计算机科学
COMPUTER SCIENCE

增强实体表示的文档级关系抽取方法研究

丁肖摇, 周刚, 卢记仓, 陈静

引用本文

丁肖摇, 周刚, 卢记仓, 陈静.增强实体表示的文档级关系抽取方法研究[J]. 计算机科学, 2023, 50(8): 157-162.

DING Xiaoyao, ZHOU Gang, LU Jicang, CHEN Jing. Study on Enhanced Entity Representation for Document-level Relation Extraction [J]. Computer Science, 2023, 50(8): 157-162.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220700161

增强实体表示的文档级关系抽取方法研究

丁肖摇[1]　周　刚[1,2]　卢记仓[1,2]　陈　静[1,2]
1 战略支援部队信息工程大学　郑州 450001
2 数学工程与先进计算国家重点实验室　郑州 450001
(dingxiaoyao2006@126.com)

摘　要　文档级关系抽取是自然语言处理领域研究的热点和难点问题,基于图的模型是当前文档级关系抽取的主流方法之一,该类方法虽然能有效解决实体节点之间的长距离依赖问题,但其在构造节点时往往未充分考虑句子上下文、文档主题、实体对距离、实体对相似度等额外信息,导致关系抽取的性能较低。针对该问题,提出了基于增强实体表示的文档级关系抽取模型。首先,将原始文档作为输入,构建基础文档图结构;然后,通过图神经网络传播机制聚合邻接点的信息,将与实体关系预测相关的句子上下文、主题信息融入基础文档图的实体节点表示中,从而获得增强的实体节点表示;最后,利用增强后实体节点的图模型对实体关系进行预测。实验结果表明,所提模型在文档级关系抽取任务中的性能优于已有模型,且可解释性更好。

关键词:文档级;关系抽取;实体表示;图模型
中图法分类号　TN929.5

Study on Enhanced Entity Representation for Document-level Relation Extraction

DING Xiaoyao[1],ZHOU Gang[1,2],LU Jicang[1,2] and CHEN Jing[1,2]
1 PLA Strategic Support Force Information Engineering University,Zhengzhou 450001,China
2 State Key Laboratory of Mathematical Engineering and Advanced Computing,Zhengzhou 450001,China

Abstract　Document-level relation extraction is a hot and challenging issue in natural language processing. Graph-based model is one of the mainstream methods of document-level relation extraction. Although this method can effectively solve the long-distance dependency between entity nodes,it often fails to fully consider the additional information such as sentence context,document topic,entity to entity distance and entity to similarity when constructing nodes,resulting in low performance of relationship extraction. A document-level relation extraction model based on enhanced entity representation is proposed to solve this problem. Firstly,the original document is used as input to construct the basic document graph structure. Then,the graph neural network propagation mechanism is used to aggregate the information of adjacent nodes,and the sentence context and topic information related to entity relation prediction is integrated into the entity node representation of the primary document graph,to obtain an enhanced entity node representation. Finally,the graph model of the enhanced entity node is used to predict the entity relationship. Experimental results show that the performance of the proposed model in the document-level relation extraction task is better than that of the existing models,and has better interpretability.

Keywords　Document-level,Relation extraction,Entity representation,Graph-based model

1　引言

关系抽取的目的是从一个自然语言文本中抽取出实体对的关系,它在自然语言处理下游的实际场景中发挥着重要作用,如知识图谱构建[1]、问答系统[2]、信息检索与分析[3]等。之前的研究主要关注句子级关系抽取[4-6],即在单个句子中抽取实体间的关系。但在现实世界中,大量的实体关系往往是通过多个句子来表达的,根据对维基百科中人工标注的数据的统计[7],至少 40.7% 的实体关系只能通过多个句子才能抽取到,鉴于此,目前更多的学者将注意力转移到文档级关系抽取的研究中。

与句子级关系抽取相比,文档级关系抽取面临着更为复杂的挑战。首先,文档级关系抽取要求更多的推理技术(逻辑推理、共指推理、常识推理),这些推理为实体对提供了预测关系的路径。其次文档级关系抽取需要考虑实体对的上下文信息,复杂的上下文交互为实体关系抽取提供了重要的辅助信息。如图 1 所示,其中不同颜色字体表示不同的实体,为了简洁仅标注部分实体和关系。“Allen F. Moore”是 DocRED[7]

到稿日期:2022-07-18　返修日期:2023-05-24
基金项目:河南省自然科学基金(222300420590)
This work was supported by the Natural Science Foundation of Henan Province,China(222300420590).
通信作者:周刚(gzhougzhou@126.com)

中的一个文档例子，这里要对红色字体的实体“Monticello High School”和蓝色字体的实体“U. S. ”进行关系预测。首先需要通过句子 0 来识别“Illinois”和“U. S. ”的所属关系，以及句子 2 中“Illinois”和“Monticello”的所属关系，这两个句内关系的抽取是比较容易的，因为不涉及跨句子信息的推理；其次确定句子 0 中“Illinois”和句子 2 中的“Illinois”是共指关系，即这两个句子中“Illinois”的共同实体都是“Illinois”。为了预测“Monticello High School”和“U. S. ”的关系，还需要知道句间实体“Monticello High School”和“Monticello”是所属关系，最后通过逻辑推理，可以得到目标实体对的“country”关系：(Monticello High School, country, U. S.)。

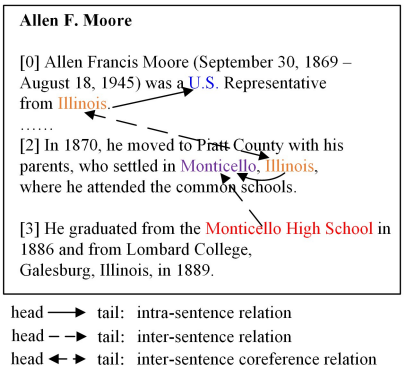

图 1 DocRED 中的例子(电子版为彩图)

Fig. 1 A case in DocRED

本文提出了增强实体表示的文档级关系抽取模型。我们将文档转化成图结构，图中包含任何两个存在关系的实体节点，从而可以较好地利用图模型处理长距离依赖推理。模型的亮点在于实体节点表示不仅包含了实体自身表示，还利用注意力机制得到了实体所在句子的上下文信息和文档的主题信息，另外通过关系图卷积网络(Relational Graph Convolutional Networks, R-GCN)[8]聚合了实体的邻居节点信息。模型通过上述技术能够更好地处理句内关系推理和句间关系推理。其中句子上下文信息的引入可进一步提升句内实体关系抽取的性能，文档主题信息的引入能够较好地应对句间实体关系推理，这也是目前文档级关系抽取研究的重点和难点。

本文提出的模型在大型的文档级关系抽取数据集 DocRED 上进行了实验，实验结果表明所提模型优于现有的文档级关系抽取模型。

2 相关工作

基于图的模型[9-13]已被广泛用于文档级关系抽取中，并且在抽取性能上展现出了巨大的优势。

Christopoulou 等[9]提出了一种面向边的文档级关系抽取模型，通过将提及、实体和句子作为节点，将节点之间的联系作为边构造成异构图，3 种不同的节点都采用平均池化的方式生成，而边的构建采用的是启发式规则，通过其他边的交互生成实体到实体的边进而推断目标实体的关系。Nan 等[10]提出了潜在结构细化模型，创新性地引入了元路径节点，作为最短路径上的元路径节点尽可能地丢弃了无用的噪声信息，紧接着通过图卷积网络(Graph Convolutional Networks, GCNs))[11]有效地收集邻居节点的信息，但这种方法并没有区分不同类型的边，给相似边类型的实体关系判断造成了困难。Zeng 等[12]提出了一种区分句内和句间关系推理的模型，可以发现节点的构建类型与之前的研究没有差别，在区分边的类型上以句内和句外为边界，生成了句内实体边和句间实体边，利用边类型的不同构造了提及图和实体图，最终的实体关系预测通过实体图来完成。Xu 等[13]提出了新颖的编码器-分类器，文中对图中的真实依赖路径进行了重构，根据推理方式的不同，将实体对路径分为模式识别路径、逻辑推理路径以及共指推理路径；重构路径为了增强有关系实体的分类概率，弱化没有关系实体对的分类概率，在分类指标上使用重构器来辅助分类关系，这种技术显著提高了文档级关系抽取的性能。

以上基于图的文档级关系抽取方法实现了较好的性能，但是存在两个问题：1)对于节点的构造只是简单的平均池化操作，没有考虑与节点相关的信息，导致节点的上下文语义信息不丰富；且对于句内关系和句间关系没有从上下文信息上进行区分，造成了句内关系和句间关系只是机械地统一处理。2)没有利用图中邻接节点信息，图中的实体信息是相互影响的，特别是邻接的实体，在语境表达上很相近，但是之前的模型没有考虑此问题，导致实体间的信息割裂。针对以上两个问题，本文模型给出了解决方法。对于问题 1)，我们将实体所在句子的上下文信息以及文档主题信息融入实体表示中，当实体关系在同一句话中，句子的上下文信息就发挥了作用，对于跨句子的实体关系分类，就需要结合两者。通过此方法可以有效解决实体表示单一的问题。对于问题 2)，我们采用 R-GCN 聚合邻居节点，将邻居节点的信息以广播的形式传递到目标实体节点上，这样实体节点就包含了邻居节点的信息，提高了目标实体关系预测的准确性。

3 任务描述

对于一个给定的输入文档 D，并且该文档已经对实体进行了标注，文档 $D=\{s_1, s_2, s_3, \cdots, s_N\}$ 由 N 个句子组成，第 i 个句子包含了 j 个单词 $s_i=\{w_1, w_2, w_3, \cdots, w_j\}$，那么输入文档 D 最终可以由单词表示为：$D=\{w_1, w_2, w_3, \cdots, w_j, \cdots, w_k\}$。输入文档也可以表示为由 n 个实体 $\mathscr{E}=\{e_1, e_2, e_3, \cdots, e_n\}$ 组成。文档级关系抽取任务的目的是从实体集 $\mathscr{E}$ 中取其中两个实体 (e_s, e_o) 以分类其关系类型，这里的关系类型集合 $\mathscr{R}$ 是预定义好的，如 Yao 等[7]对维基百科文档和维基数据知识库分析总结后发现，所涉及的关系类型包括“capital of”“spouse”“head of government”等共计 96 种，实体对关系集合通常可表示为：

$$\{(e_s, r_{so}, e_o) \mid e_s, e_o \in \mathscr{E}, r_{so} \in \mathscr{R}\} \tag{1}$$

4 增强实体表示的文档级关系抽取模型

如图 2 所示，整个模型包含 3 个部分，即编码层、增强实体表示层以及分类层。在编码层中，我们采用 BERT(Bidirectional Encoder Representations from Transformers)[14]预训练模型对输入的文档进行编码。增强实体表示层是模型的

核心,为了提升句内关系抽取的能力,在实体表示中加入了句子上下文表示;为了增强句间推理能力,构造的实体表示中增加了文档主题信息表示,这样实体的最终表示由实体表示本身、句子上下文表示以及文档主题信息表示构成。然后通过聚合邻居节点的信息进一步增强实体节点的表示能力。在分类层中采用双线性函数对实体对进行分类。

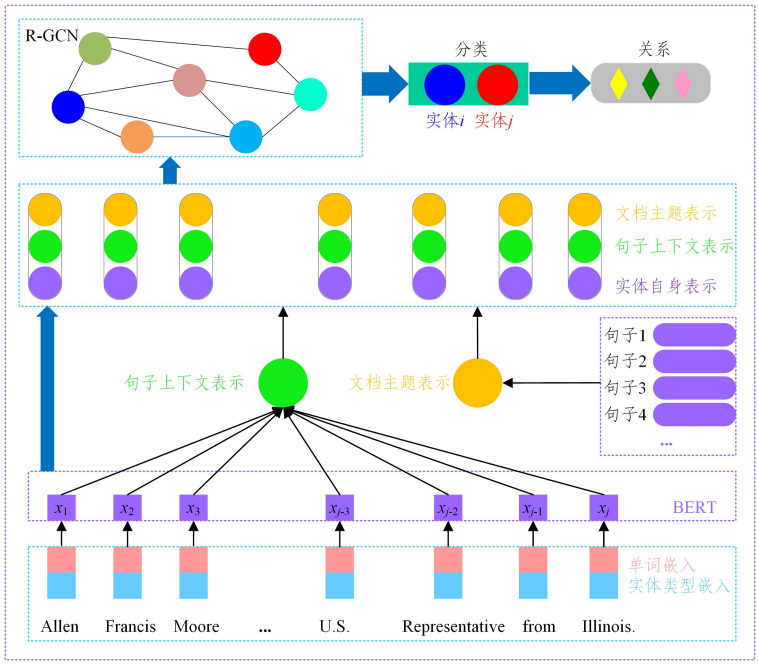

图 2 模型框架

Fig. 2 Framework of the proposed model

4.1 编码层

文档中的每一个单词 w_i 都需要映射成一个向量。考虑到实体类型对实体对的预测有很大的帮助[15-16],我们选择 Yao 等[7]定义的 6 种实体类型(PER,ORG,LOC 等),并将单词的嵌入表示与相对应的实体类型嵌入表示进行连接操作。对于没有实体类型的单词,用 NONE 代替,单词的嵌入表示如下:

$$x_i=[E_w(w_i);E_t(t_i)] \tag{2}$$

其中,$E_w(w_i)$表示单词的嵌入,$E_t(t_i)$表示单词所对应的实体类型的嵌入,[;]表示嵌入的拼接操作。

然后将每个单词嵌入送到 BERT 编码器中,得到每个单词的最终向量化表示:

$$[h_1,h_2,h_3,\cdots,h_j]=BERT([x_1,x_2,x_3,\cdots,x_j]) \tag{3}$$

其中,h_i 是单词嵌入 x_i 在 BERT 编码器的最后一层输出的隐藏表示。

4.2 增强实体表示层

为了利用图的推理能力,首先需要构造一个文档图,文档图中包含了实体节点和实体间的边,实体节点的构造是图推理增强的基础。

由于实体在文档中可能被多次提及,并且一次提及也可能包含多个单词,对于实体的表示,我们使用最大池化的平滑版本[17],对实体所包含的提及进行 LogSumExp 操作。具体地,对第 i 个实体节点,其提及集中的提及数量表示为$\{M_{e_i}\}$,那么第 i 个实体节点可以表示为:

$$e_i=\log\sum_{p=1}^{M_{e_i}}\exp(m_p) \tag{4}$$

其中,m_p 是实体所对应的提及。

实体所在的每一个句子,都对实体所表达的信息起着补充和丰富的作用,特别是对于句内关系的抽取,关系分类的两个实体对象都存在于同一个句子中,这时获取实体所在句子的上下文信息尤为重要。假设第 i 个实体所在的句子由 j 个单词组成,这里编码器同样选择 BERT,那么第 i 个实体的所在的句子可以表示为:

$$S^i=\{h_1^i,h_2^i,h_3^i,\cdots,h_j^i\}=BERT([x_1^i,x_2^i,x_3^i,\cdots,x_j^i]) \tag{5}$$

由于句子中每个单词对句子上下文信息所起到的作用是不同的,因此我们采用自注意力机制,对实体影响大的信息词分配更高的权重,反之分配较低的权重,最后通过聚合句子中的不同词形成句子的上下文表示。

$$u_j^i=\tanh(W_1h_j^i+b_1) \tag{6}$$

$$\alpha_j^i=\frac{\exp(u_j^i)}{\sum_j\exp(u_j^i)} \tag{7}$$

$$S_i=\sum_j a_j^i h_j^i \tag{8}$$

其中,W_1 和 b_1 是学习参数。

文档的主题信息也是判断实体关系的关键要素,文档的主题信息代表所有实体的上下文信息,蕴含了该文档想要表达的主旨意图。之前的模型大多直接采用句法树或者启发式规则[9,18-19]获得与文档信息相关的上下文信息,但是这种操作严重破坏了文档的原始结构,不利于实体间关系的推理。因此本文模型采用自注意力机制来获取文档的主题信息表示 c_n:

$$c_n=\mathrm{softmax}\left(\frac{h_j\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{9}$$

其中,$\boldsymbol{K}$ 和 $\boldsymbol{V}$ 是输入文档转化而来的键和值矩阵。

接下来需要将第 i 个实体节点的表示与所在句子的表示以及文档主题信息表示进行连接操作，使得实体表示具备这两者的信息，兼备句内和句间推理的能力。三者连接后的实体节点 u_i 可以表示为：

$$u_i=[e_i;S_i;c_n] \tag{10}$$

考虑到图上节点的信息传播，实体图中的每个节点需要聚合邻居节点的特征，使用 R-GCN 能够实现邻居节点信息的聚合。假定在第 l 层，h_i^l 是实体节点 u_i 的隐藏表示，N_i 是第 i 个实体节点的邻居节点集合。因此第 l 层的实体节点的特征表示可以定义为：

$$h_i^{l+1}=\sigma\left(\sum_{j\in\mathcal{N}_i}\frac{1}{|\mathcal{N}_i|}W_1^l h_j^l+W_2^l h_i^l\right) \tag{11}$$

其中，σ 是激活函数，W_1^l 和 W_2^l 是学习参数。

式(11)得到的是第 i 个实体节点在 l 层的特征表示，为了能得到该实体节点所有层的特征聚合，需要将每个层的特征表示连接起来，得到实体节点 u_i 的最终表示。

$$e_i=[h_i^{(0)};h_i^{(1)};h_i^{(2)};\cdots;h_i^{(N)}] \tag{12}$$

当 $l=0$ 时，$h_i^{(0)}$ 是实体节点 u_i 在第 0 层的隐藏表示，也就是其初始表示。

5 分类层

在经过 N 次卷积网络的聚合之后，我们可以得到图上所有实体节点的表示。对于预测两个实体的语义关系，从本质上来说可以看成一个多标签分类任务；对于每个目标实体对所对应的关系标签，需通过[0,1]之间的概率值来判断，那么 sigmoid 函数可以将标签概率控制在[0,1]之间，并且概率之间互不影响，而 softmax 函数常用于多分类，即在实体关系标签之间相互影响，因此我们使用 sigmoid 函数来计算实体对的每一种关系类型的概率。

$$P(r|e_i,e_j)=\text{sigmoid}(e_i^{\mathrm{T}}W_r e_j+b_r) \tag{13}$$

其中，W_r 和 b_r 是学习参数。

对于多标签分类任务，实体对所属的每一关系标签都可以被看成二分类问题，式(13)输出的是预测关系标签结果，当 $r=1$ 时，表示实体关系预测正确的情况，在此情况下预测正确的概率为 $P(r|e_i,e_j)$，反之 $r=0$ 时，关系预测错误的概率为 $1-P(r|e_i,e_j)$。那么可以使用二元交叉熵定义损失函数。

$$\mathcal{L}=-\sum_{r\in\mathcal{R}}\mathbb{I}(r=1)\log P(r|e_i,e_j)+\mathbb{I}(r=0)\log(1-P(r|e_i,e_j)) \tag{14}$$

其中，$\mathcal{R}$ 是关系类型集合，$\mathbb{I}(\cdot)$ 是指示函数。

6 实验

6.1 实验数据和评估

本文模型在 DocRED[7] 数据集上进行实验，DocRED 包含人工标注和远程监督的数据集，考虑到数据集的质量，本文实验只在人工标注数据集上进行。人工标注的 DocRED 数据集中包含了 5053 个文档，其中训练集包含 3053 个文档，验证集和测试集分别包含 1 000 个文档，并且拥有 96 种关系类型。

我们采用常用的 F1 值、Ign F1 值以及 AUC 值来评估本文模型。Ign F1 值是由 Yao 等[7] 提出的，其通过排除训练集、验证集和测试集共享的关系事实，使性能指标更为客观。

6.2 实验设置

模型基于 PyTorch 实现，在 NVIDIAGeForce GTX TITAN X GPU 服务器上运行调试，使用 BERT 的 BERT-Base (uncased)版本作为编码器。使用 Adam 优化器来训练模型，在不断调整参数的过程中可以发现，在三层图卷积神经网络条件下，设置丢弃率为 0.6、学习率为 5×10^{-5}、进行 100 轮次的训练能够得到训练最优的训练模型。具体参数设置如表 1 所列。

表 1 超参数设置

Table 1 Hyper-parameter settings

Hyper-parameter	Value
Batch Size	12
Encoder Hidden Size	768
Layers of GCN	3
Dropout	0.6
Optimizer	Adam
Learning Rate	5×10^{-5}
Weight Decay	0.0001
Epoch	100

6.3 基准模型

为了展示本文模型的性能，将其与以下文档级关系抽取的基准模型进行对比：

(1)LSR-BERT：Nan 等[10] 提出的潜在结构细化(Latent Structure Refinement，LSR)模型能够动态学习文档图结构，增量地捕获全局信息以完成对实体间关系的预测。

(2)HeterGSAN-BERT：Xu 等[13] 提出的显式判别推理框架抽取了 3 种不同的推理路径(句内推理、逻辑推理和共指推理)，通过比较和估计不同推理路径的概率来识别实体间的关系。

(3)BiLSTM-Multi-GCN：Wu 等[20] 提供的融入上下文的卷积模型和多头图注意卷积模型在一定程度上解决了实体长距离依赖以及特征区分问题。

(4)BSRU-ATTCapsNet：Yang 等[21] 提出的融合双向简单循环与胶囊模型引入注意力机制来尝试解决远距离依赖问题，提高了实体关系预测的精度。

(5)ESA-BERT：Yuan 等[22] 发现判断实体间的关系、充分利用句子和文档的多层次信息尤为重要，其通过设计门控机制来引导句子和文档特征，提升了预测性能。

6.4 实验结果

本文模型在 DocRED 人工标注的数据集上进行实验，实验结果如表 2 所列，可以发现 BERT 编码器的引入大幅度提升了文档级关系抽取的性能，主要原因在于 BERT 编码器可以隐式地解决实体长距离依赖问题。

本文模型实现了较好的性能。具体来说，在采用同样的 BERT 编码器下，本文模型在验证集上的 F1 值比目前最好的模型高出 0.53，并且我们也进行了 ACU 值的对比实验来考察模型对不均衡的正负样本是否敏感。通过实验发现，本文模型的 AUC 值比 ESA-BERT[22] 模型高出 2.57，反映了所提模型对 DocRED 中正负样本的综合预测能力高于现有模型。本文模型分别采用隐式的 BERT 和显式的图结构框架来

解决长距离依赖问题,且实体节点的构造方式也对提升性能起到了关键作用,使模型不仅能够处理文档中的句内实体关系,还可以妥善处理更为棘手的句间实体关系,这在接下来的实验中更能得到直接的证明。

表 2 DocRED 数据集上的实验结果

Table 2 Experiment results on DocRED

Model	Dev			Test	
	*Ign F*1	*F*1	*AUC*	*Ign F*1	*F*1
LSR-BERT[10]	52.43	59.00	—	56.97	59.05
BiLSTM-Multi-GCN[20]	—	55.43	—	—	—
BSRU-ATTCapsNet[21]	—	58.07	56.18	—	56.62
ESA-BERT[22]	56.20	58.28	56.36	55.71	58.04
HeterGSAN-BERT[13]	58.13	60.18	—	57.12	59.45
Our Model	58.76	60.71	58.93	59.03	61.28

6.5 消融实验

为了深入探究模型的有效性,我们在 DocRED 的验证集上进行消融实验,将模型分为 3 个关键构件,然后依次使某一个构件起作用,这样就可以观察模型中每个构件对性能提升的作用。按照这个思路,我们将模型分成句子上下文表示、文档主题信息表示和 R-GCN 节点聚合 3 个关键构件。观察表 3 的实验结果可知,在实体表示中去掉句子上下文表示的部分时,模型的性能明显下降,主要原因是 DocRED 数据集中存在接近 60%的句内关系,理解整个句子所要表达的语义是模型在句内关系抽取中的关键,而我们提出的句子上下文表示也就是针对句内关系抽取所给出的解决方案,因此,去掉此模块后模型的性能显著下降。当去掉文档主题信息表示时,模型性能下降的幅度小于句子上下文表示的部分,这也印证了在 DocRED 数据集中至少有 40.7%的关系需要跨句子得到,虽然下降的幅度不如句子上下文表示大,但也说明了文档主题信息对实体的句间关系预测是非常必要的,而仅仅通过句子内的细粒度信息无法从整体上感知句间关系。当去掉 R-GCN 节点聚合部分时,F1 值下降了 1.29,说明目标实体节点对关系的预测也需要其他节点信息的支持,R-GCN 有效聚合了邻居节点的信息,使实体节点的信息表示更加有效且丰富。

表 3 消融实验

Table 3 Ablation experiment

Model	*F*1
Our Model	60.71
-Sentence context representation	58.94
-Document topic representation	59.28
-R-GCN node aggregation	59.42

6.6 实体距离实验

为了进一步观察在不同实体距离下模型的抽取性能,本文模型在 DocRED 的验证集上与 Zhang 等[5]提出的图卷积网络扩展模型、Christopoulou 等[9]提出的面向边缘模型,以及 Wang 等[23]提出的全局-局部神经网络模型进行对比实验。在实体距离的实验中,距离可以通过 DocRED 中"VertexSet"的"pos"字段得到,得到全部实体对之间的距离后需要对其进行分组,首先将数据集中的两个实体按照两者的距离 dist 划分为距离为 0、距离为 1~2、距离大于或等于 3 这 3 种情况。通过对数据集进行预处理得到增加距离字段后的数据,按照距离不同分别进行实验,实验结果如图 3 所示。在 3 种不同的实体距离下,所提模型的 F1 值较之前的模型均有所提高,这主要得益于我们专门设计的句子上下文表示和文档主题信息表示技术。

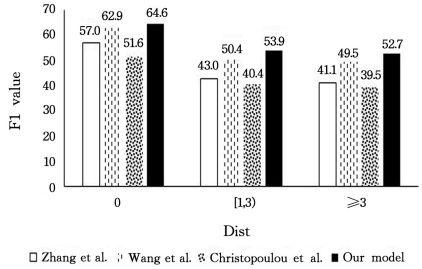

图 3 实体距离实验结果

Fig. 3 Experiment results of entity distance

结束语 本文提出了增强实体表示的文档级关系抽取模型,通过多角度的对比实验可以表明,我们的模型在处理实体间复杂交互的上下文上有着明显的优势。在未来的研究中,我们计划将概率知识图谱和实体表示进行有机的结合,概率知识图谱作为不确定性的知识,能够将实体概念更抽象地表达出来,这样目标实体对能够通过抽象概念先行判断是否存在关系,通过这种方式来提升文档级关系抽取的性能。

参 考 文 献

[1] SPEER R,CHIN J,HAVASI C. Conceptnet 5.5:An open multilingual graph of general knowledge[C]// Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017:4444-4451.

[2] YU M,YIN W,HASAN K S,et al. Improved neural relation detection for knowledge base question answering[C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017:571-581.

[3] KADRY A,DIETZ L. Open relation extraction for support passage retrieval:merit and open issues[C]// Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. 2017:1149-1152.

[4] MIWA M,BANSAL M. End-to-end relation extraction using lstms on sequences and tree structures[C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016:1105-1116.

[5] ZHANG Y,QI P,MANNING C D. Graph convolution over pruned dependency trees improves relation extraction[C]// Proceedings of the 2018 Conference on Empirical Methods in Natu-

ral Language Processing. 2018:2205-2215.
[6] GUO Z, ZHANG Y, LU W. Attention guided graph convolutional networks for relation extraction[C]// Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019:241-251.
[7] YAO Y, YE D, LI P, et al. DocRED: Alarge-scale document-level relation extraction dataset [C]// Proceedings of the 57th Conference of the Association for Computational Linguistics. 2019: 764-777.
[8] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling relational data with graph convolutional networks[C]// European Semantic Web Conference. 2018:593-607.
[9] CHRISTOPOULOU F, MIWA M, ANANIADOU S. Connecting the dots: Document-level neural relation extraction with edge-oriented graphs[C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019:4924-4935.
[10] NAN G, GUO Z, SEKULIC I, et al. Reasoning with latent structure refinement for document-level realtion extraction[C]// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020:1546-1557.
[11] KIPF T, WELLING M. Semi-supervised classification with graph convolutional networks[C]// Proceedings of the 5th International Conference on Learning Representations. 2017.
[12] ZENG S, WU Y, CHANG B. Sire: Separate intra-and inter-sentential reasoning for document-level relation extraction[C]// Findings of the Association for Computational Linguistics. 2021:524-534.
[13] XU W, CHEN K, ZHAO T. Document-level relation extraction with reconstruction [C]// Proceedings of the 33th Conference on Artificial Intelligence. 2020:14167-14175.
[14] 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. 2019:4171-4186.
[15] TANG H, CAO Y, ZHANG Z, et al. HIN: Hierarchical inference network for document-level relation extraction[C]// Proceedings of the 24th Pacific-Asia Conference. 2020:197-209.
[16] ZENG S, XU R, CHANG B, et al. Double graph based reasoning for document-level relation extraction[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020:1630-1640.
[17] JIA R, WONG C, POON H. Document-level n-ary relation extraction with multiscale representation Learning[C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. 2019:3693-3704.
[18] TAO Q, LUO X, WANG H, et al. Enhancing relation extraction using syntactic indicators and sentential contexts[C]// 2019 IEEE 31st International Conference on Tools with Artificial intelligence. 2019:1574-1580.
[19] HIRANO T, ASANO H, MATSUO Y, et al. Recognizing relation expression between named entities based on inherent and context-dependent features of relational words[C]// Proceedings of the 23th International Conference on Computational Linguistics. 2010:409-417.
[20] WU T, KONG F. Document-level relation extraction based on graph attention convolutional neural network[J]. Journal of Chinese Information Processing, 2021, 35(10):73-80.
[21] YANG C N, PENG D L. Document-level entity relation extraction method integrating bidirectional simple recurrent unit and capsule network [J]. Journal of Chinese Computer Systems, 2022, 43(5):964-968.
[22] YUAN C, HUANG H, FENG C, et al. Document-level relation extraction with entity-selection attention [J]. Information Sciences, 2021, 568:163-174.
[23] WANG D, HU W, CAO E, et al. Global-to-local neural networks for document-level relation Extraction[C]// Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. 2020:3711-3721.

DING Xiaoyao, born in 1990, Ph.D. His main research interests include relation extraction and knowledge graph.

ZHOU Gang, born in 1974, Ph.D, professor. His main research interests include big data, knowledge graph and data mining.

(责任编辑:何杨)