。

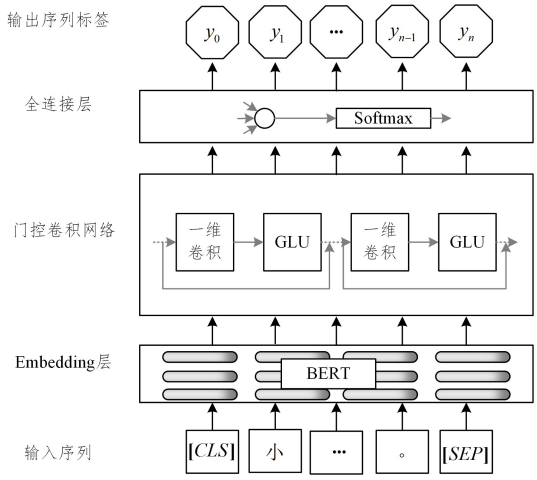

图 1 实体缺失检测模型

Fig. 1 Entity missing detection model

本文使用已有的命名实体识别的基本结构进行实体缺失检测任务。但本文没有使用在命名实体识别中广泛运用的 LSTM-CRF 结构作为标注输出层，因为 LSTM-CRF 面向存在实体字符的文本内容进行提取。在实体缺失识别中，需要检测的目标并不是实体字符，而是依据上下文推断某个字符附近是否缺失了实体。

本文提出了结合门控卷积神经网络[24]的实体缺失预测模型。该模型主要由两个部分组成：用于文本表示的 Embedding 层和用于实体缺失特征提取的门控卷积神经网络层。最后的全连接层用于将最后的输出向量提取特征降维，并使用 Softmax 函数获取相应字符的标注结果。

训练数据方面，沿用命名实体识别的数据标注方式，可发现存在 3 种等效的标注方式：在缺失实体的间隔左侧进行字符标注、在缺失实体的间隔右侧进行字符标注、在缺失实体的间隔两侧进行字符标注。详细的标注方法见本文第 4 节。

3.3 Embedding 层

本文选用了 BERT 作为 Embedding 层。BERT 由多个 Transformer[25]堆叠而成[13]。

BERT 需要在原始文本的前后分别增加两个特殊字符，即“[CLS]”和“[SEP]”，代表句子的起始和结束。其输出向量维数与预训练语言模型的训练初始参数有关。本文所使用的预训练语言模型参数为：BERT(L:12, H:768, A:12)。“L”为 Transformer 层数；“H”为隐藏层维数，即预训练语言模型输出向量的维数；“A”为多头注意力的堆叠数。

3.4 门控卷积网络层

在命名实体识别任务中使用的循环神经网络能够有效提取长距离的语义依赖信息。但是循环神经网络的主要问题是无法并行计算且容易产生梯度消失问题。长短时记忆网络引入门控机制解决了梯度消失问题，且其性能提升证实了门控机制对语义特征提取的有效性。

但基于循环神经网络的模型在单层计算中仅能从一个方向对文本进行识别计算。如果需要获取两个方向上的文本特征,则需要通过堆叠不同方向的循环神经网络来实现,增加了计算量且仅能机械地堆叠两个方向上的计算结果,仍然无法实现同时对某一字符周围的内容进行计算。

卷积神经网络能够实现并行计算,并在卷积核范围内同时包含上下文信息进行计算,能够有效消除循环神经网络的单向计算弊端。综合考虑到门控机制在语义信息提取中的优势,本文使用结合门控机制的卷积神经网络[24]进行实体缺失文本特征的提取。单层门控卷积神经网络的计算过程如图 2 所示。

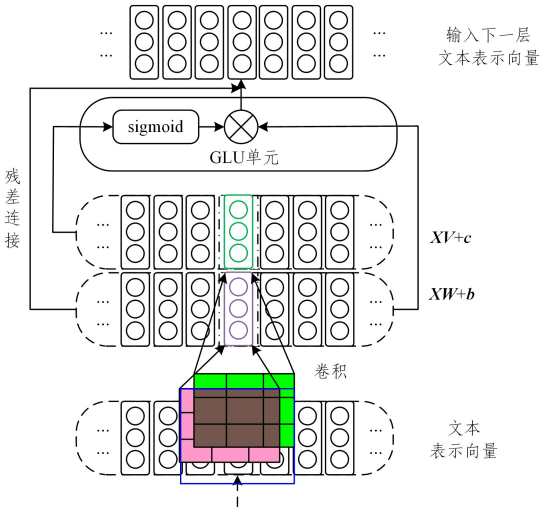

图 2　门控卷积神经网络(电子版为彩图)

Fig. 2　Gated convolutional neural network

如图 1 所示,获取 BERT 输出的文本表示向量后,门控卷积神经网络会对文本表示向量分别使用两个不同的特征图进行特征提取。

$$\boldsymbol{L}_A=\boldsymbol{X}\boldsymbol{W}+\boldsymbol{b} \tag{1}$$

$$\boldsymbol{L}_B=\boldsymbol{X}\boldsymbol{V}+\boldsymbol{c} \tag{2}$$

其中,$\boldsymbol{L}_A\in\mathbb{R}^{N\times n}$ 是卷积结果,$\boldsymbol{L}_B\in\mathbb{R}^{N\times n}$ 是经非线性函数激活后用作门控单元,N 是句子的长度,n 是字符嵌入的维度,$\boldsymbol{X}\in\mathbb{R}^{N\times m}$ 为 BERT 的输出,m 为 BERT 输出的字符嵌入维度。其中,$\boldsymbol{W}\in\mathbb{R}^{k\times m\times n}$,$\boldsymbol{V}\in\mathbb{R}^{k\times m\times n}$,$\boldsymbol{c}\in\mathbb{R}^{n}$,$\boldsymbol{b}\in\mathbb{R}^{n}$ 均为可以训练的参数,k 为卷积核大小。

其提取范围是模型超参数,在图 2 的示例中 kernel size 维度为 3,蓝色方框即为上下文内容的提取范围。

使用门控线性单元(Gated Linear Unit,GLU)对卷积结果进行处理。

$$GLU(\boldsymbol{L}_A,\boldsymbol{L}_B)=\boldsymbol{L}_A\otimes\sigma(\boldsymbol{L}_B) \tag{3}$$

其中,$\otimes$ 是矩阵的对应元素乘积,σ 代表非线性函数 sigmoid 函数。

在实际应用中需要堆叠多个卷积层以扩大实际上下文的感知范围。为避免深度卷积网络的退化问题,在每一层的门控卷积网络实现中使用残差连接。

$$\boldsymbol{H}_{(l)}=\sqrt{\boldsymbol{H}_{(l-1)}+GLU(\boldsymbol{L}_A,\boldsymbol{L}_B)} \tag{4}$$

最后使用全连接层对最后一层门控卷积神经网络的输出获得分类向量。

$$\boldsymbol{h}_{\text{out}}=\boldsymbol{W}_f\,\boldsymbol{H}_{(n)}+\boldsymbol{b}_f \tag{5}$$

其中,$\boldsymbol{h}_{\text{out}}$ 为全连接层输出;$\boldsymbol{W}_f$ 为全连接层的权重矩阵;$\boldsymbol{b}_f$ 为全连接层的偏置参数;$\boldsymbol{W}_f$ 和 $\boldsymbol{b}_f$ 均为模型中可训练的参数,随着模型的训练进行相应的调节。对于每个字符的输出使用 softmax 函数获得相应标签的概率,最终得到对应字符的标注。

4　实体缺失预测模型训练数据处理

为了训练模型对某种实体的缺失进行识别,模型需要标注实体缺失的训练数据集。但是目前公开的研究中没有符合本文需要的数据集。为此,本文提出了一种基于命名实体识别数据集的实体缺失数据训练数据集构造方法。该方法结合了已有的命名实体识别数据来构造数据。对于已在实际业务中应用了命名实体应用技术的领域,可直接利用已有的标注数据积累训练领域内的可用模型。

构造实体缺失预测模型训练数据需要准备命名实体识别数据集。首先对数据集进行过滤处理,仅保留一种标签,如图 3 中的步骤 1 所示,然后将该种标签所标注的实体字符从实体中去除。

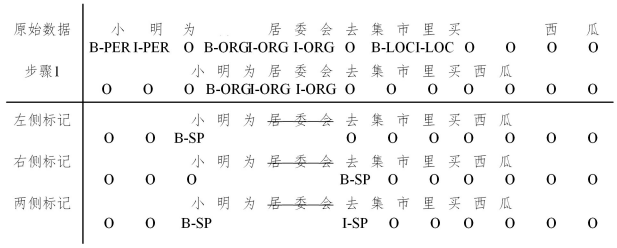

图 3　数据标注方法

Fig. 3　Annotation methods

相应地,为了训练模型预测实体缺失的能力,我们可以对被去除的实体的原位置两侧的字符进行标注。该方法的示例如图 3 中的 3 种标记方法。对于 3 种标记方法,我们都可以正确地推测出缺失实体的位置,即 3 种标记方法在应用中是等效的。

观察被去除实体后的文本,可见有 3 种标注方式:1)双侧标记法,将被去除实体两侧相邻的字符进行标记,标记时遵循 BIO 标记法,左侧标记为“B-SP”,而右侧标记为“I-SP”,其含义是两个字符间存在缺失的实体;2)右侧标记法,选择被去除实体右侧相邻的字符进行标记,标记为“B-SP”;3)左侧标记法,选择被去除实体左侧相邻的字符进行标记,标记为“B-SP”,如果实体为句子的第一个字符,则在句子的左侧添加一个句号(“。”),并标注为“B-SP”。

在图 3 的示例中,步骤 1 获得了除被标记为组织名(ORG)的标签外已被去除的句子。由于这一数据仅有一种实体标签,因此可称其为单标签数据集。

单种实体缺失检测数据有利于精确对比模型对不同种类实体缺失的识别能力,排除了多种缺失的相互干扰,为模型的性能比较提供了统一基准。

对于 ORG 单标签数据集,首先删除 ORG 标签所标记的字符,在示例中为“居委会”。其次根据不同的标记方法对剩下的字符进行标记。

左侧标记法在原位置左侧的字符标记了“B-SP”标记，即字符“为”，含义为该标记右侧应存在一个组织名实体。右侧标记法在原位置右侧的字符标记了“B-SP”标记，即字符“去”，含义为该标记左侧应存在一个组织名实体。而两侧标记法在“为”和“去”字符上都进行标记。按照命名实体识别的标记规范，“为”标记为“B-SP”，“去”标记为“I-SP”，含义为在两个标记之间应存在一个组织名实体。

由上述定义可知：3 种标记方法均能准确提示实体缺失的位置。在本文的实验中，将利用 3 种标记方式所构造的不同数据集训练模型，并从多个指标评估模型的预测性能，以确定最佳的实体缺失数据构造方式。

上述构建处理方法基于标准的命名实体识别数据进行处理，能够有效模拟实体单纯缺失的情况，可有效测定模型对实体缺失的检测能力。

5 实验

5.1 实验原始数据集

我们选取了中文命名实体识别任务的公开数据集 MSRA[26] 和人民日报 1998 年中文标注数据集。

其中，“MSRA”代表 MSRA 数据集，“PD”代表人民日报 1998 年中文标注数据集。为简化表述，后文均使用上述缩写代表具体数据集，“train”表示训练集，“test”则为测试集。MSRA 数据集中训练集和测试集分别有 45 000 条和 3 442 条语句。PD 数据集的训练集和测试集分别有 20 864 条和 4 636 条语句。图 4 给出了数据集中的句子长度分布情况。表 1 列出了数据集的基本信息。

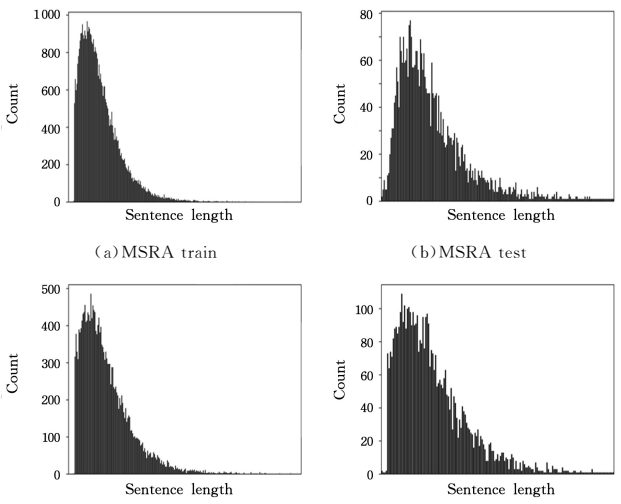

图 4 句子长度分布

Fig. 4 Sentences length distribution

表 1 数据集信息

Table 1 Information of datasets

数据集	实体数量			句子总数	平均长度
	LOC	ORG	PER		
MSRA train	36 860	20 584	17 615	45 000	48.26
PD train	16 571	9 277	8 144	20 864	46.93
MSRA test	2 886	1 331	1 973	3 442	50.15
PD test	3 658	2 185	1 864	4 636	47.28

5.2 实验数据集

由于两种数据中均含有 3 种标签：“PER”“ORG”和“LOC”。如果在一个模型中同时实现多种类型的实体缺失的预测，训练集中所去除的文本字符内容过多，不利于测定模型对具体种类实体缺失的识别能力，同样不利于测定不同标注方式对不同种类实体缺失的具体性能。因此，遵循上文中提出的实体缺失训练数据预处理方法。MSRA 数据集与 PD 数据集各被分为 3 份文本数据一致但仅带有一种标签的数据集。每份数据中仅有一种标签（“PER”“ORG”或“LOC”），即本文第 3 节所定义的单标签数据集。

在后文中，以数据集名称加上实体标签代表特定的单标签数据集。如“MSRA LOC”代表 MSRA 数据集仅保留 LOC 标签的单标签数据集。

根据第 3 节提出的 3 种不同的标注方法，对每一份数据各进行 3 种标注，形成 3 种不同标注方法的数据集。不同标注方法的数据集都将以同样的参数在多个实验模型上进行测试，以分析最佳的标注方式与缺失检测模型。

本文在实验结果数据表格中使用“tag-B”来代表该数据使用双侧标记法，而“tag-R”代表该数据使用右侧标记法，“tag-L”代表该数据使用左侧标记法。“LOC”代表该数据仅保留地点名实体，“ORG”代表该数据仅保留组织名实体，“PER”代表该数据仅保留姓名实体。

5.3 实验设置与模型超参数

本文对比了基于 BERT、基于 BERT 和 LSTM 以及基于 BERT 和 LSTM-CRF 的模型，它们均为执行与实体缺失检测任务相似的命名实体识别任务的典型模型。

本文每组实验训练时使用的超参数配置相同，Batch Size 为 16，Epoch 为 10，学习率设置为 10^{-6}，使用线性增长动态控制学习率（Warmup[27]），使用 ADAM[28] 作为优化器，投影层隐藏单元数量为 100，Dropout 为 0.1。

BERT+GCNN 模型共有 3 个堆叠的卷积层，kernel_size 分别为 4，6，8。

所有实验中使用的 LSTM 均为双向 LSTM，输出隐藏层维度为 256，LSTM 层的 Dropout 为 0.1。

5.4 评价指标

本文选取精准率、召回率、F1 作为模型的评价指标。其中精准率的计算式如式(6)所示，召回率的计算式如式(7)所示，F1 的计算式如式(8)所示。

$$precision=\frac{T_P}{T_P+F_P}\times 100\% \tag{6}$$

$$recall=\frac{T_P}{T_P+F_N}\times 100\% \tag{7}$$

$$F_1=2\cdot\frac{precision\cdot recall}{precision+recall} \tag{8}$$

其中，T_P 是分类正确的正样本，F_P 是分类错误的正样本，F_P 是分类错误的正样本，F_N 是分类错误的负样本。

5.5 实验结果

表 2 列出了使用 BERT 时，4 种模型在不同数据集上的 F1。

实验结果证明，无论是哪一种类型的标签，在使用 BERT+GCNN 模型时，使用左侧标记法标记数据，标注能获得最佳的预测效果。表 2 中加粗的数据结果为一种数据集中该种

标签所达到的最佳预测结果。

另外，对比不同模型在使用“来源相同但处理方法不同的数据集”进行训练后的表现可发现：仅使用 BERT 的模型比 BERT＋LSTM＋CRF 模型的效果更佳。使用 BERT＋LSTM 模型的效果相比 BERT＋LSTM＋CRF 也有所提升。但是使用左侧标记法的数据，BERT＋LSTM＋CRF 模型的效果通常要优于 BERT＋LSTM。

6　左侧标记法的优势分析

表 2 所列的结果都证实了左侧标记法的优势。在不具备 LSTM＋CRF 时，模型学习文本上下文信息的能力相对较弱，但实验却证明不具备 LSTM-CRF 层时效果更佳。

表 2　模型实验结果

Table 2　Experiment results

数据集	LOC				ORG				PER			
	BERT＋GCNN	BERT	BERT＋LSTM	BERT＋LSTM＋CRF	BERT＋GCNN	BERT	BERT＋LSTM	BERT＋LSTM＋CRF	BERT＋GCNN	BERT	BERT＋LSTM	BERT＋LSTM＋CRF
MSRA tag-B	74.70	73.53	72.15	69.03	75.51	72.94	71.59	67.27	80.60	80.83	76.88	72.94
MSRA tag-R	72.31	72.19	72.04	69.71	76.61	76.55	75.25	74.62	80.48	79.73	78.82	76.76
MSRA tag-L	**78.84**	78.70	75.76	77.10	**78.67**	77.80	75.53	77.13	**84.55**	83.78	82.38	83.70
PD tag-B	75.45	68.55	67.19	63.06	78.72	76.17	71.79	66.82	82.40	81.23	76.49	69.75
PD tag-R	74.89	73.08	72.29	67.77	79.43	79.01	75.29	76.98	83.33	83.03	80.43	81.32
PD tag-L	**80.45**	79.65	78.55	72.29	**83.02**	81.99	79.55	80.45	**86.75**	86.53	82.54	84.38

因此考虑如下问题：模型预测实体的缺失时是否更关注特定的字符，而不是潜在实体位置的上下文？如果模型更关注字符本身，被标注字符的频率分布是否与模型的性能存在相关性？

从这一问题出发，本文对被标注的字符进行了统计，即统计实体左侧与右侧出现字符的频数，以探索左侧标记法具有优势的原因。

图 5 给出了 PD 数据集和 MSRA 数据集中的 ORG 实体左侧和右侧出现频次不少于 100 的字符。曲线的数据点越多，意味着在该数据集中出现频次不少于 100 的字符就越多。但每一条曲线中在同一个 X 轴坐标上的字符并不相同而直接叠加了不同的数据曲线，以直观展示字符频次的分布差别。

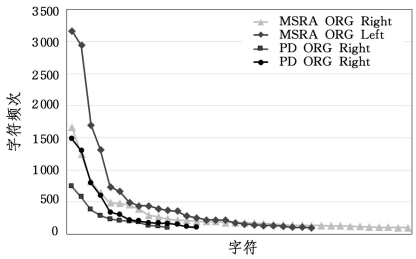

图 5　字符频次比较图

Fig. 5　Characters frequency comparison

从图 5 可以发现，无论是 PD 数据集还是 MSRA 数据集中，标记为“Left”的曲线显著高于对应的标记为“Right”的曲线，这意味着实体左侧的字符分布相对集中：因为实体数量一致，所以实体两侧字符的数量一致。曲线越偏右上部分（即曲线越“高”），意味着字符分布越集中。同时，曲线的数据点越多，意味着字符分布越集中。虽然 MSRA ORG Right 曲线的长度比 MSRA ORG Left 更长，但是 MSRA ORG Left 曲线头部的字符频次显著多于 MSRA ORG Right。

为直观形象地展现实体两侧被标注字符集中于哪些具体的字符，本文选择了 PD LOC 数据集进行统计展示。该数据集在两种标记法中所标记的字符中，频次超过 100 的字符数量相同，便于直观展示两种标记方法的分布差异性，统计结果如图 6 所示。

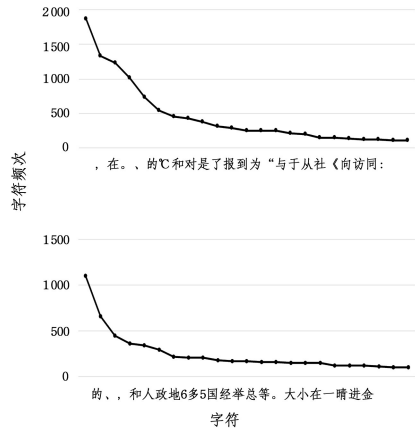

图 6　字符频次分布

Fig. 6　Characters frequency distribution

图 6 中横坐标中的字数完全一样，即两侧字符中出现的频次超过 100 的字符数量一致。可观察到，左侧字符的分布显然更加集中。

在左侧标记法（LEFT）中，出现次数最多的前 3 个字符的频次都超过了 1 200。而在右侧标记法（RIGHT）中，出现次数最多的字符的频数也未达到 1 200。在左侧标记法所标记的字符中，“在”“到”等字符一定程度上反映了中文的基本语言习惯。关于其中出现的“℃”字符，人工检查数据集发现，是训练集中出现了天气预报内容信息，恰好与大量地名信息邻接。

为精确量化实体两侧字符的分布集中程度，本文统计了所有训练集数据实体两侧的字符分布情况。

表 3 列出了训练集中每种实体左侧与右侧的字符分布情况。表 3 上半部分为每种类别的实体左侧或右侧字符中出现频次不少于 100 的字符个数，下半部分为这些字符占所有被标注字符的比例。其比例越大，意味着字符的频数分布更为集中，即该种类实体左侧或右侧的字符倾向于特定的少数字符。

表 3 实体两侧字符分布

Table 3 Characters distribution around entities

数据集	出现频次不小于100的字符个数			频次不少于100的字符占被标注字符的比例		
	LOC	ORG	PER	LOC	ORG	PER
MSRA Left	41	26	28	4.8%	4.0%	4.5%
MSRARight	52	36	28	3.8%	3.5%	3.0%
PD Left	23	14	14	3.6%	2.9%	3.1%
PD Right	23	11	17	2.2%	1.4%	2.5%

表 3 中,“Left”代表该行统计结果为实体左侧字符,“Right”则为相应的实体右侧字符统计结果。

表 3 中的数据说明:实体左侧的字符无论是在任何一个数据集中的任何一个类别的实体,其分布都无一例外比实体右侧字符更为集中。结合表 2 中左侧标记法相对于右侧标记法的优势,本文认为,被标注字符的字频分布是否集中与标注效果间存在相关性。

另外,可观察到 PER 类型实体和 ORG 类型实体两侧出现频次不小于 100 的字符个数显著少于 LOC 类型的实体,在频次不少于 100 的字符占被标注字符的比例相近的情况下,意味着预测 PER 和 ORG 类型实体缺失的特征更为集中,即模型学习所需的特征应更为简单。相应地,在表 2 中可观察到,预测 PER 实体缺失较预测另外两种实体缺失的性能更优,而预测 LOC 实体缺失的性能最差。

如果假设模型通过“记忆”实体两侧的字符来标注实体的缺失,即假设字符分布越集中,“记忆”相关字符的难度就相对降低,那么应当对如何提高模型的泛化性能进行研究。

结束语 本文聚焦于文本中的实体信息缺失,并提出了基于序列标注任务,使用门控卷积神经网络进行文本实体缺失的识别。

对于训练实体缺失识别模型,本文考虑了 3 种实体缺失识别模型的训练数据构造标记方法。实验证明,在本文提出的 3 种等效标记方法中,以 F1 为衡量指标,左侧标记法所构造的训练集能够最好地识别相关实体的缺失。

为探索左侧标记法在实体缺失识别上的优势,本文从被标记字符的字频进行统计分析,结果展示了左侧标记法的字符字频的分布集中性。结合在多个模型中的实验结果,本文认为被标记字符字频的分布集中性与左侧标记法的性能优势存在相关性。接下来将进一步研究如何提升实体缺失识别准确率与模型识别实体缺失的具体机制。

参 考 文 献

[1] FAN M, FENG C, GUO L, et al. Product-Aware Helpfulness Prediction of Online Reviews[C]// The World Wide Web Conference(WWW '19). ACM Press, 2019: 2715-2721.

[2] YANG Y, CHEN C, BAO F S. Aspect-Based Helpfulness Prediction for Online Product Reviews[C]// 2016 IEEE 28th International Conference on Tools with Artificial Intelligence(ICTAI). IEEE, 2016: 836-843.

[3] ALIKANIOTIS D, YANNAKOUDAKIS H, REI M. Automatic Text Scoring Using Neural Networks[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers). Association for Computational Linguistics, 2016: 715-725.

[4] TAY Y, PHAN M C, TUAN L A, et al. SkipFlow: Incorporating Neural Coherence Features for End-to-End Automatic Text Scoring[C]// Thirty-Second AAAI Conference on Artificial Intelligence. Association for the Advancement of Artificial Intelligence, 2018: 5948-5955.

[5] SUN F, ZHANG J. Research on Grammar Checking System Using Computer Big Data and Convolutional Neural Network Constructing Classification Model[J]. Journal of Physics: Conference Series, 2021, 1952(4): 042097, 1-9.

[6] HAO S, HAO G. A Research on Online Grammar Checker System Based on Neural Network Model[J]. Journal of Physics: Conference Series, 2020, 1651(1): 012135, 1-8.

[7] LAMPLE G, BALLESTEROS M, SUBRAMANIAN S, et al. Neural Architectures for Named Entity Recognition[C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Association for Computational Linguistics, 2016: 260-270.

[8] LIU Y, GOPALAKRISHNAN V. An Overview and Evaluation of Recent Machine Learning Imputation Methods Using Cardiac Imaging Data[J]. Data, 2017, 2(1): 8, 1-15.

[9] BIESSMANN F, RUKAT T, SCHMIDT P, et al. DataWig: Missing Value Imputation for Tables[J]. Journal of Machine Learning Research, 2019, 20(175): 1-6.

[10] LI F, GUI Z, WU H, et al. Big enterprise registration data imputation: Supporting spatiotemporal analysis of industries in China[J]. Computers, Environment and Urban Systems, 2018, 70: 9-23.

[11] GRAVES A, MOHAMED A R, HINTON G. Speech Recognition with Deep Recurrent Neural Networks[C]// 2013 IEEE International Conference on Acoustics, Speech and Signal Processing. IEEE, 2013: 6645-6649.

[12] MA X, HOVY E. End-to-end Sequence Labeling via Bi-directional LSTM-CNNs-CRF[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics(Volume 1: Long Papers). Association for Computational Linguistics, 2016: 1064-1074.

[13] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1(Long and Short Papers). 2019: 4171-4186.

[14] LI X, ZHANG H, ZHOU X H. Chinese clinical named entity recognition with variant neural structures based on BERT methods[J]. Journal of Biomedical Informatics, 2020, 107: 103422, 1-7.

[15] LIU S, YANG H, LI J, et al. Chinese Named Entity Recognition Method in History and Culture Field Based on BERT[J]. International Journal of Computational Intelligence Systems, 2021, 14(1): 163.

[16] FU J, LIU P, ZHANG Q. Rethinking Generalization of Neural

Models：A Named Entity Recognition Case Study[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020：7732-7739.

[17] AGARWAL O，YANG Y，WALLACE B C，et al. Interpretability Analysis for Named Entity Recognition to Understand System Predictions and How They Can Improve[J]. Computational Linguistics，2021，47(1)：117-140.

[18] CHEN H，HE B. Automated Essay Scoring by Maximizing Human-Machine Agreement[C]// Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics，2013：1741-1752.

[19] TAGHIPOUR K，NG H T. A Neural Approach to Automated Essay Scoring[C]// Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics，2016：1882-1891.

[20] KUMARASWAMY R，WAZALWAR A，KHOT T，et al. Anomaly Detection in Text：The Value of Domain Knowledge [C]// Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference. Association for the Advancement of Artificial Intelligence，2015：225-228.

[21] CICHOSZ P. Unsupervised modeling anomaly detection in discussion forums posts using global vectors for text representation [J]. Natural Language Engineering，2020，26(5)：551-578.

[22] RUFF L，ZEMLYANSKIY Y，VANDERMEULEN R，et al. Self-Attentive，Multi-Context One-Class Classification for Unsupervised Anomaly Detection on Text[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics，2019：4061-4071.

[23] RUFF L，KAUFFMANN J R，VANDERMEULEN R A，et al. A Unifying Review of Deep and Shallow Anomaly Detection[J]. Proceedings of the IEEE，2021，109(5)：756-795.

[24] DAUPHIN Y N，FAN A，AULI M，et al. Language Modeling with Gated Convolutional Networks[C]// Proceedings of the 34th International Conference on Machine Learning. 2017：933-941.

[25] VASWANI A，SHAZEER N，PARMAR N，et al. Attention is All You Need[C]// The 31st International Conference on Neural Information Processing Systems. Curran Associates Inc.，2017：6000-6010.

[26] LEVOW G A. The Third International Chinese Language Processing Bakeoff：Word Segmentation and Named Entity Recognition[C]// Proceedings of the 5th SIGHAN Workshop on Chinese Language Processing. Association for Computational Linguistics. 2006：108-117.

[27] HE K，ZHANG X，REN S，et al. Deep Residual Learning for Image Recognition[C]// the IEEE Conference on Computer Vision and Pattern Recognition. IEEE，2016：770-778.

[28] KINGMA D P，BA J. Adam：A Method for Stochastic Optimization[C]// 3rd International Conference on Learning Representations. 2015.

YE Han，born in 1999，postgraduate. His main research interests include natural language processing and deep learning.

LI Xin，born in 1977，Ph. D，associate professor. His main research interests include big data processing and information communication.

（责任编辑：喻藜）