CGR-BERT-ZESHEL:基于中文特征的零样本实体链接模型

潘建, 吴志伟, 李燕君

引用本文

潘建, 吴志伟, 李燕君. CGR-BERT-ZESHEL:基于中文特征的零样本实体链接模型[J]. 计算机科学, 2025, 52(4): 262-270.

PAN Jian, WU Zhiwei, LI Yanjun. CGR-BERT-ZESHEL:Zero-shot Entity Linking Model with Chinese Features [J]. Computer Science, 2025, 52(4): 262-270.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240100119

CGR-BERT-ZESHEL:基于中文特征的零样本实体链接模型

潘　建[1,2]　吴志伟[1]　李燕君[1]
1 浙江工业大学计算机科学与技术学院　杭州 310023
2 浙江工业大学之江学院　浙江 绍兴 312030

摘　要　目前,在实体链接任务的研究中,对中文实体链接、新兴实体与不知名实体链接的研究较少。此外,传统的 BERT 模型忽略了中文的两个关键方面,即字形和部首,这两者为语言理解提供了重要的语法和语义信息。针对以上问题,提出了一种基于中文特征的零样本实体链接模型 CGR-BERT-ZESHEL。该模型首先通过引入视觉图像嵌入和传统字符嵌入,分别将字形特征和部首特征输入模型,从而增强词向量特征并缓解未登录词对模型性能的影响;然后采用候选实体生成和候选实体排序两阶段的方法得到实体链接的结果。在 Hansel 和 CLEEK 两个数据集上进行实验,结果表明,与基线模型相比,CGR-BERT-ZESHEL 模型在候选实体生成阶段的性能指标 Recall@100 提高了 17.49%和 7.34%,在候选实体排序阶段的性能指标 Accuracy 提高了 3.02%和 3.11%;同时,在 Recall@100 和 Accuracy 指标上的性能均优于其他对比模型。

关键词:实体链接;中文零样本;BERT;候选实体生成;候选实体排序
中图分类号　TP391

CGR-BERT-ZESHEL:Zero-shot Entity Linking Model with Chinese Features

PAN Jian[1,2],WU Zhiwei[1] and LI Yanjun[1]
1 College of Computer Science and Technology,Zhejiang University of Technology,Hangzhou 310023,China
2 Zhijiang College of Zhejiang University of Technology,Shaoxing,Zhejiang 312030,China

Abstract　Currently,the research on entity linking tasks is less on Chinese entity links,emerging entities and unknown entity links. Additionally,traditional BERT models ignore two crucial aspects of Chinese,namely glyphs and radicals,which provide important syntactic and semantic information for language understanding. To solve the above problems,this paper proposes a zero-shot entity linking model based on Chinese features called CGR-BERT-ZESHEL. Firstly,the model incorporates glyph and radical features by introducing visual image embedding and traditional character embedding,respectively,to enhance word vector features and mitigate the effect of out-of-vocabulary words. Then,a two-stage method of candidate entity generation and candidate entity ranking is used to obtain the results. Experimental results on the two datasets which include Hansel and CLEEK show that compared with the baseline model,the performance metric Recall@100 is improved by 17.49% and 7.34% in the candidate entity generation stage,and the performance metric accuracy is improved by 3.02% and 3.11% in the candidate entity ranking stage. Meanwhile,the proposed model also outperforms other baseline models in both Recall@100 and Accuracy metric.

Keywords　Entity linking,Chinese zero-shot,BERT,Candidate entity generation,Candidate entity ranking

1　引言

实体链接(Entity Linking)[1]任务的目标是将文本中的实体指称项链接到知识库中对应的实体,以明确文本中实体指称项所代表的具体实体,从而消除实体指称项的歧义性。实体链接是问答[2]、知识库推理[3]和对话[4]等应用程序的基础组件,在信息检索、问答、推荐系统、文本摘要等应用中发挥着重要作用。

目前,在实体链接任务的研究中,对中文实体链接的研究相对较少,并存在难以准确链接新兴实体和不知名实体的挑战。零样本实体链接问题由 Logeswaran 等[5]提出,旨在将实体指称项链接到在训练期间未出现的实体,以提高模型对新兴实体和不知名实体进行准确链接的能力。

为了解决中文新兴实体链接的问题,本文提出了一种基于中文特征的零样本实体链接模型 CGR-BERT-ZESHEL。该模型使用字形和部首两种中文独特的特征来补充字符特征语义的不足。通过引入视觉图像嵌入和传统字符嵌入,分别将字形和部首特征输入模型,对词向量进行增强,从而缓解未登录词对模型的影响。该模型采用候选实体生成和候选实体排序两阶段的方法来实现零样本实体链接。与传统的实体

到稿日期:2024-01-15　返修日期:2024-05-15
基金项目:浙江省自然科学基金探索项目(LGF20F020015)
This work was supported by the Natural Science Foundation of Zhejiang Province,China(LGF20F020015).
通信作者:潘建(pj@zjut.edu.cn)

链接任务不同，模型没有使用任何的外部信息和别名表等辅助数据进行链接判断，因此能够更好地适应零样本实体链接任务。本文的主要贡献如下：

1）提出了一种基于中文特征的零样本实体链接模型，选择了字形和部首两种特征，有效地增强了中文特征信息；

2）在 Hansel 和 CLEEK 数据集上取得了较好的结果，性能与类型增强的 Dual Encoder 和 Cross-Attention Encoder 模型相当。

本文第 2 章对相关工作进行了概述；第 3 章定义了零样本实体链接问题；第 4 章详细介绍了本文提出的模型；第 5 章分析了实验结果；最后进行总结并对未来的研究进行了展望。

2　相关工作

2.1　大规模预训练模型

近年来，NLP 领域对大规模预训练模型开展了大量的研究工作。预训练模型是一种深度学习模型，通过训练大规模文本数据来学习词汇、语法和语义等语言特征。BERT[6] 建立在 Transformer 架构[7] 之上，利用掩码语言模型（Masked Language Model，MLM）任务和下一句预测（Next Sentence Prediction，NSP）任务，在大规模未标记文本语料库上进行了无监督学习和预训练。在 BERT 模型的基础上，研究人员对掩码策略[8]、预训练任务[9] 以及模型结构[10] 进行了改进。例如 RoBERTa[11] 移除了下一句预测任务，改进了掩码策略，并使用更大规模的语料库进行训练，显著提升了模型在多项 NLP 任务上的性能。GPT 系列[12-13]、Gopher[14]、GPT-J[15] 和其他 BERT 变体[16-18] 将大规模无监督预训练模型应用于机器翻译、文本摘要和对话生成等文本生成任务，并在这些任务中表现出卓越的性能。

与英文不同，中文是象形字，在句法和词汇方面具有独特的特点。因此，英文的预训练模型不能直接应用于中文的文本处理。Li 等[19] 提出使用汉字作为基本单位，而不是使用英语中的单词或词根。ERNIE[20] 使用了 3 种不同类型的掩码策略，包括字符级别掩码、短语级别掩码和实体级别掩码，以增强对多粒度语义捕获的能力，更好地理解不同层次和维度上的语义信息。Cui 等[21] 使用全词掩码策略对模型进行预训练，即将中文词语中的所有字符完全遮掩，以更好地理解和处理中文语义信息，提高模型的处理能力和准确性。Chinese BERT[22] 将中文的字形和拼音信息结合到语言训练模型中，通过从视觉上捕获汉字语义和拼音来解决中文中的异音现象，提升模型的性能。Instruct GPT[23] 使用基于人类反馈的强化学习进行训练，并通过两步微调使语言模型与用户意图保持一致。ChatGPT[24] 使用语言建模任务在大量文本数据（包括书籍、文章和网站）上进行预训练，学习自然语言中单词和短语之间的模式和关系，使其能够在对话中生成连贯和准确的答案。

2.2　实体链接

实体链接的定义中假设实体指称项是用户或检测系统给定的，而知识库中的实体集包含数万甚至数百万个实体，因此实体链接是一项具有挑战性的任务。在实体链接任务的研究中，许多实体链接系统都依赖于别名表、频率统计信息、结构化数据等资源或假设。

2013 年，He 等[25] 使用深度神经网络模型首次解决了实体链接任务，所提方法在英文实体链接数据集上取得了良好的性能。这一突破，推动了英语领域实体链接任务的快速发展。然而，相对于英文任务，中文实体链接由于知识库和数据集的缺乏，研究进展相对缓慢。

为了提取实体链接的主题特征，Chen 等[26] 构建了主题关系图模型，该模型在主题建模方面优于传统的 TF-IDF 和 LDA 等主题模型。同时，他们还构建了 4 种不同类型的中文短文本实体链接数据集，包括 HQA，NTF，NLPCC 和 CNDL，促进了中文实体链接任务的发展。Ouyang 等[27] 使用上述 4 种中文短文本实体链接数据集，提出了一种基于 GRU 的多交叉匹配模型，为解决短文本实体链接问题提供了新的思路。Hua 等[28] 提出了一种名为 XREF 的基于注意力机制的模型，用于链接新闻评论和文章，并通过在大规模未标记语料库上进行自监督训练，实现了对实体链接任务的有效应用。

为了从庞大的知识库中准确并有效地识别目标实体，大多数实体链接方法[29-30] 由两个阶段组成：一是候选实体生成，利用检索器从知识库中检索出可能的候选实体，如使用频率信息[31] 或基于稀疏模型[32] 等方法来快速检索候选实体；二是候选实体排序，利用排序器对候选实体重新排序并选择最有可能的实体，如使用神经网络对候选实体进行排序。

Wiatrak 等[33] 提出了一种基于代理的度量学习损失与对抗性正则化器相结合的方法，为候选检索阶段的硬负抽样提供了一种有效的替代方案。Xu 等[34] 在候选实体排序阶段提出了基于 LUKE 的交叉编码器，将实体作为附加信息输入到交叉编码器中，进而提高实体链接的性能。

Wang 等[2] 和 Sevgili 等[35] 分别对 2015 年前和 2015 年后的实体链接方法进行了详细的调查。Wang 等[2] 阐述了实体生成、实体排序和不可链接实体预测 3 个模块中使用的算法和特征，介绍了实体链接系统的评估方法，并讨论了未来的发展方向。Sevgili 等[35] 全面描述了最新的实体链接系统，提炼了一个通用的实体链接系统架构，总结了每个部分的主要方法以及通用架构的各种变体。

2.3　零样本实体链接

Logeswaran 等[5] 在实体链接领域首次提出了零样本实体链接任务，并发布了一个英文的零样本实体链接数据集。该任务的目标是在不使用任何已标注的训练数据的情况下，将实体指称项链接到新兴实体。在该模型中，对于每个实体指称项，首先使用 BM25 算法[32] 生成 64 个候选实体，然后将每个候选实体和实体指称项及其上下文拼接输入到 BERT[6] 中，对候选实体进行排序。在 Logeswaran 等[5] 提出的模型的基础上，Petar 等[36] 进一步提出了一种名为 KG-ZESHEL 的知识图增强的零样本实体链接方法，通过引入实体指称项信息和实体辅助信息，提升了零样本实体链接的性能。

Xu 等[37] 构造了中文零样本实体链接的数据集，填补了中文零样本实体链接任务的空白，并在候选实体生成阶段使用别名表（Alias Table）得到候选实体集结果，在候选实体排序阶段使用 Cross-Attention Encoder 模型对上一阶段的结果进行排序并预测结果，从而实现零样本实体链接。Huang

等[38]提出了基于词典的 Coarse-to-fine 实体检索器，以有效地检索候选实体，并使用基于 BERT 的双编码器对候选实体进行重新排序。Zhou 等[39]提出了基于 ERNIE 和对抗性训练的中文零样本实体链接模型，其通过在训练过程中加入对抗性扰动，实现对中文零样本实体的准确链接。

上述模型存在两个问题：一是对模型输入中未登录词的处理过于简单，如 TF-IDF 和 BM25 等模型采用直接忽略或使用[UNK]标记的方法处理未登录词；二是忽略了中文语言特征对模型的重要性，如传统的 BERT 模型在中文编码时使用字符级别的分词方法将文本分解成子词序列，而忽略了中文语言的独特特征。针对以上问题，本文提出一种中文零样本实体链接模型——CGR-BERT-ZESHEL，从字形和部首两个方面增强中文特征，缓解未登录词的问题，并增强输入的有效性。

3 问题定义

零样本实体链接任务类似于跨域分类任务，其中实体扮演类的角色。图 1 通过一个实例阐述了 CGR-BERT-ZESHEL 模型的总体结构，其中 T_B 表示 BI-ENCODER 架构，T_C 表示 CROSS-ENCODER 架构。

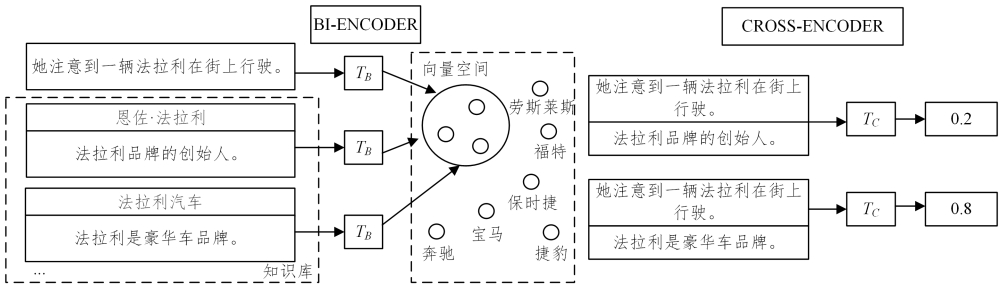

图 1 零样本实体链接任务的总体结构

Fig. 1 Overall structure of zero-shot entity linking tasks

形式上，在零样本实体链接任务中，本文定义一个实体集 ε_S，也称为可见的实体集。在训练过程中，每个实体 $e\in\varepsilon_S$ 都以训练集中实体指称项的最优实体的形式出现。同时，本文定义另一个实体集 ε_U，也称为不可见的实体集。这些实体集的具体组成如下：$\varepsilon_S=\{(e_i,d_i)\}_{i=1}^{K}$，$\varepsilon_U=\{(e_i,d_i)\}_{i=(K+1)}^{L}$。其中，$\varepsilon_S\cap\varepsilon_U=\varnothing$，$e_i$ 是实体的唯一标题，d_i 是实体的简短文本描述。训练集、验证集和测试集的具体结构如下：$D_{\text{train}}^{\text{zel}}=\{(m_i,e_i)\mid e_i\in\varepsilon_S\}_{i=1}^{N}$；$D_{\text{val}}^{\text{zel}}=\{(m_j,e_j)\mid e_j\in\varepsilon_S\}_{j=1}^{V}$；$D_{\text{test}}^{\text{zel}}=\{(m_k,e_k)\mid e_k\in\varepsilon_U\}_{k=1}^{M}$。其中，$N$，$V$ 和 M 分别代表训练集、验证集和测试集中数据的个数，并且 $D_{\text{train}}^{\text{zel}}\cap D_{\text{test}}^{\text{zel}}=\varnothing$。在上述数据集中，$m_i$，$m_j$，$m_k$ 是包含实体指称项的文本字符串；e_i，e_j，e_k 是实体指称项在知识库中链接的实体，即最优实体。其中，需要将 m_i 表示为以下元组：(左上下文，实体指称项，右上下文)，如“她注意到一辆法拉利在街上行驶”，m_i=(“她注意到一辆”，“法拉利”，“在街上行驶”)。

零样本实体链接任务可以描述为在训练集 $D_{\text{train}}^{\text{zel}}$ 上进行训练，然后使用测试集 $D_{\text{test}}^{\text{zel}}$ 中的实体指称项链接到知识库中的相应实体的过程。在零样本实体链接模型中，利用评分函数 $f_{\text{zel}}:M\times\varepsilon\rightarrow R$(其中 $\varepsilon=\varepsilon_S\cup\varepsilon_U$)计算每个候选实体的得分 $f_{\text{zel}}(m,e)$，并选择得分最高的实体作为最终答案。

4 模型

CGR-BERT-ZESHEL 模型在传统的 BERT 输入中融入了中文特征，即图像维度的字形特征 gly_i 和字符维度的部首特征 rad_i。模型的总体流程如下：

步骤 1 将实体指称项及其上下文 m_i 和实体及其描述 (e_i,d_i) 分别输入两个双编码器；

步骤 2 两个双编码器分别将 m_i 和 (e_i,d_i) 编码为密集向量，并对每个实体和实体指称项进行点积评分，得到 m_i 对应的前 k 个候选实体 $E_i=\{e_j\}_{j=1}^{k}$；

步骤 3 将双编码器检索到的 E_i 传递给交叉编码器进行候选实体排序；

步骤 4 将 m_i 分别与 E_i 中每个候选实体进行拼接作为交叉编码器的输入，并应用额外的线性层来计算每对的最终分数；

步骤 5 通过 Softmax 进行归一化并取分数最大值对应的实体作为模型的预测输出。

4.1 模型输入

模型的输入包括位置嵌入、片段嵌入、字符嵌入、字形嵌入和部首嵌入。前三者是传统 BERT 的嵌入方式，字符嵌入的处理方式类似于 BERT 中使用的令牌嵌入，但在字符粒度上，本文增加了视觉领域的字形嵌入和字形领域的部首嵌入作为中文独有的特征。图 2 展示了本文所提模型的输入结构。其中，□代表字形嵌入，○代表部首嵌入，⊕代表相加。

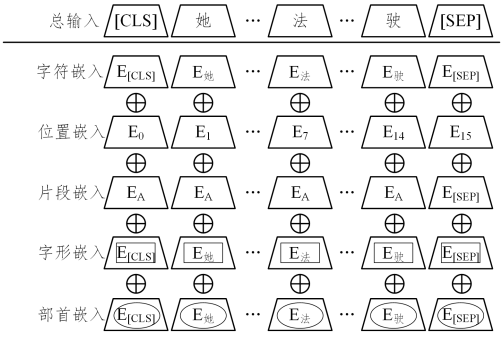

图 2 CGR-BERT-ZESHEL 的输入

Fig. 2 Input of CGR-BERT-ZESHEL

4.1.1 字形嵌入

本文使用了 GlyphCRM 模型[40]训练后的字形特征。如图 3 所示，该模型摒弃了基于 id 的字符嵌入方法，使用基于

序列字符图像的方法将每个字符渲染成 48×48 的二值灰度图像，并为其设计了双通道位置特征图，然后通过两层残差卷积神经网络生成汉字的初始字形表示，最后使用多个双向编码器作为上层结构来捕获上下文敏感信息，有效地利用了汉字的字形特征，同时减轻了未登录词的影响。本文对比了 BERT 模型和 GlyphCRM 模型训练后的特征向量，并提取了公共的中文词和对应的字形特征。在组成字形嵌入序列的过程中，公共的中文词采用 GlyphCRM 模型的字形特征作为字形嵌入，其余的字符则随机初始化一个范围为(0,1)的向量作为字形嵌入。

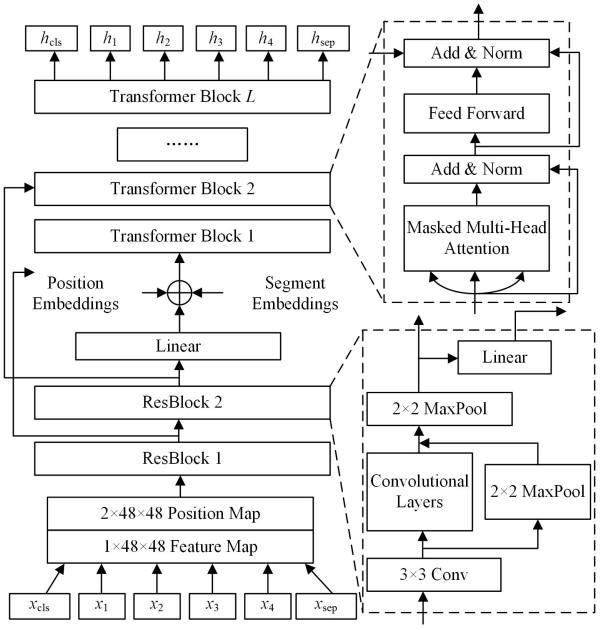

图 3　字形嵌入模块

Fig. 3　Glyph embedding module

4.1.2　部首嵌入

每个字符的部首嵌入使用传统的向量嵌入方式来增强具有相同部首的字符的语义。如图 4 所示，本文首先收集了所有中文部首并生成字符对应的部首序列。对于非中文字符，使用[NULL]进行标识，分隔符则使用[UNK]进行标识。接着，将卷积核为 3 的一维卷积层应用于部首序列，然后使用最大池化方法来获得部首嵌入特征，以确保输出维度不受输入部首序列长度的影响。其中，seq_len 代表输入序列的长度，d 代表初始嵌入的维度，ch 代表初始嵌入层经过一维卷积层后的输出通道数。最后，将输入的部首序列经过嵌入层、一维卷积层和最大池化层处理，得到一个[1，seq_len]维度的部首向量，并将其累加到其他输入中。

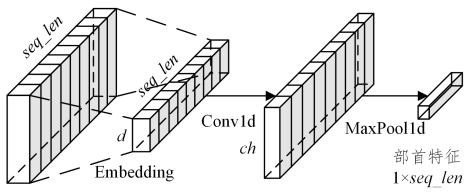

图 4　部首嵌入模块

Fig. 4　Radical embedding module

4.2　候选实体生成

交叉编码器由于对内存和计算资源的需求较大，不适合用于检索或需要快速推理的任务；而双编码器可以预先计算和缓存知识库中的实体表示，允许快速和实时的推理。因此，本文在候选实体生成阶段使用双编码器对实体指称项和知识库中的实体进行编码，并计算相似度得分。

本文使用了类似于 Humeau 等[41]描述的双编码器架构，用于建模实体指称项和实体。候选实体的生成阶段如图 5 所示，其中$char_i$，gly_i和rad_i分别代表 BERT 传统嵌入、字形嵌入和部首嵌入，⊙代表点积。输入的实体指称项上下文和候选实体分别被编码成向量 $\boldsymbol{y}_m$和 $\boldsymbol{y}_e$：

$$\boldsymbol{y}_m = red(\boldsymbol{T}_1(\boldsymbol{\tau}_m)) \tag{1}$$

$$\boldsymbol{y}_e = red(\boldsymbol{T}_2(\boldsymbol{\tau}_e)) \tag{2}$$

其中，$\boldsymbol{\tau}_m$和$\boldsymbol{\tau}_e$分别是实体指称项及其上下文和实体的输入表示，$\boldsymbol{T}_1$和$\boldsymbol{T}_2$分别是两个 Transformer 模块。$red(\cdot)$是一个函数，它将 Transformer 产生的向量序列减少为 1 个向量。根据 Humeau 等[41]的实验，本文选择的 $red(\cdot)$为[CLS]令牌输出的最后一层。

实体指称项及上下文的输入表示$\boldsymbol{\tau}_m$由围绕实体指称项的上下文词和实体指称项本身组成。具体来说，本文将每个实体指称项的输入构造为：

$$\boldsymbol{\tau}_m = [\mathrm{CLS}]ctxt_l[M_s]m[M_e]ctxt_r[\mathrm{SEP}] \tag{3}$$

其中，m，$ctxt_l$和$ctxt_r$分别是实体指称项、实体指称项左右上下文的词，$[M_s]$和$[M_e]$是标记实体指称项的特殊标记。同时，输入表示的最大长度是一个超参数。

实体的输入表示$\boldsymbol{\tau}_e$由实体标题和描述的文本组成。本文实体模型的输入为：

$$\boldsymbol{\tau}_e = [\mathrm{CLS}]title[\mathrm{ENT}]desc[\mathrm{SEP}] \tag{4}$$

其中，$title$ 和 $desc$ 是实体标题和描述的文本，[ENT]是分隔实体标题和描述的特殊标记。

然后，将推理任务简化为计算实体指称项向量和候选实体向量之间的点积，即候选实体e_i的得分 $s(m,e_i)$由点积计算得到：

$$\boldsymbol{s}(m,e_i) = \boldsymbol{y}_m \cdot \boldsymbol{y}_{e_i} \tag{5}$$

同时，该网络的训练目标是通过最优实体与同一批次(通过随机抽样选择的)实体之间的得分最大化来学习[41]。具体而言，对于批次大小为 n 的每次训练对(m_i,e_i)的损失 $L(m_i,e_i)$为：

$$\boldsymbol{L}(m_i,e_i) = -\boldsymbol{s}(m_i,e_i) + \log\sum_{j=1}^{n}\exp(\boldsymbol{s}(m_i,e_j)) \tag{6}$$

最后，将实体指称项对应的前 k 个候选实体作为该阶段的输出。

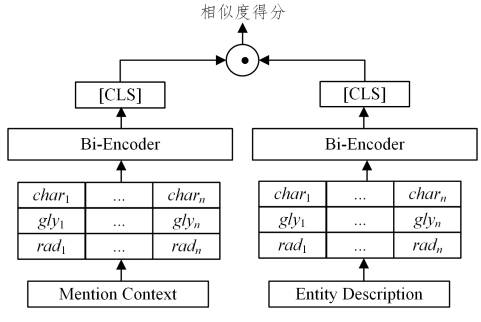

图 5　候选实体生成阶段

Fig. 5　Candidate entity generation stage

4.3 候选实体排序

本文在候选实体排序阶段使用了交叉编码器来得到候选实体的预测结果。每个样本经过双编码器生成前 100 个对应的候选实体，这些候选实体将作为交叉编码器的输入。

本文的交叉编码器类似于 Logeswaran 等[4]和 Humeau 等[41]描述的交叉编码器。候选实体的排序阶段如图 6 所示，其中 $char_i$，gly_i 和 rad_i 分别代表 BERT 传统嵌入、字形嵌入和部首嵌入，⊗代表拼接。输入是式(3)和式(4)的拼接，并从实体的输入表示中删除了[CLS]令牌，使得实体指称项和实体描述在模型训练中深度交互。具体的输入构造如下：

$$\boldsymbol{\tau}_{m,e}=[\mathrm{CLS}]ctxt_l[M_s]m[M_e]ctxt_r[\mathrm{SEP}]title[\mathrm{ENT}]desc[\mathrm{SEP}] \tag{7}$$

形式上，本文使用 $\boldsymbol{y}_{m,e}$ 来表示上下文候选嵌入向量：

$$\boldsymbol{y}_{m,e}=red(\boldsymbol{T}_{cross}(\boldsymbol{\tau}_{m,e})) \tag{8}$$

其中，$\boldsymbol{\tau}_{m,e}$ 是实体指称项和实体的输入表示，$\boldsymbol{T}_{\mathrm{cross}}$ 是一个 Transformer 模块，$red(\cdot)$ 是与 4.2 节定义相同的函数。

然后，应用一个线性层 $\boldsymbol{W}$ 来评估上下文候选嵌入向量 $\boldsymbol{y}_{m,e}$ 对候选实体的得分 $s_{\mathrm{cross}}(m,e)$：

$$s_{\mathrm{cross}}(m,e)=\boldsymbol{y}_{m,e}\boldsymbol{W} \tag{9}$$

与 4.2 节中的训练方法类似，该网络使用式(6)计算损失，以最大化最优实体的 $s_{\mathrm{cross}}(m_i,e_i)$ 来训练网络。

在 CGR-BERT-ZESHEL 模型中，前 100 个候选实体在经过交叉编码器后通过一个线性层，以计算实体指称项和候选实体之间的相似度得分。将这一组候选实体得分进行 Softmax 归一化，最终选取得分最大值对应的实体作为模型的预测输出。

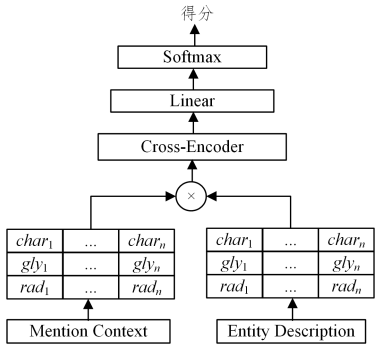

图 6 候选实体排序阶段

Fig. 6 Candidate entity ranking stage

5 实验

5.1 数据集

实验使用了 Xu 等[37]构建的 Hansel 数据集和 Zeng 等[42]构建的 CLEEK 数据集。Hansel 数据集是目前中文零样本实体链接领域广泛使用的数据集之一，包含了约 110 万条实体的知识库，用于评估实体链接模型的性能。该数据集包含训练集、验证集、少样本数据集和零样本数据集。CLEEK 数据集则是一个关于长文本的中文实体链接数据集。

本文首先使用 Hansel 的训练集和验证集对 CGR-BERT-ZESHEL 模型进行训练，然后使用 Hansel 的零样本数据集和 CLEEK 数据集评估模型的性能。

数据集的统计信息如表 1 所列，其中数据被划分为 In-KB 和 NIL 两部分。

表 1 数据集的统计信息

Table 1 Statistical information of datasets

数据集	实体指称项数量			文档数量			实体数量
	In-KB	NIL	Total	In-KB	NIL	Total	
训练集	9890000	—	9890000	1050000	—	1050000	541000
验证集	9677	—	9677	1000	—	1000	6323
零样本测试集	4208	507	4715	4200	507	4704	4046
CLEEK 测试集	2609	177	2786	100	55	100	1191

表 1 中，In-KB 指数据在知识库中的数量，NIL 指数据无法链接到知识库的数量，Total 代表总数量。Hansel 的训练集中包含 9879812 条数据样本，相对于验证集和零样本测试集来说过于庞大。因此，为了更高效地进行模型训练，本文随机采样训练集的一个子集，其中包含了 100000 条样本。在采样的同时，确保训练集与零样本测试集以及 CLEEK 测试集的交集为空集。此外，数据集中的每个实体指称项都有对应的最优实体，即模型的注意力主要集中在可以进行实体链接的部分，而忽略了包含 NIL 部分的数据。

5.2 评价指标

实体链接任务通常使用召回率(Recall)和精确度(Accuracy)作为评估指标。在模型预测的结果中：TP 为将正类预测为正类的数量；TN 为将负类预测为负类的数量；FP 为将负类预测为正类的数量；FN 为将正类预测为负类的数量。为了方便与其他模型进行比较，在候选实体生成阶段使用前 100 个候选实体的召回率作为评价指标，在候选实体排序阶段使用精确度作为评价指标。

$$Recall=\frac{TP}{TP+FN} \tag{10}$$

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{11}$$

5.3 超参数设置

使用 PyTorch 实现 CGR-BERT-ZESHEL 模型，双编码器和交叉编码器均包含 12 个编码器层，并使用 BERT 基础参数进行初始化。该模型使用 8 张 Tesla T4 显卡进行计算，每张显卡的显存为 16 GB。候选实体生成阶段的超参数参照 Gong 等[43]和 Cui 等[44]的设置，候选实体排序阶段的超参数参照 Xu 等[37]的设置，具体参数设置如表 2 所列。

表 2 模型参数设置

Table 2 Model parameter settings

阶段	参数名	取值	说明
候选实体生成	learning_rate	1×10^{-5}	学习率
	mention_length	128	实体指称项的序列长度
	entity_length	128	候选实体的序列长度
	batch_size	128	适合 GPU 的批次大小
	epoch	5	训练轮次
候选实体排序	learning_rate	1×10^{-5}	学习率
	max_length	256	序列长度
	batch_size	1	适合 GPU 的批次大小
	epoch	1	训练轮次

5.4　实验结果分析

5.4.1　对比模型介绍

本文将 CGR-BERT-ZESHEL 模型与以下基线模型进行了比较。

1)BM25 算法[32]:TF-IDF 算法的一种变体。通过计算查询到文档集合的相似度得分,来度量实体指称项和候选文档之间的相似性。

2)BERT[6]:一种采用 Transformer 架构的大规模语言模型。它包含多头自注意力机制,能够有效捕捉输入文本的上下文信息。实验中采用了其基础版本 BERT-Base 模型。

3)Chinese BERT[21]:一种基于字形和拼音信息的中文预训练模型,将字形嵌入、拼音嵌入和汉字嵌入相结合,形成一个融合嵌入,以模拟汉字独特的语义信息。

4)AT+CA[37]:Xu 等[37]在候选实体生成阶段使用的 AT 模型(别名表 Alias Table),在候选实体排序阶段使用的 CA 模型(Cross-Attention Encoder)。本文为方便描述,将两者统称为 AT+CA 模型。

5.4.2　候选实体生成阶段

实验首先在训练集上对双编码器进行训练,每个编码器使用 BERT-Base 进行初始化。为了更好地与其他模型进行比较,使用检索到的前 100 个候选实体作为排名,即使用 Recall@100 作为候选实体生成阶段的评价指标。实验结果如表 3 所列,其中 AT 模型在 CLEEK 测试集上的实验结果是通过复现 Xu 等[37]的代码得到的。从实验结果可以看出,CGR-BERT-ZESHEL 模型的性能最好,相较于 BM25,BERT,Chinese BERT 和 AT 模型,CGR-BERT-ZESHEL 模型在零样本测试集上的 Recall@100 分别提高了 27.06%,22.85%,10.52% 和 17.49%;在 CLEEK 测试集上的 Recall@100 分别提高了 30.39%,18.75%,28.06%和 7.34%。

表 3　候选实体生成阶段的实验结果

Table 3　Experimental results of candidate entity generation

(%)

模型	零样本测试集	CLEEK 测试集
BM25	69.23	57.86
BERT	73.44	69.50
Chinese BERT	85.77	60.19
AT	78.80	80.91
CGR-BERT-ZESHEL	**96.29**	**88.25**

此外,本文还对不同的 Top k 值(包括 1,4,8,16,32,50,64 和 100)在零样本和 CLEEK 测试集上进行了比较,研究了各个模型在不同的 k 值下的性能,如图 7 和图 8 所示。

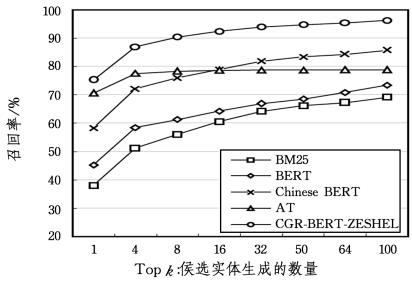

图 7　不同 Top k 在零样本测试集上的结果

Fig. 7　Results of different Top k on zero-shot

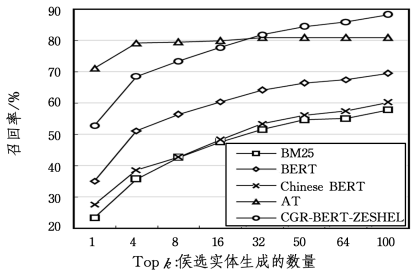

图 8　不同 Top k 在 CLEEK 测试集上的结果

Fig. 8　Results of different Top k on CLEEK

在零样本测试集上,与 BM25,BERT,Chinese BERT 和 AT 模型相比,CGR-BERT-ZESHEL 模型的 Recall 性能指标平均提高了 31.62%,27.07%,13.1%和 13.16%。

在 CLEEK 测试集上,与 BM25,BERT 和 Chinese BERT 模型相比,CGR-BERT-ZESHEL 模型的 Recall 性能指标平均提高了 30.59%,17.81%和 28.57%;与 AT 模型相比,CGR-BERT-ZESHEL 模型的 Recall 性能指标平均降低了 2.54%。但在 $k \geqslant 32$ 时,CGR-BERT-ZESHEL 模型的性能要明显优于 AT 模型。

5.4.3　候选实体排序阶段

在训练阶段,本文使用训练集中的实体指称项和双编码器检索到的前 100 个候选实体对交叉编码器进行训练。在测试阶段,使用零样本实体链接的测试集来评估交叉编码器的性能。同样地,编码器使用 BERT-Base 进行初始化,并从 100 个候选实体中选择得分最高的候选实体作为预测结果,使用精确度作为评价指标。实验结果如表 4 所列,其中 CA 模型在 CLEEK 测试集上的实验结果是通过复现 Xu 等[37]的代码得到的。

表 4　候选实体排序阶段的实验结果

Table 4　Experimental results of candidate entity ranking

(%)

模型	零样本测试集	CLEEK 测试集
BM25	38.02	23.38
BERT	72.11	61.99
Chinese BERT	69.27	59.58
CA	76.60	64.72
CGR-BERT-ZESHEL	**79.62**	**67.83**

由表 4 可知,与 BM25,BERT,Chinese BERT 和 CA 模型相比,CGR-BERT-ZESHEL 模型在零样本测试集上的精确度性能指标提高了 41.6%,7.51%,10.35% 和 3.02%;在 CLEEK 测试集上的精确度性能指标提高了 44.45%,5.84%,8.25%和 3.11%。引入了字形和部首特征后,CGR-BERT-ZESHEL 模型的性能显著提高。

5.5　消融实验

为验证本文对模型改进的有效性,分别移除字形嵌入得到 CR-BERT-ZESHEL 模型,移除部首嵌入得到 CG-BERT-ZESHEL 模型,以分析字形嵌入和部首嵌入对模型性能的影响。

5.5.1　候选实体生成阶段

实验结果如表 5 所列。在候选实体生成阶段,移除字形嵌入导致 Recall@100 在零样本和 CLEEK 测试集上分别下

降了1.99%和0.42%；移除部首嵌入导致Recall@100在零样本和CLEEK测试集上分别下降了2.99%和2.21%。由此可见，在候选实体生成阶段，这两个模块都起到了重要作用。

表5 候选实体生成阶段的消融实验结果

Table 5 Ablation results of candidate entity generation

(%)

模型	零样本测试集	CLEEK测试集
CGR-BERT-ZESHEL	**96.29**	**88.25**
CG-BERT-ZESHEL	93.30	86.04
CR-BERT-ZESHEL	94.30	87.83

另外，本文还对不同的Top k值(包括1,4,8,16,32,50,64和100)在零样本和CLEEK测试集上进行了比较，研究了两个子模块在不同的k值下的性能，如图9和图10所示。

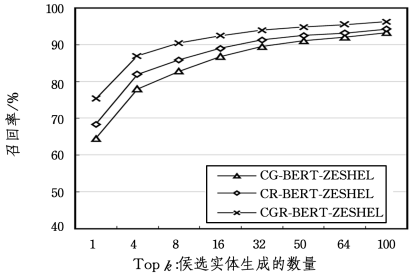

图9 不同Top k在零样本测试集上的消融实验结果

Fig. 9 Ablation results of different Top k on zero-shot

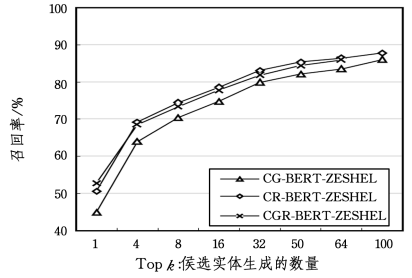

图10 不同Top k在CLEEK测试集上的消融实验结果

Fig. 10 Ablation results of different Top k on CLEEK

在零样本测试集上，与CG-BERT-ZESHEL和CR-BERT-ZESHEL模型相比，CGR-BERT-ZESHEL模型的Recall性能指标平均提高了5.96%和3.65%。

在CLEEK测试集上，CGR-BERT-ZESHEL模型与CR-BERT-ZESHEL模型性能相当；与CG-BERT-ZESHEL和CR-BERT-ZESHEL模型相比，CGR-BERT-ZESHEL模型的Recall性能指标平均提高了3.41%和0.33%。这表明在候选实体生成阶段，部首嵌入的作用略大于字形嵌入的作用，原因在于部首嵌入是由CNN卷积生成的，而字形嵌入则是预训练的结果。

5.5.2 候选实体排序阶段

实验结果如表6所列。在候选实体排序阶段，移除字形嵌入导致零样本和CLEEK测试集在Recall@100的候选实体中找到最佳候选实体的精确度分别降低了0.88%和1.27%；移除部首嵌入导致零样本和CLEEK测试集在Recall@100的候选实体中找到最优实体的精确度分别降低了6.33%和2.9%。由此可见，在候选实体排序阶段，这两个模块都起到了重要作用。

表6 候选实体排序阶段的消融实验结果

Table 6 Ablation results of candidate entity ranking

(%)

模型	零样本测试集	CLEEK测试集
CGR-BERT-ZESHEL	**79.62**	**67.83**
CG-BERT-ZESHEL	73.29	64.93
CR-BERT-ZESHEL	78.74	66.56

结束语 本文提出了一种改进的基于中文特征的零样本实体链接模型CGR-BERT-ZESHEL。首先，为了增强中文的独有特征并缓解未登录词的问题，通过从视觉图像上增加字形嵌入和从传统字符嵌入上增加部首嵌入两方面来补充词向量的更多特征。其次，将两阶段实体链接的方法应用到中文领域。在不需要任何特定于任务的启发式方法或外部实体知识的情况下，该模型在最新的中文零样本实体链接数据集Hansel和CLEEK上取得了比基线模型更好的效果。在候选实体生成阶段，Recall@100分别提高了17.49%和7.34%；在候选实体排序阶段，精确度分别提高了3.02%和3.11%。

本文未来的工作是设计一个特定专业领域的数据集来进一步验证该模型的性能，同时深入对NIL样本的研究。

参 考 文 献

[1] BUNESCU R, PASCA M. Using encyclopedic knowledge for named entity disambiguation[C]// 11th Conference of the European Chapter of the Association for Computational Linguistics. 2006: 9-16.

[2] DE C N, AZIZ W, TITOV I. Question answering by reasoning across documents with graph convolutional networks[C]// 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 2019: 2306-2317.

[3] SHEN W, WANG J Y, HAN J W. Entity linking with a knowledge base: issues, techniques, and solutions[J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 27(2): 443-460.

[4] CURRY A C, PAPAIOANNOU I, SUGLIA A, et al. Alana v2: Entertaining and informative open-domain social dialogue using ontologies and entity linking [C]// Alexa Prize Proceedings. 2018.

[5] LOGESWARAN L, CHANG M W, LEE K, et al. Zero-shot entity linking by reading entity descriptions[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3449-3460.

[6] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding [C]// Proceedings of NAACL-HLT. 2019.

[7] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [J]. arXiv: 1706.03762, 2017.

[8] YANG Z L, DAI Z H, YANG Y M, et al. Xlnet: generalized autoregressive pretraining for language understanding [C]// Pro-

ceedings of the 33rd International Conference on Neural Information Processing Systems. 2019：5753-5763.

[9] CLARK K，LUONG M T，LE Q V，et al. ELECTRA：pre-training text encoders as discriminators rather than generators[C]// International Conference on Learning Representations. 2019.

[10] LAN Z Z，CHEN M D，GOODMAN S，et al. ALBERT：a lite BERT for self-supervised learning of language representations [C]// International Conference on Learning Representations. 2019.

[11] LIU Y H，OTT M，GOYAL N，et al. Roberta：a robustly optimized bert pretraining approach [J]. arXiv：1907. 11692，2019.

[12] RADFORD A，WU J，CHILD R，et al. Language models are unsupervised multitask learner-s [J]. OpenAI blog，2019，1(8)：9.

[13] BROWN T，MANN B，RYDER N，et al. Language models are few-shot learners [J]. Advances in Neural Information Processing Systems，2020，33：1877-1901.

[14] RAE J W，BORGEAUD S，CAI T，et al. Scaling language models：methods，analysis & insights from training gopher [J]. arXiv：2112. 11446，2021.

[15] WANG B. Mesh-Transformer-JAX：model-parallel implementation of transformer language model with JAX [EB/OL]. https：//github. com/kingoflolz/mesh-transformer-jax 2021.

[16] LEWIS M，LIU Y H，GOYAL N，et al. BART：denoising sequence-to-sequence pre-training for natural language generation，translation，and comprehension [C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020：7871-7880.

[17] BAO H B，DONG L，WEI F R，et al. Unilmv2：pseudo-masked language models for unified language model pre-training[C]// International Conference on Machine Learning. PMLR，2020：642-652.

[18] ZHU J H，XIA Y C，WU L J，et al. Incorporating BERT into neural machine translation [C]// International Conference on Learning Representations. 2019.

[19] LI X Y，MENG Y X，SUN X F，et al. Is word segmentation necessary for deep learning of chinese representations? [C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019：3242-3252.

[20] SUN Y，WANG S H，LI Y K，et al. Ernie：enhanced representation through knowledge integration [J]. arXiv：1904. 09223，2019.

[21] CUI Y M，CHE W X，LIU T，et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ ACM Transactions on Audio，Speech，and Language Processing，2021，29：3504-3514.

[22] SUN Z J，LI X Y，SUN X F，et al. ChineseBERT：chinese pretraining enhanced by glyph and pinyin information[C]// Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1：Long Papers). 2021：2065-2075.

[23] OUYANG L，WU J，JIANG X，et al. Training language models to follow instructions with human feedback [J]. Advances in Neural Information Processing Systems，2022，35：27730-27744.

[24] ABDULLAH M，MADAIN A，JARARWEH Y. ChatGPT：fundamentals，applications and social impacts[C]// 2022 Ninth International Conference on Social Networks Analysis，Management and Security(SNAMS). IEEE，2022：1-8.

[25] HE Z Y，LIU S J，LI M，et al. Learning entity representation for entity disambiguation [C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics(Volume 2：Short Papers). 2013：30-34.

[26] CHEN Y，TAN Y S，WU Q B，et al. TGCEL：a chinese entity linking method based on topic relation graph[C]// 2017 6th International Conference on Computer Science and Network Technology(ICCSN T). IEEE，2017：226-230.

[27] OUYANG X Y，CHEN S D，ZHAO H，et al. A multi-cross matching network for chinese named entity linking in short text [C]// Journal of Physics：Conference Series. IOP Publishing，2019.

[28] HUA X Y，LI L，HUA L F，et al. XREF：entity linking for chinese news comments with supplementary article reference[C]// Automated Knowledge Base Construction. 2020.

[29] MURTY S，VERGA P，VILNIS L，et al. Hierarchical losses and new resources for fine-grained entity typing and linking[C]// Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics(Volume 1：Long Papers). 2018：97-109.

[30] RADFORD A，NARASIMHAN K，SALIMA-NS T，et al. Improving language understanding with unsupervised learning[J]. Citado，2018，17：1-12.

[31] YAMADA I，SHINDO H，TAKEDA H，et al. Joint learning of the embedding of words and entities for named entity disambiguation[C]// 20th SIGNLL Conference on Computational Natural Language Learning，CoNLL 2016. Association for Computational Linguistics(ACL)，2016：250-259.

[32] ROBERTSON S，ZARAGOZA H. The probabilistic relevance framework：BM25 and beyond[J]. Foundations and Trends © in Information Retrieval，2009，3(4)：333-389.

[33] WIATRAK M，ARVANITI E，BRAYNE A，et al. Proxy-based zero-shot entity linking by effective candidate retrieval[C]// Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis(LOUHI). 2022：87-99.

[34] XU Z R，CHEN Y L，SHI S B，et al. Enhancing entity linking with contextualized entity embeddings [C]// CCF International Conference on Natural Language Processing and Chinese Computing. Cham：Springer Nature Switzerland，2022：228-239.

[35] SEVGILI Ö，SHELMANOV A，ARKHIPOV M，et al. Neural entity linking：a survey of models based on deep learning [J]. Semantic Web，2022，13(3)：527-570.

[36] RISTOSKI P，LIN Z Z，ZHOU Q Z. KG-ZESHEL：knowledge graph-enhanced zero-shot entity linking[C]// Proceedings of the 11th on Knowledge Capture Conference. 2021：49-56.

[37] XU Z R, SHAN Z F, LI Y X, et al. Hansel: a chinese few-shot and zero-shot entity linking benchmark[C]// Proceedings of the Sixteenth ACM International Conference on Web Search and Data Mining. 2023: 832-840.

[38] HUANG S J, WANG B B, QIN L B, et al. Improving few-shot and zero-shot entity linking with coarse-to-fine lexicon-based retriever[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland, 2023: 245-256.

[39] ZHOU H Y, SUN C J, LIN L, et al. ERNIE-AT-CEL: a chinese few-shot emerging entity linking model based on ERNIE and adversarial training[C]// CCF International Conference on Natural Language Processing and Chinese Computing. Cham: Springer Nature Switzerland, 2023: 48-56.

[40] LI Y X, ZHAO Y, HU B T, et al. Glyphcrm: bidirectional encoder representation for chinese character with its glyph [J]. arXiv: 2107. 00395, 2021.

[41] HUMEAU S, SHUSTER K, LACHAUX M A, et al. Poly-encoders: architectures and pre-training strategies for fast and accurate multi-sentence scoring[C]// International Conference on Learning Representations. 2019.

[42] ZENG W X, ZHAO X, TANG J Y, et al. Cleek: a chinese long-text corpus for entity linking[C]// Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 2026-2035.

[43] GONG S, XIONG X, LI S, et al. Chinese entity linking with two-stage pre-training transformer encoders[C]// 2022 International Conference on Machine Learning and Knowledge Engineering (MLKE). IEEE, 2022: 288-293.

[44] CUI Y, CHE W, LIU T, et al. Pre-training with whole word masking for chinese bert[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504-3514.

PAN Jian, born in 1976, Ph.D, associate professor, postgraduate supervisor, is a member of CCF (No. 26947M). His main research interests include natural language processing, intelligent information processing and Internet of Things.

(责任编辑: 柯颖)