◆自然语言理解◆

融合生成式模型的知识增强实体链指方法*

乔胤博,杨志豪**,林鸿飞
(大连理工大学计算机科学与技术学院,辽宁大连 116024)

摘要:未链接实体分类是实体链指(Entity Linking,EL)任务中的重要研究内容之一。现有方法存在上下文语义信息不充分、分类准确率低等问题,导致实体链指任务表现不佳。本研究提出一种融合生成式模型的知识增强实体链指方法。该方法将实体链指分为两个子模块,即候选实体排序模块和未链接实体分类模块。本研究基于高精度的候选实体排序模块,获得高质量的知识扩展信息,并对未链接实体分类任务进行知识增强;针对未链指实体提及的分类问题,提出一套生成式框架,该框架能够取得超过基线模型的性能。本研究方法在2020年全国知识图谱与语义计算大会(CCKS2020)评测任务二的中文短文本实体链指数据集上取得了目前最佳性能(整体 F 值为91.76%),证明知识增强和生成式框架的引入能提高模型的泛化能力,缓解未链接实体分类中的信息不充分问题。

关键词:生成式;实体链指;知识增强;实体分类;实体排序

中图分类号:TP391.4 **文献标识码**:A **文章编号**:1005-9164(2023)01-0061-10
DOI:10.13656/j.cnki.gxkx.20230308.007

当前世界存在众多结构化数据和非结构化数据,如何将两者有效集成是一个非常困难的问题,实体链指(Entity Linking,EL)[1]就是解决该问题的技术之一。知识图谱是结构化数据集的一种,它以符号形式描述现实世界中的概念及相关关系,实体和关系是其基本组成单位。实体链指指给定输入文本和相关的指称提及(Mention),然后从知识图谱或知识库中检索正确的实体。例如,"乔丹是美国著名篮球运动员。",实体链指会将"乔丹"链接到知识图谱中的"迈克尔·乔丹"实体上。

一般来说,有知识图谱的地方离不开实体链指。实体链指主要应用在知识图谱问答、舆情分析、内容推荐、搜索引擎和知识库拓展等方面[1]。近年来,随着中文短文本实体链指数据集的提出,越来越多研究者关注中文短文本实体链指研究。已有的实体链指研究主要针对长文本实体链指场景[2-4],针对中文短文本实体链指的研究较少。但是随着网络的发展,出现了越来越多的短文本,这些短文本具有极大的研究

收稿日期:2022-11-15 修回日期:2022-12-09
*中央高校基本科研业务费项目(DUT22ZD205)资助。
【第一作者简介】
乔胤博(1996-),男,在读硕士研究生,主要从事知识图谱问答、自然语言处理研究,E-mail:yinboqiao@mail.dlut.edu.cn。
【**通信作者】
杨志豪(1973-),男,教授,主要从事文本挖掘、机器学习、知识图谱研究,E-mail:yangzh@dlut.edu.cn。
【引用本文】
乔胤博,杨志豪,林鸿飞.融合生成式模型的知识增强实体链指方法[J].广西科学,2023,30(1):61-70.
QIAO Y B,YANG Z H,LIN H F. Knowledge Enhanced Entity Linking Method Integrating Generative Model [J]. Guangxi Sciences,2023,30(1):61-70.

价值。短文本实体链指任务的特点是语义信息缺少、上下文信息缺失等[5]。现已有研究关注短文本实体链接的问题,如 Nie 等[6]为解决短文本全局信息不充分的问题,使用多种相似度衡量指标挖掘实体提及和候选实体之间的语义信息;Sakor 等[7]利用大型外部知识库构建扩展知识图谱,基于扩展知识图谱进行候选实体集合生成和候选实体排序。但是这些工作对于未链接实体分类问题缺乏研究和探讨,而未链接实体分类对实体链指整体性能提高具有积极意义,候选实体排序模型性能仍有很大的提升空间。

为解决上述问题,本研究将中文短文本实体链指任务分为两个子模块,分别是候选实体排序模块和未链接实体分类(也称 NIL 实体分类)模块。在分析候选实体排序模型和未链接实体分类模型时发现,提高候选实体排序模块与未链接实体分类模块之间的交互,有助于提升未链接实体分类模块的性能。候选实体排序模块会产生正确的链接实体,利用这部分高质量的已链接实体信息对未链接实体分类任务进行知识增强。另外,生成式模型对传统分类任务有极大的促进作用,因此借助生成式模型,针对未链接实体分类问题提出一种融合生成式模型的框架,充分利用知识库信息并挖掘短文本上下文的信息,提高未链接实体分类的准确率。

1 相关工作

1.1 实体链指

近年来实体链指任务大都基于深度学习算法,使用神经网络模型计算实体指称上下文与候选实体上下文之间的语义相似度,通过相似度得分情况,从候选实体中选择目标实体。研究人员一般将实体指称上下文与候选实体上下文相似度排序问题建模为二分类问题,通过构造正负例样本训练模型的判别能力,从而找到真正的目标实体。Ganea 等[8]提出了局部实体链指注意力机制,将注意力集中在少数能为实体链指提供决策信息的词语上,减少噪声词的影响。Le 等[9]在 Ganea 等[8]工作的基础上,挖掘实体指称之间的潜在关系,假设实体指称之间存在 k 种关系,通过关系矩阵将关系信息融入模型中。Fang 等[10]将全局链接转换为序列决策问题,并提出一个强化学习模型,从全局角度进行决策。Martins 等[11]认为实体识别和实体链指具有强相关性,提出一种联合学习框架,将识别的实体与知识库中的实体进行链接。

近年来一些短文本实体链指工作受到广泛关注。Cheng 等[12]提出了基于 BERT 预训练模型[13]的实体链指方法,使用预训练语言模型提取实体提及和候选实体上下文的特征,提高实体识别和实体链指两个子任务的性能。詹飞等[14]构建多任务框架,在实体链指任务的基础上,引入 NIL 实体分类作为辅助任务,提高了模型的泛化表达能力。本研究提出一种改进的实体链指方法,优化实体分类任务的性能表现,并通过知识增强来增加上下文语义信息。

1.2 生成式模型

机器学习模型可以分为两类模型:判别式模型和生成式模型。目前,深度学习可归纳为 7 个主流模型来解决其对应的问题,比如分类模型、文本匹配模型、阅读理解模型、掩码语言模型等[15]。通过一个统一模型来解决所有任务是众多研究者的目标,尤其在预训练模型提出后,这种目标更有可能实现。因为形式灵活、只需要设计目标序列即可使用,seq2seq 生成式模型逐渐在多个任务中显示出广泛的有效性。Sun 等[15]通过实验剖析了 seq2seq 生成式模型在多个任务中的广泛有效性,认为该生成式模型有希望成为统一模型的通用范式。seq2seq 生成式模型在实体链指任务中也得到了应用,比如 De Cao 等[16]提出一种自回归的端到端实体链指方式,将生成式模型应用到实体链指任务中,将实体检索的多分类任务转换为自回归的生成式任务;Yang 等[17]将生成式模型引入多标签分类任务中,充分考虑标签之间的关系,并取得了优秀的性能。

受生成式模型在实体链指任务中有效应用的启发,本工作在中文短文本的实体链指中引入生成式模型来建模未链接候选实体分类问题。具体使用的生成式框架将会在下一章详细介绍。

2 实体链指模型

2.1 数据预处理

2.1.1 候选实体集生成

常用的构建候选实体集的方法是将所有候选实体加入候选实体集中[18],典型方法有基于别名词典、编辑距离和词向量计算等。

通过对训练集和交叉验证集的数据分析可知,知识库中实体名或实体别名能够覆盖带有实体 ID 标识的全部实体提及。因此,可使用字典匹配方式来获取候选实体集。例如,“小品《战狼故事》中,吴京突破重重障碍解救爱人,深情告白太感人”文本中“吴京”是一个实体指称,通过字典匹配,发现实体 ID 为

“91342”“227925”和“159056”的3个实体名称为“吴京”,因此这3个候选实体构成了“吴京”这个实体指称的候选实体集。

2.1.2 候选实体排序数据集构造

为充分利用知识库信息,尽可能地融入更多的语义信息,将每个知识库实体的属性信息进行拼接。原始数据集中每个实体属性包含多个<predicate,object>二元组,用来存储实体的属性信息。对每个知识库实体,将其所有实体属性信息进行拼接,从而获得一个完整的实体描述。以“吴京”(导演吴京)为例,将其“出生地”“外文名”“摘要”等属性字段进行拼接,得到“吴京”的描述:“职业:演员、导演 代表作品:流浪地球、战狼、战狼Ⅱ……”。接着,将蕴含实体指称的原始文本(下文称为“实体提及上下文”)和所有候选实体属性信息文本(下文称为“候选实体上下文”)进行组合,以构造候选实体排序数据集。该数据集由3个部分组成:实体提及上下文、候选实体上下文和标签。其中标签值为0和1,即将正确的目标实体标签标记为“1”,错误的标记为“0”,一个实体指称的候选实体排序集中只有一个正例,其余皆为负例,参考示例如表1所示。

表1 候选实体排序数据集

Table 1 Candidate entity ranking dataset

实体提及上下文 Entity mention context	候选实体上下文 Candidate entity context	标签 Label
小品《战狼故事》中,吴京突破重重障碍解救爱人,深情告白太感人	职业:演员、导演 代表作品:流浪地球、战狼、战狼Ⅱ……	1
小品《战狼故事》中,吴京突破重重障碍解救爱人,深情告白太感人	摘要:吴京,1934年4月9日出生,祖籍江苏苏州,台湾成功大学土木系毕业……	0
小品《战狼故事》中,吴京突破重重障碍解救爱人,深情告白太感人	主要成就:中国矿业大学杰出校友 职称:高级工程师 摘要:吴京,1914年12月20日生……	0

2.1.3 实体分类数据集构造

通过对2020年全国知识图谱与语义计算大会(CCKS2020)中文短文本实体链指任务数据集的分析可知,知识库中不包含的实体占比11%左右,这部分的实体分类数据非常少,如果只使用这部分实体相关数据训练NIL实体分类模型是不充足的。因此,考虑所有的实体提及,并构造实体分类数据集。数据集中每条数据分为3部分:实体提及上下文、实体提及、类别标签。

利用候选实体排序模块的结果来补充可用的语义信息。如表2所示,通过候选实体排序模型,可成功将“小品”和“战狼故事”链接到知识库中的正确实体上,同时可以利用已链接成功的实体信息来补充实体指称上下文信息。使用括号来包含补充信息,那么表2中的文本就变为“小品《战狼故事》中,吴京(中国内地男演员、导演)突破……”。

表2 实体分类数据集

Table 2 Entity classification dataset

实体提及上下文 Entity mention context	实体提及 Entity mention	标签 Label
小品《战狼故事》中,吴京(中国内地男演员、导演)突破重重障碍解救爱人,深情告白太感人	小品	Other
小品《战狼故事》中,吴京(中国内地男演员、导演)突破重重障碍解救爱人,深情告白太感人	战狼故事	Work
小品《战狼故事》中,吴京(中国内地男演员、导演)突破重重障碍解救爱人,深情告白太感人	吴京	Person

2.2 实体链指模型

如图1所示,本研究提出的实体链指模型包含两个模块,分别是候选实体排序模块(下文简称为“模块1”)和未链接实体分类模块(下文简称为“模块2”),这两个模块通过交互提升实体链指模型的性能表现。

2.2.1 候选实体排序模型

候选实体排序模型整体结构如图2所示。该模型基于候选实体生成方法,生成候选实体集。例如,“小品《战狼故事》中,吴京突破重重障碍解救爱人,深情告白太感人”的候选实体集包含“91342”“227925”“159056”等候选实体,这些候选实体名称都为“吴京”,但是无法从语义层面判断究竟哪一个是目标实体。于是,将短文本和候选实体集都通过候选实体排序模型进行排序,通过语义相似度得分进行候选实体排序,从而确定“159056 吴京 中国内地男演员、导演”是最终的目标实体。本研究将候选实体打分任务认为是Point-wise排序任务。模型由输入层、BERT编码层和线性层组成,图2中[CLS]和[SEP]是BERT预训练模型的特殊标签,用以表示句子的整体表征向量和句对分隔位置。

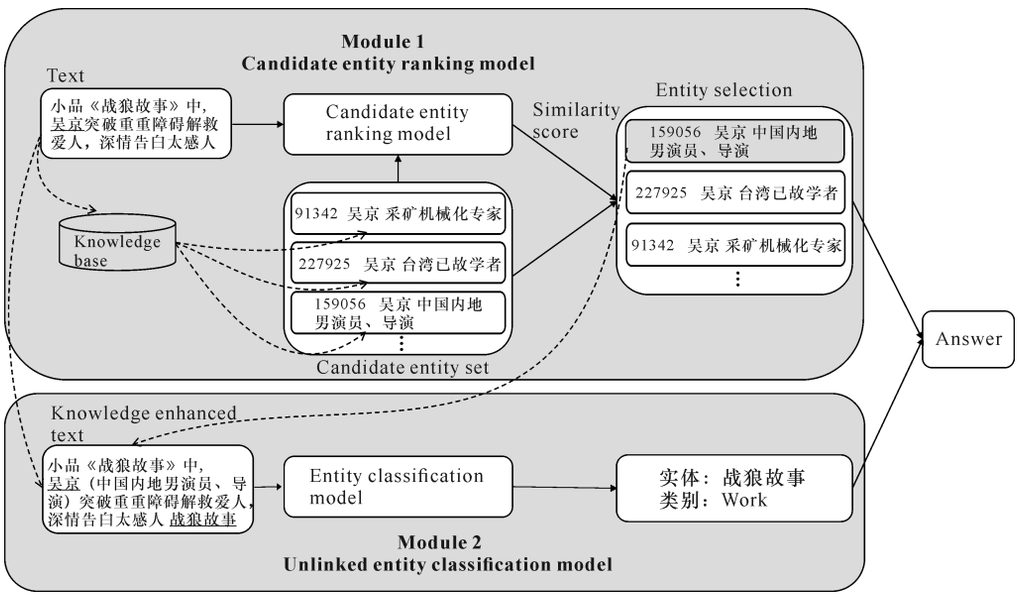

图 1　整体模型图

Fig. 1　Overall model diagram

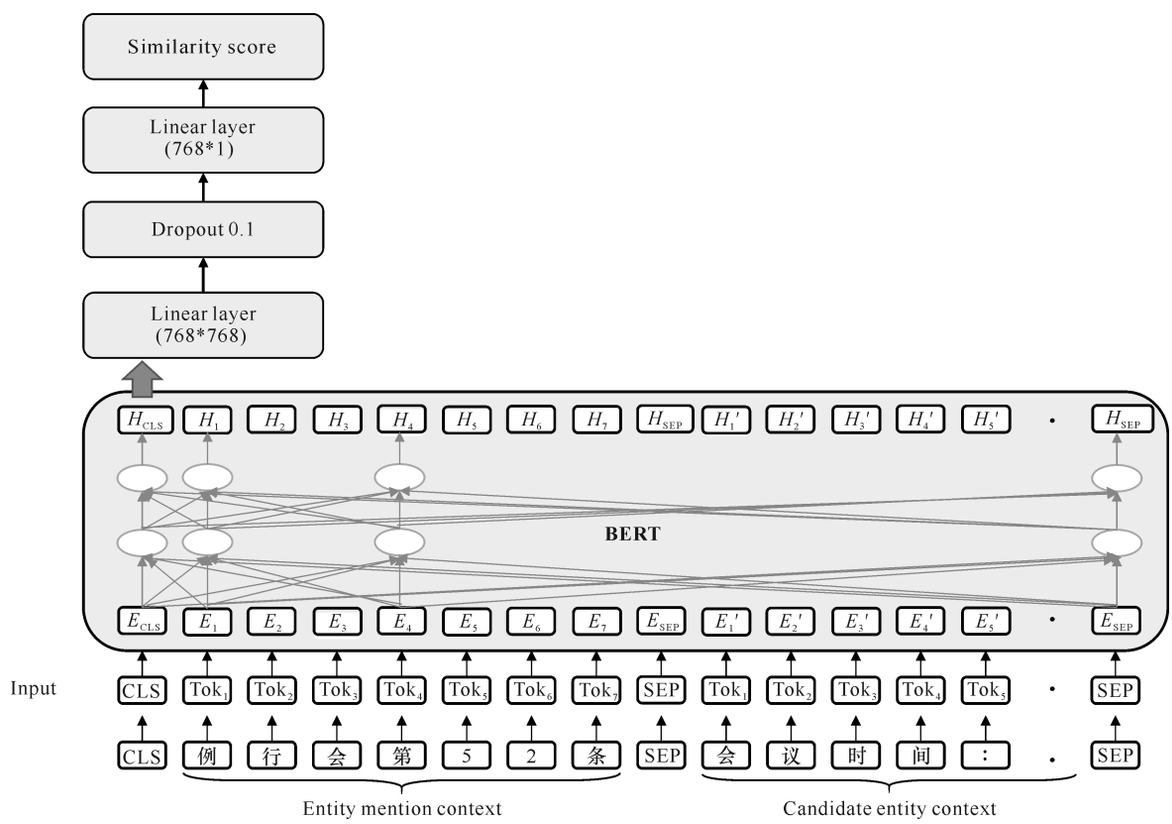

图 2　候选实体排序模型

Fig. 2　Candidate entity ranking model

输入层：输入层由实体提及上下文和候选实体上下文作为模型输入。这里使用的是语句对分类任务，使用[SEP]分隔符连接实体提及上下文和候选实体上下文。

BERT[13]编码层：BERT 预训练模型是优秀的语义特征提取器，对于实体分类任务来说，BERT 的作用是提取输入序列的语义特征。使用 BERT 的结构，由 12 层 Transformer 编码器[19]构成。编码单元最重要的模块就是自注意力（self-attention）部分，如公式（1）所示：

$$(Q,K,V)=\text{softmax}(\frac{QK^{\mathrm{T}}}{\sqrt{d_k}})V, \quad (1)$$

其中,Q,K,V 是输入向量经过线性变换得到的 3 个矩阵向量,$\sqrt{d_k}$ 是归一化参数。

对于输入序列,经过 BERT 编码层后得到输入序列的语义向量表示,其过程如公式(2)所示:

$$\text{CLS}=\text{BERT}(\{x_1,x_2,\cdots,x_n\})。 \quad (2)$$

线性层:通过两个线性层和一个随机失活(Dropout)层得到一维向量,如公式(3)所示:

$$\hat{y}=\text{sigmod}(W_{l_2}\cdot\text{Dropout}(W_{l_1}\cdot\text{CLS})), \quad (3)$$

其中,W_{l_1},W_{l_2} 是权重矩阵,是可训练的参数。

候选实体排序模型的损失函数采用交叉熵损失函数,如公式(4)所示:

$$L_1=-\frac{1}{n}\sum_{i=1}^{n}(y_i\ln\hat{y}_i+(1-y_i)\ln(1-\hat{y}_i)), \quad (4)$$

其中,y_i 表示句对的真实标签,$\hat{y}_i$ 表示模型得到的预测结果。真实标签可参考表 1,即该候选实体是否是正确的目标实体,如果是目标实体,标签为 1,反之则为 0。训练的目标是最小化损失函数。

2.2.2 生成式实体分类模型

候选实体排序模型解决了候选实体排序集不为空的实体指称的链指问题,但是文本中如果存在一些无法找到候选实体的相关实体指称时,则需要将这部分实体指称进行实体分类。由于第一部分的候选实体链接可以提供额外的知识,利用这部分的知识就能对实体分类进行一定程度的知识增强。于是,将候选实体排序模型中链接成功的实体信息摘要信息拼接到实体提及上下文中,形成生成式实体分类模型的知识增强输入文本,并通过实体分类模型进行实体分类。

CCKS2020 中文短文本实体链指数据集中存在"NIL_Work|NIL_Other"这样的多标签问题,而生成式模型可以很好地解决多标签分类问题;另外,生成式模型不需要考虑复杂的正负例构造问题,因此可将传统的分类任务建模成一个生成式任务。生成式实体分类模型整体框架如图 3 所示,该模型主要包含 4 个模块:输入层、BERT 编码层、编码器、解码器。

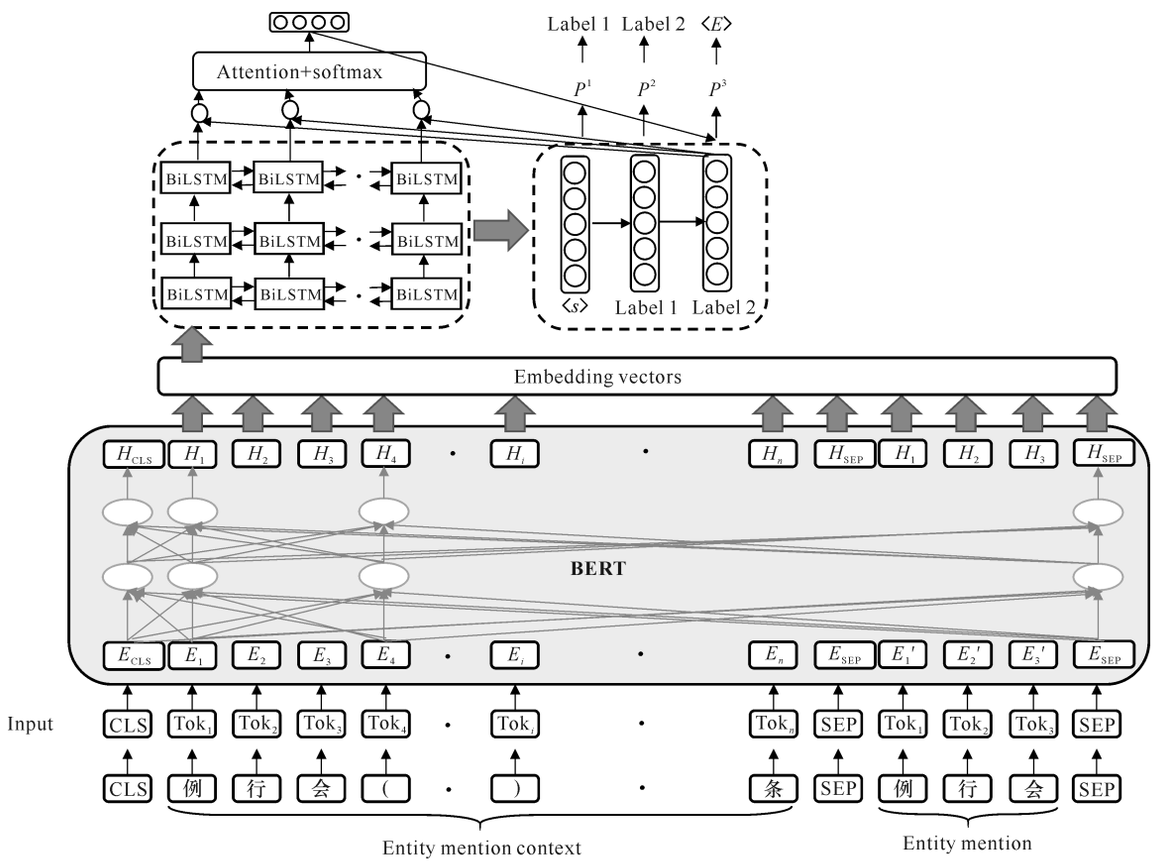

图 3 生成式实体分类模型

Fig. 3 Generative entity classification model

输入层:输入由特殊符号[SEP]连接实体提及上下文和实体提及构成。

BERT 编码层:对于输入序列 $X=\{x_1,x_2,\cdots,x_n\}$,BERT 编码后每个词的向量表示为句子编码向量 $E=\{e_1,e_2,\cdots,e_n\}$。

编码器:编码器结构由 3 层双向长短时记忆(Bi-directional Long Short-Term Memory,BiLSTM)网络构成。BiLSTM 网络包含两个子网络,分别是前向网络和后向网络。对于 BERT 得到的句子编码向量 $E=\{e_1,e_2,\cdots,e_n\}$,BiLSTM 网络产生一个与上下文相关的隐藏向量 h,同时也会得到每个字符对应的编码状态 $S=\{s_1,s_2,\cdots,s_n\}$。

解码器:解码器采用的是基于注意力的序列到序列(Attention-based seq2seq)模型[20]。编码器基于 BiLSTM 单元将输入序列压缩到单个隐藏向量中,在句子长度较长的情况下,这个过程中会发生语义信息的丢失。这一情况导致解码器难以从被压缩的片面信息中解码出相关信息。因此,本研究解码器利用注意力 seq2seq 机制,在解码过程中通过注意力分数判断哪部分对标签生成更为重要。按照时间步进行解码,通过公式(5)-(8)计算出时间 t 对应的输出向量 o^t。

$$\mathrm{score}(h_t,s_k)=h_t^{\mathrm{T}}s_k, \tag{5}$$

$$a_k^{(t)}=\frac{\exp(\mathrm{score}(h_t,s_k))}{\sum_{i=1}^{n}\exp(\mathrm{score}(h_t,s_i))},k=1,2,\cdots,n, \tag{6}$$

$$c^{(t)}=a_1^{(t)}s_1+\cdots+a_n^{(t)}s_n=\sum_{k=1}^{n}a_k^{(t)}s_k, \tag{7}$$

$$o^t=\tanh(W_c\cdot[h_t,c^{(t)}]), \tag{8}$$

其中,$a_k^{(t)}$ 表示在 t 时刻第 k 个字符的注意力得分;$c^{(t)}$ 表示 t 时刻所有字符及其注意力得分的加权向量;h_t 表示第 t 时刻 BiLSTM 模型的隐藏状态;s_k 表示 k 时刻的 BiLSTM 模型输出向量;W_c 是一个权重矩阵,是可训练的参数。

生成式实体分类模型采用交叉熵损失函数,假设当前输入序列对应的标签为 $y=(y_1,y_2,\cdots,y_m)$,对应的独热向量为 o^*,损失函数如公式(9)所示:

$$L_2=-\sum_{i=1}^{m}o_i^*\log(o_i^t)。 \tag{9}$$

3 实验仿真及分析

3.1 数据集

本文选用 CCKS2020 评测任务二的中文短文本实体链指数据集,该数据集包含知识库文件与标准数据集。标准数据集包括训练集、验证集,其中训练集 7 万条数据,验证集 1 万条数据。

3.2 评价指标

在给定输入文本序列 $\{x_1,x_2,\cdots,x_t\}$(其中 t 是输入文本的字符数)情况下,假设文本中有 n 个实体提及 $M=\{m_1,m_2,\cdots,m_n\}$,每个实体提及对应知识库中的实体 ID 为 $E_n=\{e_1,e_2,\cdots,e_n\}$,模型预测的每个实体提及所对应的知识库中实体 ID 为 $E_n^{'}=\{e_1^{'},e_2^{'},\cdots,e_n^{'}\}$,则模块 1 的准确率、召回率和 F 值定义如公式(10)、公式(11)所示:

$$P_M=R_M=\frac{\sum_{n\in N}|E_n\cap E_n^{'}|}{\sum_{n\in N}|E_n^{'}|}, \tag{10}$$

$$F_M=\frac{2(P_M\times R_M)}{P_M+R_M}, \tag{11}$$

其中,N 表示候选实体排序数据集的全体数据。

对于未链接实体,在给定输入文本序列中,假设有 k 个未链接实体提及 $C=\{c_1,c_2,\cdots,c_k\}$,每个实体提及对应的类别为 $U_k=\{u_1,u,\cdots,u_k\}$,模型预测的每个实体提及类别为 $U_k^{'}=\{u_1^{'},u_2^{'},\cdots,u_k^{'}\}$,模块 2 的准确率、召回率和 F 值定义如公式(12)(13)所示:

$$P_C=R_C=\frac{\sum_{k\in K}|U_k\cap U_k^{'}|}{\sum_{k\in K}|U_k^{'}|}, \tag{12}$$

$$F_C=\frac{2(P_C\times R_C)}{P_C+R_C}, \tag{13}$$

其中,K 表示未链接实体分类数据集中的全体数据。

总体的评价指标是模块 1 和模块 2 的综合准确率、召回率和 F 值。综合准确率、召回率和 F 值定义如公式(14)(15)所示:

$$P=R=\frac{\sum_{n\in N}|E_n\cap E_n^{'}|+\sum_{k\in K}|U_k\cap U_k^{'}|}{\sum_{n\in N}|E_n^{'}|+\sum_{k\in K}|U_k^{'}|}, \tag{14}$$

$$F=\frac{2(P\times R)}{P+R}。 \tag{15}$$

CCKS2020 标准数据集中实体指称是给定的,不需要进行实体识别,因此 $P=R=F$,将 F 值作为模型最终的评价指标。

3.3 实验设置

学习率设定为 5×10^{-5},使用 Adam 优化器[21]优化模型参数,使用的 LSTM 有 3 层。模块 1 的批量大小(batch)设置为 32,模块 2 的批量大小设置为 16。

实验采用 Ubuntu 操作系统,CPU 为 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20 GHz,GPU 为 NVIDIA Geforce RTX3090 和 TAITAN XP,python 版本为 3.7,pytorch 版本为 1.10,tensorflow 版本为 1.15.0。

3.4 基线模型

本文选取 5 个基线模型作为对照组进行实验,实验结果都基于 CCKS2020 中文短文本实体链指数据集中的交叉验证集结果。

模型 1[22]:提出了一个流水线方法,首先将实体链指任务建模为实体排序任务,如果一个实体不能被链接到知识库中,那么将该实体进行 NIL 实体分类。

模型 2[23]:使用飞桨框架实现多任务模型,文本中称多任务模型为多任务层投影模型,一定程度上节省了多任务场景下的参数量,并提高了实体链指的准确率。

模型 3[24]:设计了一个多特征因子融合的实体链指模型,先对短文本中的实体提及进行类别预测,然后将其构造为 NIL 实体,加入该实体的候选实体集中,最后利用多层感知机将多特征因子融合打分,根据候选实体和实体提及上下文进行相似度匹配,选择得分最高的候选实体作为实体链指结果。

模型 4[25]:将 NIL 实体加入候选实体集中参与实体排序,最终得到链接实体,并针对 NIL 实体任务,提出基于问答的 NIL 实体判断模型。

模型 5[15]:将多任务学习方法引入短文本实体链指任务中,将 NIL 实体分类任务作为实体链指的辅助任务,提高模型的泛化能力,优化模型在实体链指中的表现。

3.5 实验结果

参考官方评价指标,使用 F 值来衡量候选实体排序模块(模块 1)和未链接实体分类模块(模块 2)的性能。5 个基线模型并没有都将实体链指任务划分为两个模块分别评价,因此模块 1 和模块 2 的 F 值都为空。实验结果如表 3 所示,本文模型在中文短文本实体链指数据集上取得了目前最好的结果,在 CCKS2020 中文短文本实体链接数据集上比当前最高结果高约 2.5%,说明知识增加策略和生成式模型能够很好地融合有效信息,提取相关特征并提升模型的整体表现。

另外,模块 1 的 F 值达到 92.7%,比总体 F 值高约 1%,说明构建的候选实体排序模型非常有效。模块 2 的 F 值能达到 87.86%,接近 5 个基线模型总体 F 值的性能表现,进一步说明知识增强策略和生成式模型的有效性。

表 3 实验结果

Table 3 Experimental results

模型 Model	总体 F 值(%) Overall F value (%)
Model 1[22]	89.27
Model 2[23]	88.19
Model 3[24]	89.29
Model 4[25]	88.01
Model 5[15]	88.49
Ours	91.76

3.6 方法有效性分析

3.6.1 对比实验分析

本文采用 BERT 预训练模型对输入文本序列进行语义信息的提取。为比较不同预训练模型在本文数据集任务上可能存在的结果差别,选择 5 种不同的主流预训练模型 ALBERT[26]、ERNIE[27]、RoBERTa[28]、XLNet[29] 和 BERT[13] 进行实验对比,实验结果如表 4 所示。通过表 4 可知,在候选实体排序模块和实体分类模块上,预训练语言模型都能取得优秀的结果,其中基于 BERT 模型的效果最好。

表 4 不同预训练模型的实验结果

Table 4 Experimental results of different pre-training models

模型 Model	F 值(%) F value (%)		
	模块 1 Module 1	模块 2 Module 2	总体 Overall
ALBERT[26]	89.27	78.45	87.15
ERNIE[27]	92.46	87.70	91.53
RoBERTa[28]	92.46	87.70	91.53
XLNet[29]	92.46	87.70	91.53
BERT[13]	92.71	87.86	91.76

本文两个模块实际上都属于文本分类任务,为进一步说明本文提出模型的有效性,选取文本分类领域通用的几个模型进行对比。针对模块 1,选择 BERT + CNN、BERT + RNN、BERT + RCNN 作为对比,比较 F 值,实验结果如表 5 所示。从表 5 可知,采用本文的候选实体排序模型效果最好,取得了目前最好的结果,说明该模型对本文的任务有效。

表 5　模块 1 对比实验

Table 5　Comparative experiment of module 1

模型 Model	F 值(%)F value (%)
BERT + CNN	86.92
BERT + RNN	82.07
BERT + RCNN	92.46
Ours	92.71

针对模块 2,选取 BERT + CNN、BERT + RNN、BERT + RCNN、BERT + Dense 和本文的模型 BERT + seq2seq + Attention 进行对比,评价指标为 F 值,实验结果如表 6 所示,BERT + seq2seq + Attention 取得了最优异的性能表现,说明提出的方法对未链接实体多分类任务有效。

表 6　模块 2 对比实验

Table 6　Comparative experiment of module 2

模型 Model	F 值 (%) F value (%)
BERT + CNN	79.95
BERT + RNN	77.65
BERT + RCNN	87.70
BERT + Dense	85.89
Ours	87.86

3.6.2　消融实验分析

通过消融实验来研究本文方法各个部分的有效性,通过移除不同因素进行对比,比如知识增强、NIL 分类任务、生成式模型等。实验结果如表 7 所示,表中“ - w/o KE”表示去除模型中的知识增强,“ - w/o Seq”表示在没有生成式模型的情况下,使用基础的 BERT + Dense;“ - w/o NIL”表示没有 NIL 实体分类模块。通过 - w/o KE 消融实验可知,去除模型中的知识增强,模块 2 的 F 值下降约 3.8%,总体 F 值也下降约 0.9%,因此,知识增强有助于缓解短文本实体链指实体提及上下文语义信息不充分的问题。由 - w/o Seq 消融实验结果可知,模块 2 的 F 值下降约 2%,总体 F 值也下降约 0.7%,表明用本文提出的生成式模型来进行 NIL 实体分类是有效的。本文通过两个模块解决实体链指问题。如果只采用第一个模块进行候选实体排序而没有用第二个 NIL 实体分类模块,模型总体 F 值大幅下降约 9%,这充分说明模块 1 与模块 2 能够有效解决 CCKS2020 中文短文本实体链指任务。

表 7　消融实验结果

Table 7　Ablation experimental results

模型 Model	F 值(%)F value (%)		
	模块 1 Module 1	模块 2 Module 2	总体 Overall
- w/o KE	92.46	84.08	90.81
- w/o Seq	92.34	85.89	91.08
- w/o NIL	82.33	/	82.33
Ours	92.71	87.86	91.76

Note:"/" indicates that there are no corresponding experimental results

4　结论

本文提出一种融合生成式模型的知识增强实体链指方法,来解决中文短文本实体链指任务。本文模型能够准确地捕捉实体提及和候选实体之间的相似性关系,并通过知识库本身的结构化知识扩充语义信息,提升了 CCKS2020 实体链指任务的性能。一系列的实验和分析表明本文提出的模型取得了目前该数据集上最好的结果,从而证明了本文方法的有效性。

参考文献

[1] SHEN W,WANG J,HAN J. Entity linking with a knowledge base: issues, techniques, and solutions [J]. IEEE Transactions on Knowledge and Data Engineering,2015,27(2):443-460.

[2] ZHANG Q,SUN Z,HU W,et al. Multi-view knowledge graph embedding for entity alignment [C]//Proceedings of the Twenty-eighth International Joint Conference on Artificial Intelligence. [S. l.]:IJCAI,2019:5429-5435.

[3] CETOLI A,BRAGAGLIA S,O'HARNEY A D,et al. A neural approach to entity linking on wikidata [C]// AZZOPARDI L,STEIN B,FUHR N,et al. Advances in Information Retrieval. ECIR 2019. Lecture Notes in Computer Science. [S. l.]: Springer,2019(11438):78-86.

[4] 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.

[5] TRISEDYA B D,QI J,ZHANG R. Entity alignment between knowledge graphs using attribute embeddings [J]. Proceedings of the AAAI Conference on Artificial Intelligence,2019,33(1):297-304.

[6] NIE F,ZHOU S,LIU J,et al. Aggregated semantic matching for short text entity linking [C]//Proceedings of the 22nd Conference on Computational Natural Language Learning. Stroudsburg,PA: Association for Com-

putational Linguistics,2018:476-485.

[7] SAKOR A,MULANG I O,SINGH K,et al. Old is gold:linguistic driven approach for entity and relation linking of short text [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). Stroudsburg, PA: Association for Computational Linguistics,2019:2336-2346.

[8] GANEA O,HOFMANN T. Deep joint entity disambiguation with local neural attention [C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA: Association for Computational Linguistics,2017:2619-2629.

[9] LE P,TITOV I. Improving entity linking by modeling latent relations between mentions [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1:Long Papers). Stroudsburg, PA: Association for Computational Linguistics, 2018:1595-1604.

[10] FANG Z,CAO Y,LI Q,et al. Joint entity linking with deep reinforcement learning [C]//WWW'19:The World Wide Web Conference. New York: Association for Computing Machinery,2019:438-447.

[11] MARTINS P H,MARINHO Z,MARTINS A E F. Joint learning of named entity recognition and entity linking [C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop. Stroudsburg,PA: Association for Computational Linguistics,2019:190-196.

[12] CHENG J,PAN C,DANG J,et al. Entity linking for Chinese short texts based on BERT and entity name embeddings [C/OL]//China Conference on Knowledge Graph and Se-mantic Computing,2019:1-12[2022-10-15]. https://conference. bj. bcebos. com/ccks2019/eval/webpage/pdfs/eval_paper_2_1. pdf.

[13] DEVLIN J,CHANG M,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). Stroudsburg, PA: Association for Computational Linguistics, 2019:4171-4186.

[14] 詹飞,朱艳辉,梁文桐,等.基于多任务学习的短文本实体链接方法[J].计算机工程,2022,48(3):315-320.

[15] SUN T X,LIU X Y,QIU X P,et al. Paradigm shift in natural language processing [J]. Machine Intelligence Research,2022,19(3):169-183.

[16] DE CAO N,IZACARD G,RIEDEL S,et al. Autoregressive entity retrieval [Z/OL]. (2021-03-24) [2022-10-17]. https://arxiv. org/pdf/2010. 00904. pdf.

[17] YANG P,SUN X,LI W,et al. SGM:sequence generation model for multi-label classification [C]//Proceedings of the 27th International Conference on Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics,2018:39151-3926.

[18] 张晟旗,王元龙,李茹,等.基于局部注意力机制的中文短文本实体链接[J].计算机工程,2021,47(11):77-83,92.

[19] VASWANI A,SHAZEER N,PARMAR N,et al. Attention is all you need [C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach,CA:NIPS,2017.

[20] BAHDANAU D,CHO K H,BENGIO Y. Neural machine translation by jointly learning to align and translate [C]//3rd International Conference on Learning Representations. San Diego:ICLR 2015,2015.

[21] KINGMA D P,BA J. Adam:a method for stochastic optimization [C]//The 3rd International Conference on Learning Representations. San Diego: ICLR 2015, 2015.

[22] ZHU F. Improving entity linking by a novel pipeline method for chinese short text [C]//CCKS:China Conference on Knowledge Graph and Semantic Computing. [S. l.]:Springer,2020.

[23] 何长鸿,孙承杰,林磊,等.基于多任务层投影 ERNIE 的短文本实体链接及实体分类方法[C/OL]//全国知识图谱与语义计算大会论文集,2020[2022-10-19]. https://bj. bcebos. com/v1/conference/ccks2020/eval _ paper/ccks2020_eval_paper_2_8. pdf.

[24] 吕荣荣,王鹏程,陈帅.面向中文短文本的多因子融合实体链指研究[C/OL]//全国知识图谱与语义计算大会论文集,2020[2022-10-19]. https://bj. bcebos. com/v1/conference/ccks2020/eval _ paper/ccks2020 _ eval _ paper_2_1. pdf.

[25] 潘春光,王胜广,罗志鹏.知识增强的实体消歧与实体类别判断[C/OL]//全国知识图谱与语义计算大会论文集,2020[2022-10-19]. https://bj. bcebos. com/v1/conference/ccks2020/eval _ paper/ccks2020 _ eval _ paper_2_2. pdf.

[26] LAN Z,CHEN M,GOODMAN S,et al. ALBERT:a lite BERT for self-supervised learning of language representations [C]//The 8th International Conference on Learning Representations. Addis Ababa: ICLR 2020, 2020.

[27] SUN Y,WANG S,LI Y,et al. ERNIE:enhanced representation through knowledge integration [Z/OL]. (2019-04-19)[2022-10-19]. https://arxiv. org/pdf/1904. 09223. pdf.

[28] LIU Y,OTT M,GOYAL N,et al. ROBERTa:a robustly optimized bert pretraining approach [Z/OL]. (2019-07-26)[2022-10-19]. https://arxiv. org/pdf/1907. 11692. pdf.

[29] YANG Z,DAI Z,YANG Y,et al. XLNet:generalized autoregressive pretraining for language understanding [C]//Advances in Neural Information Processing Systems 32 (NeurIPS 2019). New York:NeurIPS,2019.

Knowledge Enhanced Entity Linking Method Integrating Generative Model

QIAO Yinbo,YANG Zhihao**,LIN Hongfei

(College of Computer Science and Technology,Dalian University of Technology,Dalian,Liaoning,116024,China)

Abstract:Unlinked entity classification is one of the important research contents in Entity Linking (EL) task. The existing methods have problems such as insufficient contextual semantic information and low classification accuracy,which lead to poor performance of entity linking tasks. A knowledge-enhanced entity linking method integrating generative models is proposed in this study. This method divides the entity linking into two sub-modules,namely the candidate entity sorting module and the unlinked entity classification module. Based on the high-precision candidate entity ranking module,the high-quality knowledge expansion information is obtained and the knowledge of unlinked entity classification tasks is enhanced. Aiming at the classification problem mentioned by unlinked entities,a generative framework is proposed,which can achieve better performance than the baseline model. This research method has achieved the best performance on the Chinese short text entity linking dataset of China Conference on Knowledge Graph and Semantic Computing in 2020 (CCKS2020) evaluation task 2 (the overall F value is 91.76%),which proves that the introduction of knowledge enhancement and generative framework can improve the generalization ability of the model and alleviate the problem of insufficient information in unlinked entity classification.

Key words:generative;entity linking;knowledge enhancement;entity classification;entity ranking

责任编辑:米慧芝