Computer Science and Application 计算机科学与应用, 2022, 12(4), 1202-1212
Published Online April 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.124122

Hans汉斯

基于知识库标记预训练孪生神经网络的中文实体链接

何展鹏

广东工业大学，计算机学院，广东 广州

收稿日期：2022年3月21日；录用日期：2022年4月22日；发布日期：2022年4月29日

摘　要

知识库问答(Knowledge Base Question Answering)有两个子任务：实体链接和关系预测。实体链接任务(Entity linking)是对于一个给定的文本，识别出其中实体指代(Mention)，并关联到知识库对应话题实体的过程。实体链接包括实体指代识别(MD)和实体消歧(ED)两个子任务。由于存在口语化严重、短文本信息少、实体多歧义多等问题，在中文数据集中更具挑战。传统方法并未充分利用知识库，缺少对短文本中指代和知识库实体的表示深入探究，本文提出用加入知识库标记的BERT模型进行实体识别，加入知识库标记及实体子图的BERT-SiameseFNN网络得到指代和候选实体的语义表示，进行实体消歧。通过在多个中文数据集上验证，表明该方法得到更充分利用知识库，并得到更好的匹配表示，有效提升实体链接性能。

关键词

知识库问答，实体链接，预训练语言模型

Knowledge Marker-Based Pre-Trained Language Model with Siamese Network for Chinese Entity Linking

Zhanpeng He

Computer College, Guangdong University of Technology, Guangzhou Guangdong

Received: Mar. 21st, 2022; accepted: Apr. 22nd, 2022; published: Apr. 29th, 2022

Abstract

The knowledge base question answering (KBQA) task has two sub-tasks: entity linking and rela-

文章引用: 何展鹏. 基于知识库标记预训练孪生神经网络的中文实体链接[J]. 计算机科学与应用, 2022, 12(4): 1202-1212. DOI: 10.12677/csa.2022.124122

tion detection. Entity linking (EL) is the process of linking entity mentions appearing in given text with their corresponding topic entities in a knowledge base, which includes mention detection and entity disambiguation. Due to expression diversity, short context information, different meaning between similar candidate entities, it is more challenging in Chinese Entity Linking. Traditional methods do not make full use of the knowledge base and lack further exploration of representation between mention and candidate entity. We propose a knowledge marker method for both mention detection and entity disambiguation. In entity disambiguation, we also use a BERT-Siamese FNN network to encode mention-candidate entity pairs. The experimental results on two datasets show that the EKBERT reaches the state-of-the-art models and distills rich but discriminative information.

Keywords

Knowledge Base Question Answering, Entity Linking, Pre-Trained Language Model

1. 引言

知识库问答分为两步，实体链接和关系预测。实体链接负责从问句中识别出指代并映射到知识库对应的话题实体；关系预测用于理解问句意图，预测问句问及实体的关系路径。最终通过实体及关系路径映射到知识库的一个三元组，作为问句答案。

面向中文短文本的实体链接(Entity Linking)，是自然语言处理领域的基础任务之一，即对于给定的一个中文短文本(如问答中的问句、微博短文等)识别出其中的指代，并与给定知识库中的话题实体进行关联的过程，包括实体指代识别和实体消歧两个子任务[1]。在 NLP 社区的研究中，实体链接经常被忽视。国外学者[2] [3] [4]认为通过问句中的实体指代和知识库实体在字面上的相似度匹配就能解决。本文分析中文数据集[5]发现：1) 未见实体指代多，占测试集的 85%。2) 95%的实体指代对应的知识库候选实体都存在相同的重叠词(overlap)。模型要在短文本提供极少信息下挖掘字面相似的候选实体之间的差别。3) 缺少知识库的背景知识，人类也无法给出指代映射的正确实体。例如短文本“2013 年 12 月永宁站的日进出人次是多少排名第几？”，指代“永宁站”是一个未见实体，仅出现在测试集中。指代对应的知识库候选实体是{永宁站，永宁站(新北市)，永宁站(沈吉铁路)}。均存在相同重叠词“永宁站”，从上下文信息中，只能推断问句问及一个车站的信息，但无法分辨出话题实体。

针对上述问题，已有一定的研究成果。工作[6]采用预训练模型 BERT [7]以及流水线的方式，把实体识别建模为序列批注识别出实体指代，然后将其与候选实体作为实体消歧的输入，建模为句对匹配，输出候选实体的分值。采用预训练模型一定程度上解决了未见指代首次登陆的问题，但容易识别出知识库不存在实体的错误指代。同时该工作并未利用到知识库信息。工作[8]在工作[6]基础上沿用了预训练模型 BERT 以及流水线的方式，在实体消歧上把实体描述文本拼接到候选实体后，引入知识库信息提升实体消歧的表现，在数据集[9]中获得最佳表现。但在句对表示上并未深入研究，仅简单取指代的首尾 token 作为匹配的最终表示。在实体识别中结合了 BERT、GRU 和 CNN，参数量大且结构较为复杂。

为解决在中文实体链接上的问题，本文在目前已有研究的基础上，对基于 BERT 的中文实体链接框架提出了一个优化模型 Entity Knowledge subgraph-BERT (EKBERT)。主要贡献如下：

1) 提出一种统一的知识标记方法，同时应用到实体链接的实体识别和实体消歧任务。

2) 在实体消歧任务上，引入候选实体的一跳子图增强上下文信息，结合知识标记提取指代和候选实体的片段表示，传入孪生网络得到更好的句子对表示以及候选实体得分。

3) 在公开的中文实体链接和中文知识库问答任务上进行实验，达到了目前公开的最佳效果。

2. 相关工作

2.1. 中文知识库问答

知识库由大量三元组[<主语>，<谓语>，<宾语>]构成。知识库问答的目标是把一个自然语言的问句转化成一个查询语句，通过检索知识库得到答案。目前主要分为基于语义解析和基于信息检索两种方法[10]，基于语义解析的方法直接从自然语言问句中解析出实体、关系及逻辑组合，转化为知识库上的查询语句并从知识库查询返回答案。方法[11]利用序列标注模型解析问句中的实体、利用端到端模型解析问句中的关系序列。基于语义解析的方法通常依赖大量人力进行关系分类的标注，难以预测训练集中未出现的关系。基于信息检索的方法在识别指代并链接实体的基础上，从知识库中召回候选实体的关系路径，并与问句进行语义匹配的排序，选择出最可能的路径从知识库中检索答案。方法[2]提出增强路径匹配的方法，实现问句与候选路径的多层次匹配。相比于基于语义解析的方法，基于信息检索的方法在路径选择方面具有更好的泛化能力，能够应用在较大的知识库中。

2.2. 基于预训练语言模型的实体链接

深度学习的实体链接方法[12]中，一般分为指代识别和实体消歧两步。在指代识别上，方法[13]等人使用了一个别名词典和 LSTM 语言模型进行指代识别，方法[7] [8] [14] [15]用了基于 BERT 语言模型做指代识别，方法[14] [15]基于知识库特点构建了领域词典，基于精确匹配辅助深度学习模型提高了整体的泛化能力。但未实现利用知识库融合到端到端的实体识别模型。方法[16]采用了预训练模型并引入了知识库实体信息。通过拼接知识库实体及其描述信息，输入 BERT 得到实体嵌入表示，然后与双向 GRU 网络的提取的语义表示进行拼接，输入一维卷积网络，最终输出每个 token 的二分类结果进行指代识别。但网络结构较为复杂。方法[15]在方法[17]提出的模型 FLAT 上，把知识库实体作为词典融入 lattice 特征，再基于 BERT 做实体识别，上述工作都利用知识库进行端到端的中文实体识别。

在实体消歧上，一般建模为文本匹配的排序问题。方法[4]提出一个引入 Lattice 结构的 CNN 网络，从词和字层面上编码，解决中文文本无法避免的分词错误从而得到更好的文本匹配表示。但未利用预训练语言模型，模型需要标注数据较多。方法[7]采用预训练语言模型 BERT 建模问句和候选实体的匹配任务，提高泛化能力。方法[14]在利用 BERT 得到匹配得分上，再规定了实体长度，实体出度，实体与提问词的距离等特征，构建了一个基于特征的评分机制，但超参数设置难控制。方法[8]结合 BERT 的匹配得分和基于多个特征通过 LightGBM 得到另一个匹配得分，相加得到最终得分。但仍依赖人为规定，且未利用知识库的信息。方法[18]等人提出一个基于 BERT 的孪生神经网络匹配模型，节省了参数及推理时间，并获得更好的文本匹配表示，证实了孪生网络在句子表示中的作用。但并未在短文本实体链接任务上验证。

2.3. 融入知识的预训练语言模型

最近，很多工作研究把知识融入到预训练语言模型中。方法[19]先把文本变成知识增强的句子树，然后通过软位置编码和掩盖注意力机制把知识融入到语言模型中，不影响原来学习好的参数解决知识噪声问题。在特定领域的中文序列批注和中文长文本匹配上得到验证，但未在中文实体链接任务上实验。方

法[20]等人利用维基百科额外的实体描述信息作为上下文信息，从而提高需要推理的问答任务的表现。方法[21]等人利用 BERT 以及实体信息解决实体关系分类问题。这些工作都表明注入知识到语言模型中的有效性，能提升下游任务的表现，

3. 本文方法

3.1. EKBERT 整体结构

EKBERT 模型的整体结构如图 1 所示。本实体链接方法由两个部分组成，指代识别模型和实体消歧模型。给定一个问句 Q，我们将它与知识库的实体进行精确匹配，在问句中匹配的词前后插入特殊标记“|”，然后输入指代识别模型，模型输出问句的实体指代 m。通过数据集提供的 mention2id 字典得到指代对应的所有候选实体 $E=\{e_1,\cdots,e_T\}$，把问句 Q 和逐个候选实体 e_i 组成句子对。同时，从知识库中检索候选实体 e_i 的一跳子图，作为上下文信息拼接在候选实体 e_i 后。最后，在指代和候选实体的前后插入特殊标记“#”和“$”标出位置。输入到实体消歧模型。该模块直接输出候选实体的得分 p，分值最高的候选实体 e_i 作为问句指代映射到知识库的话题实体，完成实体链接。

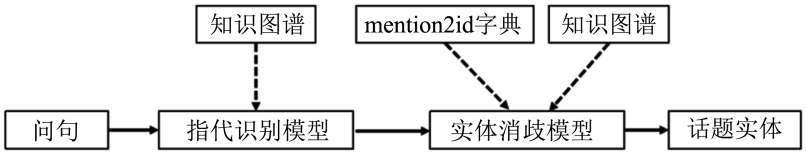

Figure 1. The overall architecture of EKBERT
图 1. EKBERT 实体链接的整体架构

3.2. 实体指代识别

EKBERT 的指代识别模型基于 BERT-Softmax 模型，输入问句 Q，输出指代 m，如图 2 所示。为了召回更多知识库中的实体同时提高指代识别的准确率，我们希望进一步挖掘知识库中的有用信息。知识库中的实体往往隐藏指代，我们将问句与知识库的实体进行精确匹配。这里会涉及 ner 任务的嵌套实体问题。比如知识库中存在实体“永宁”和“永宁站”。“进出”和“进出人次”。我们会在这些词的前后插入标记“|”，如果同一个字前(或后)有多个词边界，只会插入一次特殊标记“|”。比如问句“2013 年 12 月永宁站的日进出人次是多少排名第几?”与知识库精确匹配的实体是{2 月，永宁，永宁站，进出，进出人次}，匹配后的输入文本为“2013 年 1|2 月|永宁|站|的日|进出|人次|是多少排名第几？”。“永”字前只会插入一次“|”。标注方法采用 BIO，引入的标记“|”会根据是否在指代 span 内确定，在指代 span 以外标签为“O”，在指代 span 以内标签为“I”。另外，涉及书名号的命名实体，把书名号统一标记到命名实体的 span 中。

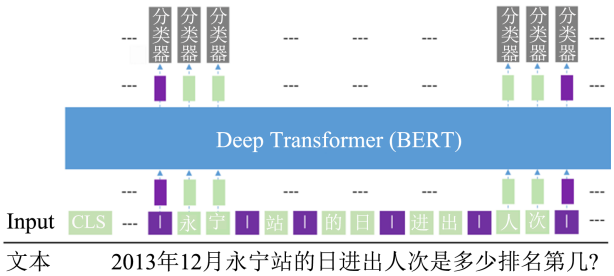

Figure 2. Mention detection model of EKBERT
图 2. EKBERT 的指代识别模型

3.3. 实体消歧

EKBERT 的实体消歧模型如图 3 所示，编码层是 BERT，对输入的句子对进行编码。中间层是一个孪生的前馈全连接神经网络，得到指代 m 和候选实体 e_i 匹配表示，输出层用 Softmax 分类器给出候选实体 e_i 的分值。

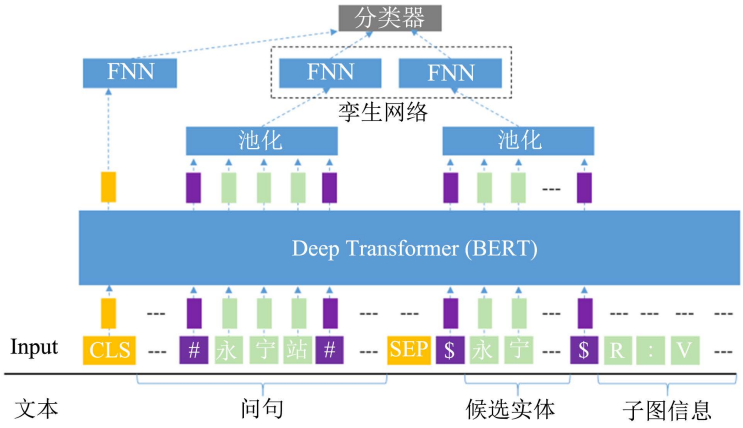

Figure 3. Entity disambiguation model of EKBERT
图 3. EKBERT 的实体消歧模型

3.3.1. 知识库信息增强

知识库信息：给定一个指代 m，通过 mention2id 字典我们得到候选实体集合 $E=\{e_1,\cdots,e_T\}$。对于该集合中的每一个候选实体 e_i，我们从提供的知识库中检索得到若干个三元组，把带指代 m 的问句 Q 作为句子 1，候选实体 e_i 和它的一跳子图拼接作为句子 2。其中，一跳子图由候选实体(头实体)直接相连的多组“关系 r：尾实体 o”(或“属性：属性值”)组成，不同的“关系：尾实体”(或“属性：属性值”)用“|”分隔开。

实体边界信息：我们希望模型能捕捉到指代和候选实体的边界，因此，在指代词前后添加标记“#”，在候选实体前后添加标记“$”。最终，我们把问句，候选实体，知识库子图按公式(1)输入 BERT 的编码器：

$$[\mathrm{CLS}]\mathrm{Q}_{\mathrm{left}}\#\mathrm{m}\#\mathrm{Q}_{\mathrm{right}}[\mathrm{SEP}]\$e_i\$r_1:o_1,\cdots,r_J:o_J \tag{1}$$

3.3.2. 池化策略及孪生神经网络

假设每个 token 的隐层表示是 $H_t\in R^d$。向量 H_i 到 H_j 表示指代 m 的隐层表示，向量 H_k 到 H_m 表示候选实体 e_i 的隐层表示。我们通过池化(平均或最大)和激活函数 tanh 得到指代 m 和候选实体 e_i 的一维向量表示。由于孪生神经网络(Siamese Transformer [18]、Siamese BiLSTM [22] [23]、Siamese CNN [24])在句子对匹配任务得到充分的验证，其结构能挖掘句子对中的深层差异，得到更好的句子嵌入。本文采用 Siamese FNN 作为中间层，经过 BERT 得到的指代和候选实体表示已经包含丰富的交互信息，因此用结构简单的 FNN ($W_1\in R^{d\times d}, b_1\in R^d$)做孪生网络的编码层。对指代 m 的 token 向量 H_i 到 H_j 进行平均池化后，得到它的最终表示 H'_m。过程公式化为：

$$H'_m=W_1\left[\tanh\left(\frac{1}{j-i+1}\sum_{t=i}^{j}H_t\right)\right]+b_1 \tag{2}$$

对候选实体 e_i 的 token 向量 H_k 到 H_m 进行平均池化后，它的最终表示为 H'_e。过程公式化为：

$$H'_e = W_1\left[\tanh\left(\frac{1}{m-k+1}\sum_{t=k}^{m} H_t\right)\right] + b_1 \tag{3}$$

[CLS]位置的 token 向量作为句子表示，我们通过另一个 FNN ($W_2 \in R^{d\times d}, b_2 \in R^d$)得到其最终表示 H'_{cls}，过程公式化为：

$$H'_{cls} = W_2\left(\tanh\left(H_{cls}\right)\right) + b_2 \tag{4}$$

最后，把这 4 个向量 $H'_{cls}, H'_m, H'_e, \left|H'_m - H'_e\right|$ 拼接，输入到 Softmax 分类器，得到每个候选实体的二分类概率。标签 1 的输出对应得分，得分越高越可能是话题实体。

在训练过程中，我们使用交叉熵作为损失函数并在每个全连接层之前应用了 dropout，防止模型过拟合。

4. 实验

4.1. 数据集和评估指标

我们在两个中文数据集中进行了实验。第一个是 NLPCC 2016 CKBQA 知识库问答数据集[5]。数据集提供了：1) 14,609 个问答对的训练集和 9870 个问答对的测试集；2) 一个接近 4300 万条 SPO 的知识库；3) 一个叫 mention2id 的指代词与候选实体的映射字典。第二个是 CCKS 2019 EL 实体链接数据集[25]。数据集提供：1) 一个约 365 万条 SPO 的知识库；2) 9 万条训练集，1 万条验证集，3 万条测试集；3) mention2id 字典。

原始数据集提供的数据格式如表 1，我们沿用工作[7]的方法，调整成适合指代识别，实体消歧任务的格式。构建出来的形式如表 2、表 3 所示。对于指代识别模型，我们采用精确率 Precision、召回率 Recall 和 F1 值进行评估。对于实体消歧模型，我们采用 Top 1、Top 2、Top 3 的准确率和平均倒排数 MRR 进行评估。对于整个实体链接任务，我们采用 Top 1、Top 2、Top 3 的准确率进行评估。

Table 1. An example in NLPCC 2016 CKBQA dataset
表 1. 原始数据集 NLPCC 2016 CKBQA 的案例格式

question	2013 年 12 月永宁站的日进出人次是多少排名第几？
triple	永宁站(新北市)|日进出人次|29,806 [1]，第 49 名(2013 年 12 月)
answer	29,806 [1]，第 49 名(2013 年 12 月)

Table 2. An example in mention detection
表 2. 指代识别的案例格式

tokens	2013 年 12 月永宁站的日进出人次是多少排名第几？
labels	O O O O B-entity I-entity I-entity O O O O O O O O O O O O O O

Table 3. An example in entity disambiguation
表 3. 实体消歧的案例格式

sentence A	2013 年 12 月永宁站的日进出人次是多少排名第几？
sentence B	永宁站(新北市)
label	1

4.2. 实验设置

4.2.1. 数据预处理

由于数据集中的有大量指代带有书名号，构建指代识别数据集时，我们统一把书名号标注到命名实体中，否则书名号《》的标注不统一会影响指代识别模型的表现。

4.2.2. 超参数设置

表 4 展示了实验使用的超参数。第一组参数用于指代识别，和工作[7]保持一致，第二组用于实体消歧。其中最大序列长度 Maxlength 由拼接了知识库子图信息的上下文平均长度决定，最大序列长度设置到 380 能覆盖数据集 95%的案例。不注入知识库信息，设置到 60 则能覆盖数据集 95%的案例。

Table 4. Parameter settings
表 4. 模型用到的超参数设置

参数	指代识别	实体消歧
Batch size	32	24
Epochs	30	1
Max length	60	380
Learning rate	1e−5	1e−5
Weight decay	1e−2	1e−2
Warmup steps	1000	1000
Dropout rate	0.1	0.1

4.3. 实验效果及分析

实验的基准模型有两类：1) 数据集的 SOTA 模型：BB-KBQA [7]和方法[8]；2) K-BERT [19]，该模型在预训练模型 BERT 中巧妙融入了知识库，并在 NER 任务 MSRA-NER、NLPCC-DBQA 和匹配任务 LCQMC、NLPCC-DBQA 中表现最好。

4.3.1. 指代识别模块效果

指代识别模型的表现如表 5 所示。NLPCC CKBQA 数据集中，BB-KBQA [7]用的是 BERT-CRF 模型，没有利用知识库，表现最低。方法[8]利用 BERT 直接推理得到实体描述表示，作为知识增强表示，序列标注过程先用 BERT-BiLSTM 编码，然后每个 token 拼接前述的实体描述表示，用一维卷积作为解码层，网络结构复杂且参数量大。K-BERT [19]融入了知识库信息，但由于在句子中实体后插入大量实体描述信息，导致知识噪声问题。本方法提出的 EKBERT-MD (mention detection)在原句子中，匹配知识库的实体前后插入知识标记“|”，不影响语义并利用知识库实体的边界信息，实验结果在两个数据集上的表现都有提升。

Table 5. Mention detection results (%)
表 5. 指代识别的表现(%)

模型	NLPCC2016 CKBQA			CCKS 2019EL		
	P	R	F1	P	R	F1
State-of-the-Art						
BB-KBQA [7]	96.96	97.02	96.99	-	-	-
方法[8]	-	-	-	82.68	85.34	83.98

Continued

知识增强的预训练模型						
K-BERT [19]	97.21	97.6	97.4	81.26	84.32	82.76
EKBERT						
EKBERT-MD	**98.87**	**98.87**	**98.87**	**86.09**	**88.07**	**87.06**

4.3.2. 实体消歧模块效果

实体消歧模型的表现如表 6 所示。融入知识库信息的模型比没有利用知识库的 BB-KBQA [7]有明显提升，K-BERT [19]在句子中实体后直接插入了子图信息，当插入信息远大于文本长度，会影响整个语句的语义表示，因此提升有限。方法[8]将注入了实体描述信息，并取指代前后位置的 token 作为匹配表示，是数据集中公开的第一名。我们提出的 EKBERT-ED (entity disambiguation)注入了实体描述信息和指代、实体标记，取标记范围内的 token 池化(max 表示最大池化，ave 表示平均池化)，通过 Siamese FNN 编码输出得分。表现优于 K-BERT，并接近方法[8]。

Table 6. Entity disambiguation results (%)
表 6. 实体消歧的表现(%)

模型	NLPCC 2016CKBQA				CCKS 2019EL		
	Top 1	Top 2	Top 3	MRR	Top 1	Top 2	Top 3
State-of-the-Art							
BB-KBQA [7]	89.14	93.19	95.05	92.16	-	-	-
方法[8]	-	-	-	-	**93.93**	99.09	**99.86**
知识增强的预训练模型							
K-BERT [19]	89.58	92.58	94.41	91.46	93.38	98.16	99.21
EKBERT							
EKBERT-ED-max	**93.39**	**96.98**	**98.4**	**95.94**	93.83	**99.14**	99.82
EKBERT-ED-ave	93.28	96.79	98.36	95.87	93.76	99.02	99.79

4.3.3. 实体链接效果

联合指代识别和实体消歧模型，整体的实体链接结果如表 7 所示，在两个数据集上的表现都比现有方法有所提升。

Table 7. Entity linking results (%)
表 7. 实体链接的表现(%)

模型	NLPCC 2016CKBQA			CCKS 2019EL		
	Top 1	Top 2	Top 3	Top 1	Top 2	Top 3
State-of-the-Art						
BB-KBQA [7]	86.27	90.18	91.97	-	-	-
方法[8]	-	-	-	80.16	84.56	85.2

Continued

知识增强的预训练模型						
K-BERT [19]	86.52	89.83	91.72	78.73	82.76	83.48
EKBERT						
EKBERT	**92.92**	**96.31**	**97.56**	**80.32**	**85.03**	**85.78**

4.3.4. 知识库问答

为了证明本文的实体链接模型能为整个 KBQA 任务带来提升，我们进行了完整实验。我们采用与工作[7]相同的关系预测模型和检索答案的计分公式，只替换实体链接模型。关系预测模型给出候选关系得分 S^T，实体链接模型给出候选实体得分 S^e，通过加权和得到最终得分。我们复现的关系预测模型 accuracy 表现是 Top 1 (94.75%)，Top 2 (97.99%)，Top 3 (98.82%)，贴近公开数据(表 8)。最终，联合我们提出的实体链接模型。在 CKBQA 任务[5]中 F1 值上为 87.34%，比工作[7]提升了 3%。

Table 8. Comparison of public results and our reproduced results (%)
表 8. 公开数据和复现数据对比(%)

模型	Top 1	Top 2	Top 3	F1
BB-KBQA [7]公开数据	94.81	97.68	98.60	84.12
本文复现数据	94.75	97.99	98.82	84.18

4.4. 消融实验

为了理解除 BERT 以外的其余组件给实体消歧带来的提升，我们在数据集[5]设置了消融实验。我们把除去实体边界信息的实体消歧模型标为 EKBERT-ED-w/o etoken，把除去实体边界信息和知识库信息的实体消歧模型标为 EKBERT-ED-w/o etoken + kg。消融实验结果如表 9 所示，从趋势发现，去掉更多组件，表现逐步变差。注入实体的知识库信息和实体边界信息的方法分别带来 2.18%和 2.08%的提升。

Table 9. Comparison of our model with different components (%)
表 9. 不同组件的模型对比(%)

模型	Top 1	Top 2	Top 3	MRR
EKBERT-ED	**93.39**	**96.98**	**98.40**	**95.94**
EKBERT-ED-w/o etoken	91.31	95.49	97.02	94.56
EKBERT-ED-w/o etoken + kg	89.13	93.15	94.99	92.51

4.5. 案例分析

进一步分析实验结果，我们给出几个有代表性的案例分析，如表 10 所示。

Table 10. Representative examples in experiment results
表 10. 具有代表性的实验案例分析

id	问题	BB-KBQA	Our	分析
1	动物地鸠属是属于什么目呀？	×	√	提出的指代识别模型预测成功

Continued

2	你知道赵文卓甄子丹事件都有谁吗？	×	√	提出的指代识别模型预测成功
3	告诉我高等数学的出版时间是什么时候？	×	√	提出的实体消歧模型预测成功
4	苹果的拉丁学名是什么呀？	×	√	提出的实体消歧模型预测成功

案例 1 至 3 展示指代识别模型的效果。案例 1 和案例 2 都涉及嵌套实体问题，“动物地鸠属”由命名实体“动物”“地鸠属”组成，“赵文卓甄子丹事件”由“赵文卓”，“甄子丹”，“事件”组成。而目标指代分别是“地鸠属”和“赵文卓甄子丹事件”。我们引入的特殊 token“|”的方法让模型识别出正确的指代。方法[5]未识别出正确指代。案例 4 和 5 展示实体消歧模块的能力，案例 4 中，本文和方法[5]均接收正确指代“高等数学”，本文模型从 18 个都具有重叠词“高等数学”的候选实体中给出正确的实体“高等数学”，方法[5]预测的是“高等数学(北大版高等数学)”。其中 15 个候选实体都具有“出版时间”关系。说明本文模型能分辨出与问句有相同重叠词(高等数学、出版时间)的候选实体之间的差别。案例 5，两个模型均接收正确指代“苹果”，问句涉及到背景知识“拉丁学名”，本文方法从 22 个候选实体中预测出话题实体“苹果”，方法[5]预测错误实体“苹果(蔷薇科苹果属果树)”。得益于问句和知识库信息，模型能识别出候选实体中 4 个包含关系“拉丁学名”的实体，同时得益于引入实体边界的池化和孪生网络，使模型能进一步区别这 4 个候选实体。

5. 结束语

本文提出了一个基于知识标记的预训练孪生神经网络实体链接模型，提出一种知识标记方法，在指代识别中融合实体边界信息。在实体消歧中融合了实体边界和知识库信息，通过孪生网络得到更好的匹配表示。实验证明，EKBERT 是一个能有效解决未见实体指代和利用知识库区分相似候选实体的实体链接方法，并兼容流水线的知识库问答框架。在 NLPCC 2016 CKBQA 和 CCKS 2019 EL 数据集上达到并超过最好的方法。在未来，我们计划改进模型的规模和参数规模，并在更多中文数据集上评估，经一步提高模型的性能。

基金项目

广东省自然科学基金资助项目(2021A1515012556)。

参考文献

[1] 韩先培, 等. CCKS2019 知识库评测技术报告: 实体、关系、事件及问答[EB/OL]. 中文信息学报. https://arxiv.org/pdf/2003.03875, 2017.

[2] Yu, M., Yin, W., Hasan, K.S., *et al*. (2017) Improved Neural Relation Detection for Knowledge Base Question Answering. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 571-581. https://doi.org/10.18653/v1/P17-1053

[3] Wu, P., Huang, S., Weng, R., *et al*. (2019) Learning Representation Mapping for Relation Detection in Knowledge Base Question Answering. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 6130-6139. https://doi.org/10.18653/v1/P19-1616

[4] Lai, Y., Feng, Y., Yu, X., *et al*. (2019) Lattice CNNS for Matching Based Chinese Question Answering. *Proceedings of the AAAI Conference on Artificial Intelligence*, **33**, 6634-6641. https://doi.org/10.1609/aaai.v33i01.33016634

[5] Duan, N. (2016) Overview of the NLPCC-ICCPOL 2016 Shared Task: Open Domain Chinese Question Answering. In: *Natural Language Understanding and Intelligent Applications*, Springer, Cham, 942-948. https://doi.org/10.1007/978-3-319-50496-4_89

[6] Liu, A., Huang, Z., Lu, H., *et al*. (2019) BB-KBQA: BERT-Based Knowledge Base Question Answering. In: *China National Conference on Chinese Computational Linguistics*, Springer, Cham, 81-92.

https://doi.org/10.1007/978-3-030-32381-3_7

[7] Devlin, J., Chang, M.W., Lee, K., *et al*. (2019) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of NAACL-HLT* 2019, Minneapolis, 2-7 June 2019, 4171-4186.

[8] 汪洲, 侯依宁, 等. 基于特征融合的中文知识库问答方法[EB/OL]. https://bj.bcebos.com/v1/conference/ccks2020/eval_paper/ccks2020_eval_paper_1_4_1.pdf, 2020.

[9] 2020 全国知识图谱与语义计算大会评测及任务介绍新冠知识图谱构建问答[EB/OL]. http://sigkg.cn/ccks2020/?page_id=516

[10] Chakraborty, N., Lukovnikov, D., Maheshwari, G., *et al*. (2021) Introduction to Neural Network-Based Question Answering over Knowledge Graphs. *Wiley Interdisciplinary Reviews*: *Data Mining and Knowledge Discovery*, **11**, e1389. https://doi.org/10.1002/widm.1389

[11] Wang, Y., Zhang, R., Xu, C. and Mao, Y. (2018) The APVA-Turbo Approach to Question Answering in Knowledge Base. *Proceedings of the* 27*th International Conference on Computational Linguistics*, Santa Fe, 20-26 August 2018, 1998-2009.

[12] Li, T., Liu, M., Zhang, Y., *et al*. (2021) A Review of Entity Linking Research Based on Deep Learning. *Journal of Peking University*: *Health Sciences*, **57**, 91-98.

[13] Zhou, B., Sun, C., Lin, L., *et al*. (2018) LSTM Based Question Answering for Large Scale Knowledge Base. *Journal of Peking University*: *Health Sciences*, **54**, 286-292.

[14] Yang, Y., He, X., Zhou, K., *et al*. (2019) Multi-Module System for Open Domain Chinese Question Answering over Knowledge Base.

[15] 张鸿志, 李如寐, 王思睿, 黄江华. 基于预训练语言模型的检索-匹配式知识库问答系统[EB/OL]. https://bj.bcebos.com/v1/conference/ccks2020/eval_paper/ccks2020_eval_paper_1_4_2.pdf, 2020.

[16] CCKS&百度 2019 中文短文本的实体链指第一名方案[EB/OL]. https://github.com/panchunguang/ccks_baidu_entity_link, 2019.

[17] Li, X.N., Yan, H., Qiu, X.P. and Huang, X.J. (2020) FLAT: Chinese NER Using Flat-Lattice Transformer. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, Association for Computational Linguistics, Stroudsburg, 6836-6842. https://aclanthology.org/2020.acl-main.611

[18] Reimers, N. and Gurevych, I. (2019) Sentence-BERT: Sentence Embeddings Using Siamese BERT-Networks. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, 3-7 November 2019, 3982-3992. https://doi.org/10.18653/v1/D19-1410

[19] Liu, W.J., Zhou, P., Zhao, Z., *et al*. (2020) K-bert: Enabling Language Representation with Knowledge Graph. *Proceedings of the AAAI Conference on Artificial Intelligence*, **34**, 2901-2908. https://doi.org/10.1609/aaai.v34i03.5681

[20] Xu, Y., Zhu, C., Xu, R., *et al*. (2020) Fusing Context into Knowledge Graph for Commonsense Reasoning.

[21] Wu, S. and He, Y. (2019) Enriching Pre-Trained Language Model with Entity Information for Relation Classification. *Proceedings of the* 28*th ACM International Conference on Information and Knowledge Management*, Beijing, 3-7 November 2019, 2361-2364. https://doi.org/10.1145/3357384.3358119

[22] Conneau, A., Kiela, D., Schwenk, H., *et al*. (2017) Supervised Learning of Universal Sentence Representations from Natural Language Inference Data. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, 7-11 September 2017, 670-680. https://doi.org/10.18653/v1/D17-1070

[23] Mueller, J. and Thyagarajan, A. (2016) Siamese Recurrent Architectures for Learning Sentence Similarity. 30*th AAAI Conference on Artificial Intelligence*, Phoenix, 12-17 February 2016, 2786-2792.

[24] Kim, Y. (2014) Convolutional Neural Networks for Sentence Classification. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1746-1751. https://doi.org/10.3115/v1/D14-1181

[25] CCKS2019&百度. CCKS2019 中文短文本的实体链指[EB/OL]. https://www.biendata.xyz/competition/ccks_2019_el, 2019.