doi:10.3969/j.issn.1673-9833.2020.04.010

基于 BERT 和 TextRank 关键词提取的实体链接方法

詹 飞[1,2]，朱艳辉[1,2]，梁文桐[1,2]，冀相冰[1,2]

（1. 湖南工业大学 计算机学院，湖南 株洲 412007；
2. 湖南工业大学 智能信息感知及处理技术湖南省重点实验室，湖南 株洲 412007）

摘 要：提出一种基于 BERT（bidirectional encoder representations from transformers）和 TextRank 关键词提取的实体链接方法。将 BERT 预训练语言模型引入实体链接任务，进行实体指称上下文和候选实体相关信息的关联度分析，通过提升语义分析的效果来增强实体链接的结果。采用 TextRank 关键词提取技术增强目标实体综合描述信息的主题信息，增强文本相似度度量的准确性，从而优化模型效果。使用 CCKS2019 评测任务二的数据集对模型效果进行验证，实验结果表明，所提方法的实体链接效果明显优于其他实体链接方法，能有效解决实体链接问题。

关键词：实体链接；BERT 预训练语言模型；语义分析；TextRank；关键词提取

中图分类号：TP391 **文献标志码**：A **文章编号**：1673-9833(2020)04-0063-08

引文格式：詹 飞，朱艳辉，梁文桐，等 . 基于 BERT 和 TextRank 关键词提取的实体链接方法 [J]. 湖南工业大学学报，2020，34(4)：63-70.

Entity Linking Via BERT and TextRank Keyword Extraction

ZHAN Fei[1,2]，ZHU Yanhui[1,2]， LIANG Wentong[1,2]，JI Xiangbing[1,2]

（1. College of Computer Science，Hunan University of Technology，Zhuzhou Hunan 412007，China；
2. Hunan Key Laboratory of Intelligent Information Perception and Processing Technology，
Hunan University of Technology，Zhuzhou Hunan 412007，China）

Abstract：This paper proposes a method based on BERT (bidirectional encoder representations from transformers) and TextRank keyword extraction for entity linking. BERT pre-training language model is introduced into the entity linking task for an analysis of the correlation between entity reference context and related information of candidate entity, thus enhancing the result of entity linking by improving the effect of semantic analysis. By using TextRank keyword extraction, an enhancement can be achieved of the subject information of the comprehensive description information of the target entity, with the accuracy of text similarity measurement increased, and the effect of the model optimized as well. Based on the verification of the model effect by the data set of CCKS2019 evaluation task II, the experimental results show that the proposed method, which can effectively solve the entity linking problem, is

收稿日期：2019-10-10

基金项目：科技创新 2030—“新一代人工智能”基金资助重大项目（2018AAA0100400），国家自然科学基金资助项目（61702177），湖南省自然科学基金资助项目（2018JJ2098，2020JJ6089），湖南省教育厅基金资助重点项目（19A133）

作者简介：詹 飞（1993-），男，河南三门峡人，湖南工业大学硕士生，主要研究方向为自然语言处理与知识工程，E-mail：shufeixue@163.com

通信作者：朱艳辉（1968-），女，湖南湘潭人，湖南工业大学教授，主要从事自然语言处理与知识工程方面的教学与研究，E-mail：swayzhu@163.com

characterized with an entity linking effect which is significantly superior to that of other entity linking methods.

Keywords：entity linking；BERT pre-training language model；semantic analysis；TextRank；keyword extraction

1 研究背景

近年来，大规模中文通用知识图谱的发展给国内人工智能领域的发展带来了新的机遇。实体链接作为命名实体识别任务的后续任务，是知识图谱构建和补全过程中的关键一环。实体链接任务的目标是将文本中识别的实体指称和该实体指称在给定知识库中对应的实体相关联，通常可以将实体链接分解为两个串行的子任务：候选实体生成和候选实体排序。候选实体生成阶段为当前实体指称过滤掉知识库中的大部分不相关实体，得到候选实体集。候选实体集中通常包含多于一个候选实体，在候选实体排序阶段对候选实体集中的实体和当前实体指称进行相似度打分并排序，得分最高的实体即为当前实体指称的目标链接实体。实体链接任务的关键挑战即为如何有效利用实体指称和候选实体的相关信息来对二者进行相似度打分。

现有实体链接工作的重点集中在候选实体排序阶段。随着深度学习的发展，深度学习技术被广泛地应用到自然语言处理领域的多项任务中，并取得了很好的效果。针对实体链接任务，He Z. Y. 等 [1] 提出一种基于深度神经网络（deep neural networks，DNN）的方法来进行实体链接，通过深度神经网络自主学习实体和上下文的特征表示，端到端地进行实体链接，避免了人工设计特征，当时在两个公开实体链接数据集上取得了最先进的性能。M. Francis-Landau 等 [2] 使用卷积神经网络（convolutional neural networks，CNN）来捕获实体指称上下文和目标实体上下文的语义信息，并利用多个粒度的卷积来比较两者之间的语义相似度。T. H. Nguyen 等 [3] 提出结合循环神经网络和卷积神经网络的联合模型来同时获取实体指称上下文局部特征和全局主题特征，用卷积神经网络获取局部相似性，用循环神经网络获取全局一致性，该模型在多个数据集上被证明是有效的。Liu C. 等 [4] 提出一种新型的注意力机制来获取给定实体指称周围重要的文本，并且结合一种前向－后向算法获取文本主题信息来提高实体链接的准确率。Hu S. Z. 等 [5] 提出具有双重注意力机制的对称 Bi-LSTM（bidirectional long short-term memory）模型，该模型能有效利用结构信息和注意力机制更全面地提取实体特征，并结合上下文特征和结构特征作为实体的特征表示。

预训练语言模型出现之前，使用深度学习方法解决自然语言处理问题的研究思路，大多是针对特定的目标任务来设计对应的模型。BERT（bidirectional encoder representations from transformers）出现之前，已经有了一些专家学者对预训练语言模型进行了相关研究工作，如 ULMFiT（universal language model fine-tuning）[6] 和 OpenAI GPT[7] 模型，但由于单向语言模型的限制，它们不能对上下文语义信息进行充分利用。J. Devlin 等 [8] 对现有预训练语言模型 [7] 进行改进，提出新的预训练语言模型 BERT，目前，该模型在许多下游任务上取得了较优效果。本研究将 BERT 引入实体链接任务中，将预训练的 BERT 语言模型作为实体链接模型的一部分。

关键词能够反映出文本主题信息，强化文本相似度比较的效果。将关键词提取技术加入到实体链接过程中，辅助进行实体指称和候选实体相关信息的相似度比较，能够增强文本相似度度量的准确性，从而优化模型效果。TextRank 关键词提取算法将关键词提取问题转化到图模型中进行处理，能够考虑到相邻词的语义关系，提取出的关键词能够更好地反映文本的主题信息。因此，本文将 TextRank 关键词提取算法融合到实体链接过程中。

基于 BERT 模型的实体链接方法在 NLP（natural language processing）任务上的优秀表现和关键词提取对文本相似度比较的强化效果，本文提出一种基于 BERT 和 TextRank 关键词提取的实体链接模型。该模型的特点是将 BERT 预训练语言模型引入实体链接任务，通过 BERT 来获取句子的向量表示，从而进行实体指称和候选实体相关信息的关联度分析。同时，使用 TextRank 关键词提取技术来获得目标实体描述文本的关键词，作为目标实体综合描述的一部分，输入到 BERT 中，这能够增强目标实体综合描述的主题信息，从而优化模型的效果。

2 BERT-TextRank 模型

本研究提出基于 BERT 和 TextRank 关键词提取的深度神经网络模型进行实体链接，模型整体结构

如图 1 所示，主要包括 TextRank 关键词提取部分、BERT 层和输出层。

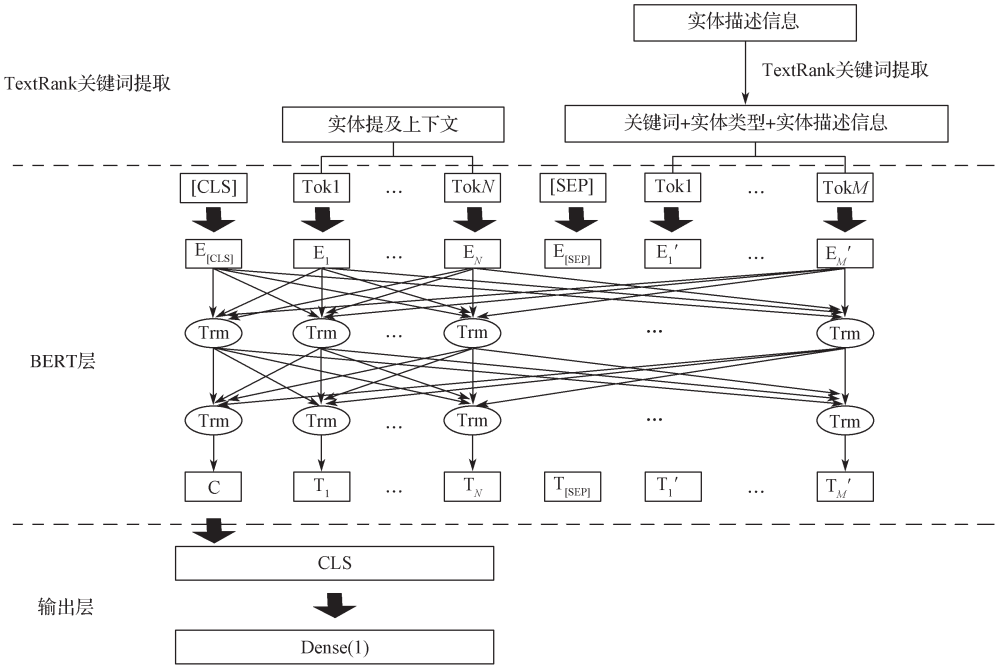


图 1　基于 BERT 和 TextRank 关键词提取的实体链接网络模型

Fig. 1　Entity linking model based on BERT and TextRank keyword extraction

将实体指称上下文和候选目标实体的综合描述用 [SEP] 分隔符隔开作为 BERT 的输入，实体指称上下文为当前实体指称所在的句子，候选目标实体的综合描述由关键词、实体类型和实体描述信息组成。关键词由实体描述信息通过 TextRank 关键词提取得到，实体类型和实体描述信息从目标知识库中获取。然后取 BERT 输出中 CLS 位置对应的向量作为下一个全连接层的输入，使用 sigmoid 函数进行激活，把文本语义相似性问题抽象为二分类问题。

2.1　BERT 预训练语言模型

BERT 模型结构如图 2 所示。

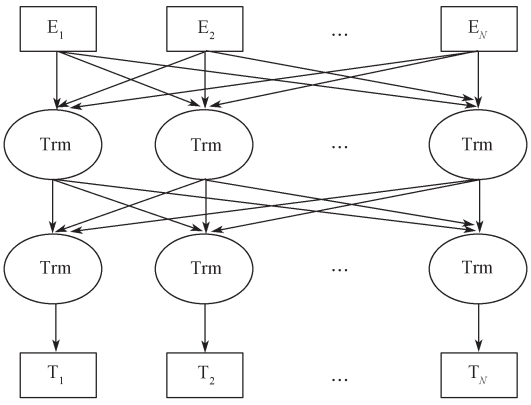


图 2　BERT 模型结构图

Fig. 2　BERT model structure illustration

图 2 所示模型借鉴了 A. Vaswani 等 [9] 提出的“多层双向 Transformer 编码器”思想，以双向 Transformer 的 Encoder 作为模型的基本组成单元。

BERT 模型虽然和之前的预训练语言模型 OpenAI GPT 一样都使用了 Transformer，但不同的是 OpenAI GPT 模型使用的是单向的注意力机制，BERT 模型则针对这一不足进行了改进，使用双向 Transformer 的 Encoder 作为基本组成单元，BERT 的这种结构能够联合所有层中的左右两个方向的上下文信息进行训练。

BERT 模型使用的 Transformer 基于多头注意力机制（multi-head attention）。多头注意力机制的结构如图 3 所示。

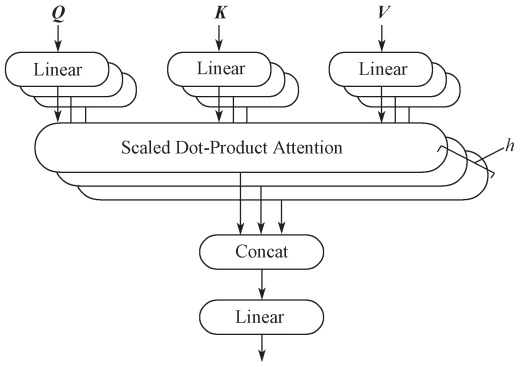


图 3　多头注意力机制结构图

Fig. 3　Multi-head attention structure mechanism

由图 3 的结构形式可知，多头注意力机制可以帮

助模型捕获更多层面的语义特征，将各个注意力头单独进行计算，然后将其结果进行拼接，得到最终结果。

$$head_i=Attention(\boldsymbol{Q}\boldsymbol{W}_i^{\boldsymbol{Q}}, \boldsymbol{K}\boldsymbol{W}_i^{\boldsymbol{K}}, \boldsymbol{V}\boldsymbol{W}_i^{\boldsymbol{V}}), \quad (1)$$

$$MultiHead(\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V})= Concat(head_1, head_2, \cdots, head_h)\boldsymbol{W}^{O}。 \quad (2)$$

式（1）~（2）中：$\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 为输入量；

$\boldsymbol{W}$ 为变换参数。

对多头注意力的输入量 $\boldsymbol{Q}, \boldsymbol{K}, \boldsymbol{V}$ 分别进行线性变换，每次线性变换的参数 $\boldsymbol{W}$ 取值不同，分别为 $\boldsymbol{W}_i^{\boldsymbol{Q}}$，$\boldsymbol{W}_i^{\boldsymbol{K}}$ 和 $\boldsymbol{W}_i^{\boldsymbol{V}}$，线性变换得到的结果输入 Scaled Dot-Product Attention 中得到 $head_i$，重复做 h 次；然后将 h 次 Scaled Dot-Product Attention 得到的结果 $head_1, head_2, \cdots, head_h$ 进行拼接，并对拼接的结果进行线性变换，得到多头注意力的最终结果，线性变换的参数为 $\boldsymbol{W}^{O}$。

受 Y. Bengio 等 [10] 研究结论的启发，BERT 的训练方式不同于之前的预训练语言模型，而是通过大量未标注的百科文本语料进行训练，得到预训练语言模型，然后根据具体需要，针对特定目标任务对 BERT 模型进行微调。新的预训练方法也是 BERT 表现优于之前的预训练语言模型的重要因素，它不再采用传统的单向语言模型来进行预训练，而是提出两个新任务来进行预训练，即通过 MLM（masked language model）和“下一句预测”（next sentence prediction）两个新的任务分别捕捉词语和句子级别的特征。

MLM 用来克服之前的预训练语言模型的单向性所具有的局限，对于输入序列中 15% 的数据，随机地将这些输入序列中的一部分单词用 [mask] 标记进行遮盖，然后以预测这些被遮盖的单词为目标来对模型进行训练，这样能够同时在左右两个方向上融合上下文信息。通过 MLM 任务的训练，模型能够同时对左右两侧的语义特征进行提取，通过联合所有层中的左右两个方向的上下文信息进行训练，得到深度双向 Transformer 转换。但是用于遮蔽单词的特殊标记 [mask] 在实际的 NLP 任务中并不存在，用从语料中随机获取的词和预测位置的原词按照一定比例对需要 [mask] 遮蔽的词进行替换，从而可以保证训练过程和实际任务保持一致。用特殊标记“[mask]”来替换 80% 的目标单词，用从语料中随机获取的一个词来替换 10% 的目标单词，剩余 10% 的目标单词不进行任何操作。

对于“下一句预测”任务捕捉词语和句子级别的特征，是为了让模型能够更好地捕捉句子级别的语义特征。每条训练数据为连续的两个句子 M 和 N，概率为 50% 的句子 N 是原文中的正确句子，概率为 50% 的句子 N 会被替换为语料中的一条随机语句来作为负样本进行训练，然后再做二分类来判断输入的句子 N 是正确的还是随机产生的。

2.2 TextRank 关键词提取

使用 TextRank 算法进行关键词提取的思路是将关键词提取问题转化到图模型中进行处理，这样能够考虑到相邻词的语义关系。使用 TextRank 算法提取得到的关键词能够增强句子的主题信息，从而优化文本相似度度量的效果。

TextRank 算法是以 PageRank 算法为蓝本，针对自然语言处理的特点进行修改而形成的。使用 TextRank 算法进行关键词提取的思路是将关键词提取问题转化到图模型中进行处理，这样能够考虑到相邻词语的语义关系。并根据各个词之间的相互联系判断其对于文本整体重要性的高低，得到各个词的重要性得分，然后根据其得分从高到低进行排序，设定阈值 H，重要性得分较高的 H 个词即可视为提取出来的文本关键词。将文本看成是句子集合 $T=\{S_1, S_2, \cdots, S_n\}$，其中的每个句子 $S_i \in T$，又可以看作词的集合 $S_i=\{N_1, N_2, \cdots, N_m\}$，构建图模型 $G=(V, E)$，其中 $V=S_1 \cup S_2 \cup \cdots \cup S_n$，当两个词共同出现在一个句子中时，对应的节点有边，否则无边。词的重要性得分计算方法如下：

$$score(N_i)= (1-d)+d\times \sum_{j\in In(N_i)} \frac{1}{\left|Out(N_j)\right|} score(N_j), \quad (3)$$

式中：$In(N_i)$ 是指向节点 i 的节点集合；

$Out(N_j)$ 是节点 j 指向的节点组成的集合；

d 为阻尼系数；根据实际情况对阻尼系数进行赋值，通常取 0.85。

在使用 TextRank 进行关键词提取时，以词为节点，以共现关系建立节点之间的链接来进行图模型的构建。这里的图模型与 PageRank 模型不同的是，PageRank 构建的是有向图，而 TextRank 构建的图是无向图。首先对图中的每个节点指定任意初始值，然后进行迭代训练直至收敛，这样就能够计算出各节点的最终权重。

3 实验与结果分析

3.1 实验数据

本研究采用 CCKS2019（2019 全国知识图谱与语义计算大会）任务二提供的训练语料和知识库 [11-12]。训练语料中每条数据包含一条文本和该文本中包含

的实体指称，以及各个实体指称在给定知识库中对应的目标实体。知识库中包含每个实体的别名、实体类别和实体描述信息。本研究仅评价数据集中的非“NIL”型实体指称，即在目标知识库中存在链接实体的实体指称。

训练语料由训练集和验证集组成，其中训练集包括 9 万条短文本标注数据，验证集包括 1 万条短文本标注数据，数据通过百度众包标注生成。标注数据集主要来自于真实的互联网网页标题数据，这些标题数据来源于用户检索 Query 对应的有展现及点击的网页，短文本平均长度为 21.73 个中文字符，覆盖了不同领域的实体，如人物、电影、电视、小说、软件、组织机构、事件等。

3.2　评价指标

实体链接评价指标选用精确率 P、召回率 R、F 值（F-score），具体说明如下：

给定输入文本集 Q，对于 Q 中每条输入文本 q，此输入 q 中有 N 个实体指称即，$M_q=\{m_1, m_2, m_3, \cdots\}$，每个实体指称链接到知识库的实体编号为 $E_q=\{e_1, e_2, e_3, \cdots\}$，实体链接系统输出的链接结果为 $E_q'=\{e_1', e_2', e_3', \cdots\}$，则实体链接的准确率 P，召回率 R 和 F 值定义如下：

$$\begin{cases} P=\dfrac{\sum_{q\in Q}\left|E_q\cap E_q'\right|}{\sum_{q\in Q}\left|E_q'\right|}, \\ R=\dfrac{\sum_{q\in Q}\left|E_q\cap E_q'\right|}{\sum_{q\in Q}\left|E_q\right|}, \\ F=\dfrac{2PR}{P+R}。\end{cases} \tag{4}$$

3.3　实验环境

本研究中的软硬件实验环境如下：操作系统为 Ubuntu16.04，GPU 显卡为 NVIDIA RTX 2080Ti(11 GB)，python 版本为 3.6，tensorflow 版本为 1.12.0，内存为 16 GB，硬盘容量为 1 TB。

3.4　参数设置

本研究所使用的 BERT 为包含 12 层的 Transformer 的 BERTBASE，学习率为 1e-5，最大序列长度为 512，训练 batch_size 为 4。

3.5　实验结果

为了验证本研究中所提出的基于 BERT 和 TextRank 关键词提取的实体链接方法的有效性，本研究复现了经典的句子语义建模方法，如 TextRNN 和 TextRNN 方法，用 TextRNN[13] 和 TextRCNN[14] 进行实体链接，与本文中的 BERT-TextRank 模型类似，在这两个模型中都将实体链接中的文本语义相似性问题抽象为二分类问题进行处理。分别将 TextRNN 和 TextRCNN 模型与 TextRank 关键词提取算法相结合，然后进行对比实验。

3.5.1　关键词个数 K 的取值实验

分别使用 TextRNN-TextRank、TextRCNN-TextRank 和 BERT-TextRank 3 组模型进行实体链接实验，TextRNN-TextRank 表示将 TextRNN 模型和 TextRank 关键词提取算法进行结合，其他两个模型名称含义与其类似。K 值表示 TextRank 算法提取的关键词个数，以步长为 1，在区间 [0, 6] 内对参数 K 做取值实验，K 值为 0 时表示不进行关键词提取。随着 K 值的取值变化，上述 3 个模型的实体链接效果如表 1 所示。

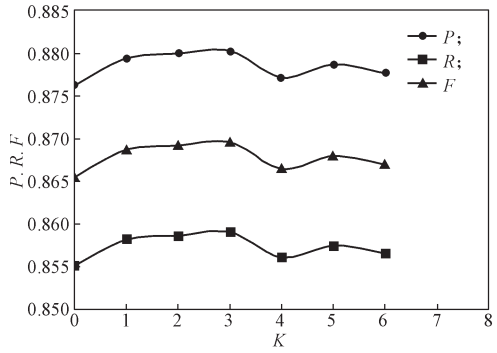
表 1　关键词个数 K 取值实验结果

Table 1　Experimental results corresponding with values of keywords K

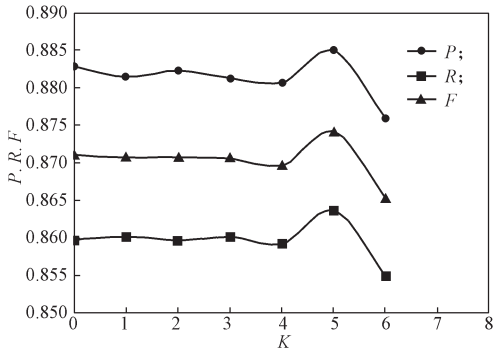
模　型	参数	K=0	K=1	K=2	K=3	K=4	K=5	K=6
TextRNN-TextRank	P	0.876 3	0.879 4	0.880 0	0.880 3	0.877 2	0.878 7	0.877 7
	R	0.855 2	0.858 2	0.858 7	0.859 1	0.856 1	0.857 5	0.856 6
	F	0.865 6	0.868 7	0.869 2	0.869 6	0.866 5	0.868 0	0.867 0
TextRCNN-TextRank	P	0.882 9	0.881 5	0.882 3	0.881 3	0.880 8	0.884 9	0.876 0
	R	0.859 7	0.860 2	0.859 6	0.860 1	0.859 2	0.863 6	0.854 9
	F	0.871 1	0.870 7	0.870 8	0.870 6	0.869 7	0.874 1	0.865 3
BERT-TextRank	P	0.884 1	0.881 7	0.884 2	0.885 3	0.883 8	0.884 8	0.884 1
	R	0.897 6	0.895 2	0.897 8	0.898 8	0.897 3	0.898 3	0.899 5
	F	0.890 8	0.888 4	0.890 9	0.892 0	0.890 5	0.891 5	0.891 7

各模型的 TextRank 关键词个数 K 调节实验结果如图 4 所示。

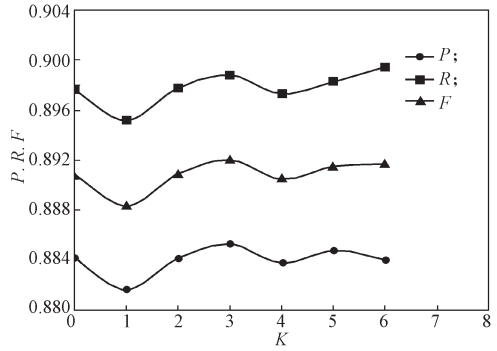
分析对比图 4a、b、c 的实验结果表明，结合 TextRank 关键词提取算法后，3 个模型的实体链接效果都有所提高，且本文提出的 BERT-TextRank 方法的实验效果优于其他两个模型实验结果。TextRNN-TextRank 和 BERT-TextRank 模型在关键词个数 K=3 时 F 值达到最大，而 TextRCNN-TextRank 模型在关键词个数 K=5 时 F 值达到最大值。



a）TextRNN-TextRank



b）TextRCNN-TextRank



c）BERT-TextRank

图 4 关键词个数 *K* 取值实验结果

Fig. 4 Experimental results corresponding with values of keyword *K*

当在 *F* 值达到峰值后继续增加关键词个数会导致主题信息比较分散，从而导致 *F* 值有所降低。这说明利用 TextRank 模型提取关键词，从而增强知识库中实体描述文本的主题信息，对于实体链接是有效的，但是不同模型对于关键词个数的敏感性不同，模型 *F* 值取得峰值时对应关键词个数 *K* 也并不完全相同。因此，接下来关键词个数 *K* 分别选取各个模型的最佳值进行对比实验，即 *K* 值分别选取 3, 5, 3。

3.5.2 相似度阈值 *Y* 取值实验

分析实验结果发现，存在一部分实体指称在目标知识库中对应的候选实体集合不为空，但是候选实体集合中不存在正确的目标实体，即知识库中没有该实体指称对应的实体，导致错误链接。

模型的输出层为全连接层，使用 sigmoid 函数进行激活，把文本语义相似性问题抽象为二分类问题进行处理。将模型输出值记为 *y*，*y* 即为实体指称链接到当前目标实体的概率，也是实体指称上下文和当前目标实体综合描述信息的相似度得分。设定相似度阈值 *Y*，对其定义如下：

$$\begin{cases}\text{将实体指称链接到当前实体}, y \geqslant Y;\\ \text{将实体指称链接为“NIL”}, y < Y。\end{cases} \quad (5)$$

当候选实体上下文与目标实体特征描述的相似度得分 *y* 大于阈值 *Y* 时，将实体指称链接到当前实体；当 *y* 小于阈值 *Y* 时，即认为知识库中不存在此实体指称的目标链接实体，将其链接目标标记为“NIL”。

由上述实验确定 TextRNN-TextRank 模型的参数 *K*=3，TextRCNN-TextRank 模型的参数 *K*=5，BERT-TextRank 模型的参数 *K*=3 后，对阈值 *Y* 进行取值实验，实验区间设置为 [0, 0.5]，以步长为 0.1 进行 *Y* 取值实验，其实验结果如表 2 所示。

表 2 阈值 *Y* 取值实验结果

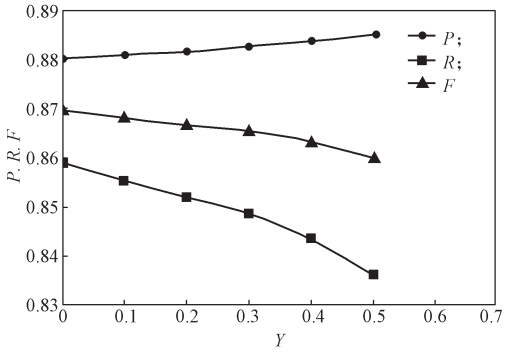
Table 2 Threshold *Y* experiments results

模 型	参数	*Y*=0	*Y*=0.1	*Y*=0.2	*Y*=0.3	*Y*=0.4	*Y*=0.5
TextRNN-TextRank	*P*	0.880 3	0.881 0	0.881 6	0.882 7	0.883 8	0.885 1
	R	0.859 1	0.855 4	0.852 1	0.848 7	0.843 5	0.836 2
	F	0.869 6	0.868 0	0.866 6	0.865 4	0.863 2	0.860 0
TextRCNN-TextRank	*P*	0.884 9	0.886 6	0.887 6	0.889 5	0.892 2	0.895 6
	R	0.863 6	0.859 0	0.853 5	0.846 2	0.836 3	0.823 3
	F	0.874 1	0.872 6	0.870 2	0.867 3	0.863 4	0.857 9
BERT-TextRank	*P*	0.885 3	0.891 0	0.894 4	0.897 6	0.901 1	0.901 5
	R	0.895 5	0.895 4	0.892 2	0.887 8	0.882 6	0.875 7
	F	0.892 0	0.893 2	0.893 3	0.892 7	0.891 7	0.890 1

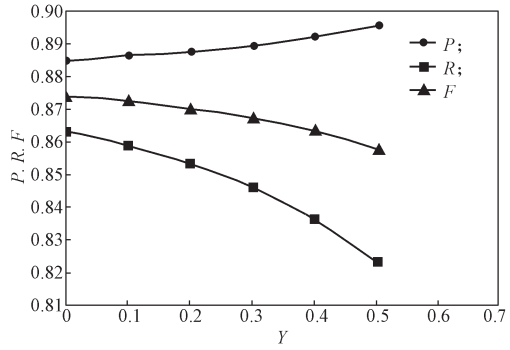
各模型的阈值 *Y* 调节实验结果如图 5 所示。

分析对比图 5a、b、c 的实验结果表明，3 组模型中的 *P* 值均随着 *Y* 值的增大呈上升的变化趋势，而 *R* 值均随着 *Y* 值的增大呈下降的变化趋势，但是 TextRNN-TextRank 模型和 TextRCNN-TextRank 模型的 *P* 值增加幅度不够大，导致其 *F* 值呈单调下降趋

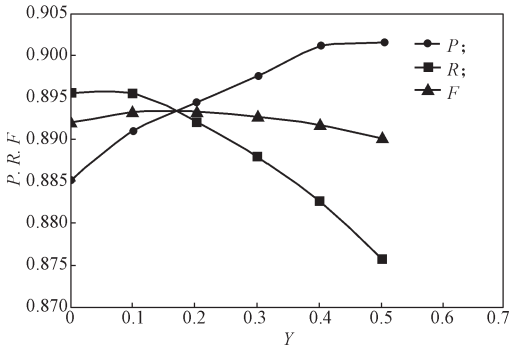
势，BERT-TextRank 模型的 F 值随着阈值 Y 的增大先呈现出上升的变化趋势，在 Y 取 0.2 时其 F 值达到最大值，然后呈下降趋势。证明随着阈值 Y 的增大，正确链接应为“NIL”的实体指称被更多的识别出来，实体链接准确率提高。但是一部分正确链接为非“NIL”的实体指称因为相似度得分相对较低，在阈值 Y 增大的过程中被链接为“NIL”，从而导致实体链接召回率逐渐降低。在 TextRNN-TextRank 模型和 TextRCNN-TextRank 模型中，P 值增大幅度相对较小，而其 R 值也逐渐减小，从而导致其 F 值呈单调下降的变化趋势。因此 TextRNN-TextRank 模型和 TextRCNN-TextRank 模型的阈值 Y 应选择 0，即不设定阈值，但是对于本研究提出的 BERT-TextRank 方法，根据实验结果选定阈值 Y 为 0.2，能够提升模型的实体链接效果。



a）TextRNN-TextRank



b）TextRCNN-TextRank



c）BERT-TextRank

图 5　阈值 Y 取值实验结果

Fig. 5　Experimental results of threshold Y

3.5.3　不同模型对比实验

TextRNN、TextRCNN 和 BERT 三种模型结合关键词提取方法和设定相似度阈值前后对比实验如表 3 所示。

表 3　3 个模型结合关键词提取和阈值控制前后的实验结果

Table 3　Experimental results before and after keyword extraction with three models under threshold control

模　型	无关键词和阈值（K=0，Y=0）			K、Y 取最佳值		
	P	R	F	P	R	F
TextRNN	0.876 3	0.855 2	0.865 6	0.880 3	0.859 1	0.869 6
TextRCNN	0.882 9	0.859 7	0.871 1	0.884 9	0.863 6	0.874 1
BERT	0.884 1	0.897 6	0.890 8	0.894 4	0.892 2	0.893 3

对比分析表 3 的实验数据表明，3 种模型结合 TextRank 关键词提取算法和选定相似度阈值 Y 后的 F 值均比结合之前有所提升，BERT-TextRank 模型相比 TextRNN-TextRank 模型和 TextRCNN-TextRank 模型的 P、R、F 值也有较大提升，有效证明了本研究构建的基于 BERT 预训练语言表征模型和 TextRank 关键词提取的实体链接模型相比较于其他模型的有效性。

4　结语

本研究提出了一种基于 BERT 和 TextRank 关键词提取的实体链接方法。该方法可以分为 TextRank 关键词提取和 BERT 句子相似度比较两部分。TextRank 关键词提取部分用来提取知识库中实体描述文本的关键词来增强文本主题信息，强化文本相似度比较的效果。BERT 句子相似度比较部分将实体指称的上下文和候选实体的特征描述进行相似度比较，候选实体的特征描述由关键词、实体类型和实体描述文本组成，关键词即为 TextRank 提取得到的结果。实验结果证明了本文所提方法的有效性，说明加入主题信息对于文本相似性的度量是有效果的。未来计划借鉴 Liu Y. 等 [15] 提出的结合词嵌入和主题模型的思想，结合主题模型和 BERT 模型，将文本主题信息融合到句子向量表示中来进行文本相似性度量和实体

链接。

参考文献：

[1] 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. Sofia：Association for Computational Linguistics，2013：30−34.

[2] FRANCIS-LANDAU M，DURRETT G，KLEIN D. Capturing Semantic Similarity for Entity Linking with Convolutional Neural Networks[EB/OL]. [2019−06−25]. https://arxiv.org/pdf/1604.00734.pdf.

[3] NGUYEN T H，FAUCEGLIA N，MURO M R，et al. Joint Learning of Local and Global Features for Entity Linking Via Neural Networks[EB/OL]. [2019−06−25]. https://www.aclweb.org/anthology/C16-1218.pdf.

[4] LIU C，LI F，SUN X，et al. Attention-Based Joint Entity Linking with Entity Embedding[J]. Information，2019，10(2)：46. https://doi.org/10.3390/info10020046.

[5] HU S Z，TAN Z，ZENG W X，et al. Entity Linking Via Symmetrical Attention-Based Neural Network and Entity Structural Features[J]. Symmetry，2019，11(4)：453. https://doi.org/10.3390/sym11040453.

[6] HOWARD J，RUDER S. Universal Language Model Fine-Tuning for Text Classification[EB/OL]. [2019−06−30]. https://arxiv.org/abs/1801.06146.pdf.

[7] RADFORD A，NARASIMHAN K，SALIMANS T，et al. Improving Language Understanding by Generative Pre-Training[EB/OL]. [2019−06−30]. https://www.docin.com/p-2176538517.html.

[8] DEVLIN J，CHANG M W，LEE K，et al. BERT：Pre-Training of Deep Bidirectional Transformers for Language Understanding[EB/OL]. [2019−07−11]. https://arxiv.org/pdf/1810.04805.pdf.

[9] VASWANI A，SHAZEER N，PARMAR N，et al. Attention Is All You Need[EB/OL]. [2019−08−03]. https://arxiv.org/pdf/1706.03762.pdf.

[10] BENGIO Y，SCHWENK H，SENÉCAL J S，et al. Neural Probabilistic Language Models[J]. Innovations in Machine Learning，2006，194：137−186.

[11] 冀相冰，朱艳辉，李 飞，等. 基于Attention-BiLSTM的中文命名实体识别[J]. 湖南工业大学学报，2019，33(5)：73−78.
JI Xiangbing，ZHU Yanhui，LI Fei，et al. Entity Recognition of Chinese Names Based on Attention-BiLSTM[J]. Journal of Hunan University of Technology，2019，33(5)：73−78.

[12] 李 飞，朱艳辉，王天吉，等. 基于医疗类别的电子病历命名实体识别研究 [J]. 湖南工业大学学报，2018，32(4)：61−66.
LI Fei，ZHU Yanhui，WANG Tianji，et al. Research on Electronic Medical Record Named Entity Recognition Based on Medical Categories[J]. Journal of Hunan University of Technology，2018，32(4)：61−66.

[13] KIM Y. Convolutional Neural Networks for Sentence Classification[EB/OL]. [2019−08−03]. https://arxiv.org/pdf/1408.5882.pdf.

[14] LAI S，XU L，LIU K，et al. Recurrent Convolutional Neural Networks for Text Classification[EB/OL]. [2019−08−03]. https://dl.acm.org/doi/10.5555/2886521.2886636.

[15] LIU Y，LIU Z，CHUA T，et al. Topical Word Embeddings[EB/OL]. [2019−08−03]. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.699.6286&rep=rep1&type=pdf.

（责任编辑：姜利民）