Computer Science and Application 计算机科学与应用, 2020, 10(12), 2361-2370
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012250

Hans汉斯

基于改进BERT的知识图谱问答研究

易诗玮

同济大学，上海
Email: yishiwei@tongji.edu.cn

收稿日期：2020年11月28日；录用日期：2020年12月23日；发布日期：2020年12月30日

摘　要

基于知识图谱的自动问答是自然语言处理领域的研究热点之一。针对现有的中文开放领域知识库问答，本文将知识图谱问答过程分为实体识别、属性抽取和答案检索三个步骤。首先采用改进BERT结合BiLSTM-CRF的命名实体识别模型来提取问句中的相关实体，然后采用改进BERT结合softmax的分类模型进行属性抽取，最后利用前两步的结果进行答案检索。实验结果显示，该方法在NLPCC-ICCPOL的KBQA数据集上取得了97.54%的F1分数。

关键词

知识图谱问答，命名实体识别，BERT，属性抽取

Research on Knowledge-Based Question Answering Based on Improved Bert

Shiwei Yi

Tongji University, Shanghai
Email: yishiwei@tongji.edu.cn

Received: Nov. 28th, 2020; accepted: Dec. 23rd, 2020; published: Dec. 30th, 2020

Abstract

Automatic Question Answering Based on knowledge map is one of the research hotspots in the field of natural language processing. In this paper, the process of knowledge extraction is divided into three steps: entity recognition, attribute extraction and answer retrieval. Firstly, the named entity recognition model based on improved Bert combined with BiLSTM-CRF is used to extract the related entities in the question, and then the classification model of improved Bert combined with softmax is used for attribute extraction. Finally, the results of the first two steps are used for

文章引用: 易诗玮. 基于改进 BERT 的知识图谱问答研究[J]. 计算机科学与应用, 2020, 10(12): 2361-2370.
DOI: 10.12677/csa.2020.1012250

answer retrieval. The experimental results show that the method achieves 97.54% F1 score on the KBQA dataset of NLPCC-ICCPOL.

Keywords

Knowledge-Based Question Answering, Named Entity Recognition, BERT, Attribute Extraction

1. 引言

随着知识图谱数据规模的爆炸式增长，各个领域的数据用户希望能够准确、快速地获取所需的信息。问答系统作为信息检索的一种形式，它能以简单、准确的自然语言结果回答自然语言问题。问答系统是目前人工智能和自然语言处理领域中一个倍受关注并具有广泛发展前景的研究方向，在开放领域[1]和许多特定领域已经取得了不错的效果，例如生物医学问答系统[2]，地理领域问答系统[3]。按照数据来源的不同，问答系统可以分为基于知识图谱的问答系统，基于阅读理解的问答系统，基于问答对的问答系统，本文研究的内容是基于知识图谱的问答系统。

问答系统的主流研究方法可以分为三类：1) 基于语义解析的方法。该方法的主体思想是将自然语言转化为格式化的逻辑形式，将问题的语义信息生成对应的逻辑形式。Liang [4]通过解析器将问句转换成 lambada-DCS 表达式，并在知识图谱中检索得出答案。2) 基于信息抽取的方法。该类方法通过提取问题中的关键信息，通过检索知识图谱得到一系列候选答案，然后分别对问句和候选答案进行特征抽取得到特征向量，建立分类器对候选答案进行筛选。Yao 等人[5]使用依存分析技术找到问句中的主要实体，利用实体信息从知识图谱中找到实体子图，抽取问句和实体子图的多种特征进行筛选，从而确定答案。3) 基于深度学习的方法。该类方法通过分布式表示把问题和候选答案都映射到同一向量空间，通过训练模型提高问题向量和正确答案向量的相似度得分。模型训练完成后则可根据候选答案的向量表示和问题表示的得分进行筛选。Dong 等人[6]基于卷积神经网络提出了 MCCNNs，模型使用答案路径、答案类型、答案上下文的向量表示之和作为问题的向量表示。

相较于英文问答系统的研究成果，中文问答系统的研究在规模和研究水平上存在很多差距。其中主要原因有两个。一方面，中文不像英文那样，词与词之间均使用空格分开，这就造成数据预处理中有额外的分词步骤。同时，这也使英文问答系统中现有的部分技术和已经产生的研究成果不可以直接投入使用。另外一方面，中文知识图谱的资源缺乏也是其一，而大规模的英文知识图谱却有很多，例如 Freebase，YAGO，WordNet 等。

NLPCC-ICCPOL 发布的中文知识图谱以及相应的问答数据集，解决了数据缺乏这一问题。因此，本文拟采用深度学习、信息抽取相结合的方法，在开放领域的中文知识图谱问答做进一步探索。本文提出的知识图谱问答系统分为命名实体识别、属性抽取和答案检索三个部分。在命名实体识别阶段，通过改进 BERT 提出的 BERDAT (Bidirectional Encoder Representation from Dynamic Attention Transformers)模型得到问句的 Embedding 向量，将此向量输入到 BLSTM-CRF 模型中，得到最佳的标记序列，从而识别出正确实体。属性抽取阶段通过 BERDAT 模型得到“问句－候选关系”对的 Embedding，通过分类器判断出正确的属性值。最后在知识图谱中进行答案检索得到最终答案。通过对比实验，本文提出的知识图谱问答模型在 NLPCC-KBQA 数据集上取得了 97.54%的 F1 分数，超越了此前公开的最佳模型。

2. 相关理论

2.1. 命名实体识别

命名实体识别(Named Entity Recognition, NER)是 NLP 中一项非常基础的任务，通过对输入文字的每个位置标注出相应的实体信息，实现实体识别的功能。常用的实体标注方法有 BIO、BIOES、BMES，本文实验采用 BIO 标注方法。BIO 将实体 X 标注为 B-X、I-X、O 的格式，其中 B-表示实体的起始位置，I——表示实体的中间或结尾，O——表示不属于实体。

早期主要采用基于规则的方法，这些方法主要基于人工定义的语义和句法规则来识别实体，从而造成人工成本高、可移植性差等问题。因此，目前的研究主要集中在基于概率统计和基于深度学习的方法上[7]。根据特征提取的方式不同，可以分为基于字符和基于词的嵌入方式。Lample 等人[8]利用 BiLSTM 提取字符级特征，与词典中的词向量融合形成最终输入向量，并将 BiLSTM 和 CRF 模型相结合进行命名实体识别，在英语、德语、西班牙语等测试语料库中取得了良好的效果。Strubel 等人[9]提出了一种基于扩张卷积神经网络(ID-CNNs)的模型，其通过 skip-n-gram 方法在 SENNA 语料上训练词嵌入向量。

这两种方法的各自优缺点也很明显，基于字符的嵌入方式切断了词的边界信息，丢失了隐藏在词中的语义特征。基于词的嵌入方法虽然保留了词的边界信息和语义特征，但是在数据训练过程中要先进行分词，如果分词错误就会影响整个模型的训练效果。为了综合各自的优点，Ma 等人[10]利用 CNN 提取单词的字符级 Embedding 向量，然后将字符和词的 Embedding 向量相连接后输入到 RNN 上下文编码器。Peters 等人[11]提出了基于字符卷积的 ELMO 词表示方法，可以根据上下文语境来生成相应词的向量表示。

2.2. BERT

BERT [12]被 Google 提出后，在多项任务中都取得了突破性的进展。其网络架构使用的是多层 Transformer 结构，Transformer 最大的特点是引入 Attention 机制，从而增加词向量模型泛化能力，充分描述字符级、词级以及句子级的关系特征。具体的模型结构如图 1 所示，其中输入的 Embedding 向量是由 PositionEmbedding，SegmentEmbedding，TokenEmbedding 相加得到。PositionEmbedding 是编码单词出现的位置，SegmentEmbedding 用于区分每一个单词属于句子 A 还是句子 B，TokenEmbedding 是每个单词的 Embedding，三个形式的 Embedding 都是通过训练学习得到。

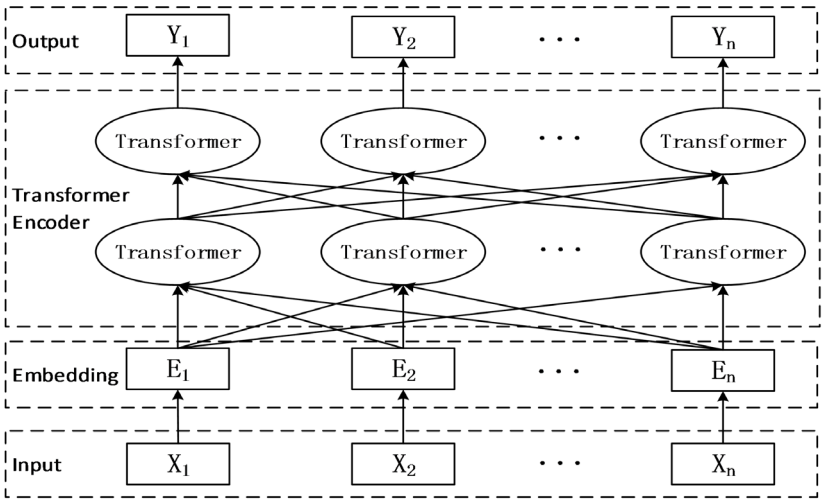

Figure 1. Network architecture of BERT
图 1. BERT 的网络结构

Transformer 结构[13]是 Google 提出的注意力机制模型，该模型分为编码器和解码器两部分。在 BERT 中只用到了 Transformer 编码器，该编码器主要分为 4 个部分：多头注意力模型、归一化层、前馈网络层和归一化层。

3. 知识图谱问答模型

3.1. 模型流程

本文提出的知识图谱问答模型分为三大模块：命名实体识别、属性抽取和答案检索。首先通过命名实体识别步骤提取问句中的实体，然后检索知识图谱返回候选答案的三元组集合，最后通过属性抽取步骤对候选三元组中的属性进行筛选排序，最终输出得分最高的三元组中的答案。具体流程如图 2 所示。其中，命名实体识别和答属性抽取模型都使用 BERDAT 模型做特征提取，然后将特征作为输入，利用不同的网络结构完成相应的功能。

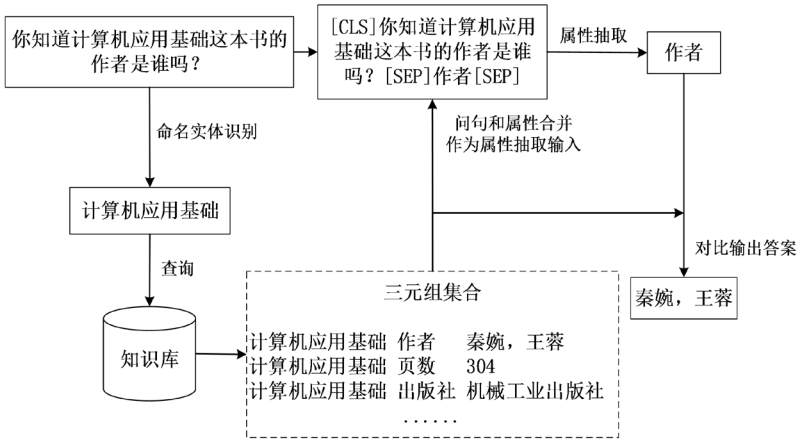

Figure 2. Flow chart of knowledge-based Question Answering
图 2. 知识图谱问答流程图

3.2. BERDAT

Jawahar 等人[14]等人通过实验验证了 BERT 每一层 Transformer 对文本的理解都有所不同，从浅层到高层可以分别学习到短语级别的，句法级别的，和语义级别的信息。因此，本文为了是输出向量更好的涵盖更多的信息，对 BERT 进行了改进，将十二层 Transformer 生成的向量进行动态融合，提出了 BERDAT 模型。

模型的设计如图 3 所示，其中[*CLS*]、Tok_i是模型对输入格式，E_i是对输入进行 Embedding 后得到的结果。$Embedding_i$是模型的第i层 Transformer 输出的结果向量。拼接层首先将每一层 Transformer 输出的向量先进行扩维，然后再进行合并，例如 $Embedding_i$的维度是(64, 64, 768)，那么 12 层 Transformer 进行拼接后的维度是(64, 64, 768, 12)。1 × 1 卷积层的作用是把拼接后的向量降维，得到维度是(64, 64, 768, 1)的向量，最终对其进行调整，得到维度是(64, 64, 768)的输出向量。这和 BERT 的输出向量是相同的维度，达到了对多层 Transformer 融合的效果，同时也便于后文在 BERDAT 的基础上构建命名实体识别模型和属性抽取模型。

3.3. 命名实体识别模型

命名实体识别模型 BERDAT-BiLSTM-CRF 可分为特征提取和实体标注两部分，具体结构如图 4 所示。在特征提取部分中，长度为N的输入问句被处理成序列$\{[CLS], Tok_1, \cdots, Tok_N\}$，经过词嵌入后得到 $N+1$

个词向量，将其输入到 BERDAT 模型中后，得到的输出向量即为提取的特征向量。实体标注部分使用常见 BiLSTM-CRF 网络，首先将特征向量输入到双向 LSTM 层，将每个时间序列的正向反向输出拼接，经过全连接层映射为一个维度为标注类型数的向量。由于本文仅定义了一种实体类型，因此标注类型有 3 种，分别是“B-ENT”、“I-ENT”、“O”。最后，CRF 层对上一层的输入向量求出条件概率最大的标注序列，将问句的每个位置打上了标注信息。

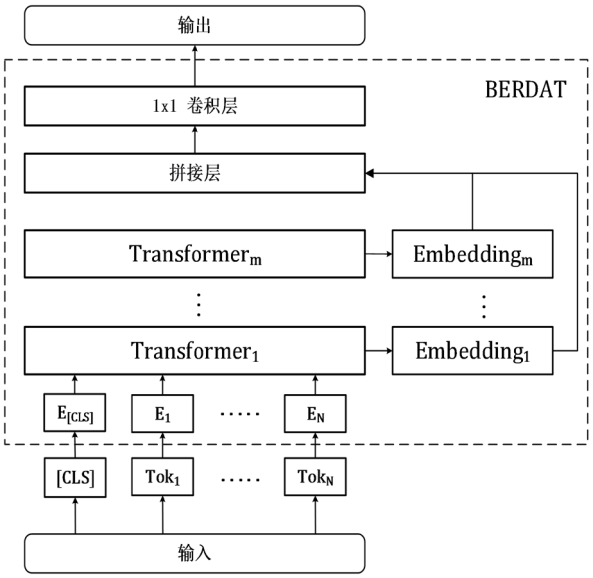

Figure 3. BERDAT model
图 3. BERDAT 模型

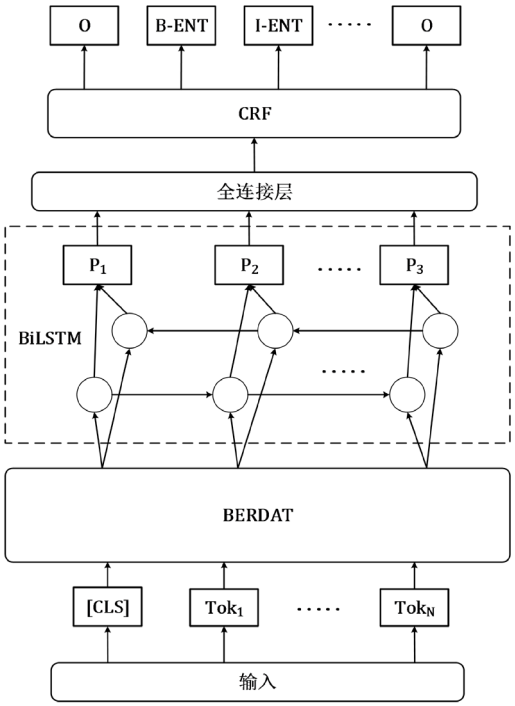

Figure 4. Named entity recognition model
图 4. 命名实体识别模型

3.4. 属性抽取模型

完成命名实体识别后，在知识图谱中检索预测实体，将包含该实体的三元组作为候选三元组集合。然后将问句和三元组中的属性值做预处理，形成一个新的输入序列：“$[CLS]$问句$[SEP]$属性值$[SEP]$”。最后通过 Softmax 进行二分类，从而判断该序列中的属性是否正确。在训练过程中，标签为 0 的 Softmax 概率小则为错误属性，标签为 1 的概率大则为正确属性。

属性抽取模型如图 5 所示。特征提取的过程与命名实体识别模型一致，经过 BERDAT 网络后，取输出向量中$[CLS]$位置的向量$E^{o}_{[CLS]}$作为输入序列的特征向量，得到一个维度是(1，隐藏层个数)的特征矩阵。然后依次经过全连接层和 Softmax 层，得到两种分类各自的概率值，通过比较两个概率值确定该属性的分类。分类为 0 代表属性和问句不符合，反之则是正确属性。

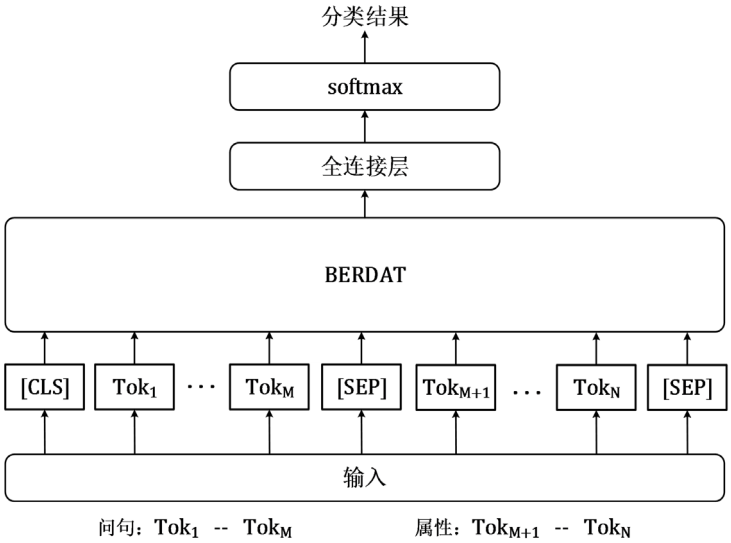

Figure 5. Attribute extraction model
图 5. 属性抽取模型

4. 实验与结果分析

4.1. 数据集及实验环境

本文使用 NLPCC-ICCPOL-2016 发布的 KBQA 数据。该数据集提供了包含 14,609 个问答对的训练集和包含 9870 个问答对的测试集。本文从总的训练集划分出 2609 个问答对作为验证集，剩余 12,000 个问答对作为训练集。但是知识图谱中的数据是从百度百科中自动抽取的，导致数据集中存在不少的噪声，因此在实验之前首先对数据进行数据预处理。比如去除属性中出现空格和部分非中文字符，将所有大写字母转化成小写，去除信息缺失和信息不匹配的三元组数据。最终，NLPCC-KBQA 数据集的划分情况如表 1 所示。

Table 1. Data set situation partition
表 1. 数据集划分情况

	训练集	验证集	测试集
预处理前	12000	2609	9870
预处理后	11869	2587	9581

本文实验运行在 CPU 为 InterE5-2650、内存为 62GB 的计算机上，模型训练所用显卡为 NvidiaGTX1080Ti，显存 22GB，所用深度学习框架为 Tensorflow1.12，操作系统为 ubuntu18.04，知识图谱数据存储和检索使用 MySQL8.0。

4.2. 命名实体识别

在命名实体识别的步骤中，使用 NLPCC-NER 数据集进行实验。该数据集是由预处理后的 NLPCC-KBQA 数据集通过实体标注生成的。具体的方法是利用问答对中的标准答案，反向查找问句对应的实体位置进行 BIO 标注。

实验中将学习率的区间设置为[3e−5, 5e−5]，优化损失函数采用带权值衰减的 Adam 优化器[15]。模型训练的超参数设置是：最大序列长度为 128，Transformer 多头数为 12，Transformer 层数为 12，Transformer 隐藏层的维度为 768，LSTM 神经元个数为 128，dropout 比率为 0.9。最终 BERDAT 模型在 NLPCC-NER 的实验中，当学习率为 4e−5 时效果最好。相应的，验证集和测试集的评价结果如表 2 所示，精确率、召回率和 F1 分数均在 97%以上，达到了很高的水平。

Table 2. Named entity recognition result (unit: %)
表 2. 命名实体识别结果(单位：%)

	评价指标	验证集	测试集
BERDAT-BiLSTM-CRF	Precision	97.30	97.50
	Recall	97.45	97.67
	F1	97.37	97.59

4.3. 属性抽取

为了训练属性抽取模型，需要使用预处理后的 NLPCC-KBQA 数据集生成属性抽取数据集。在这个过程中，不仅要有正确的“问句－属性”对，还需要人为生成错误的“问句－属性”，从而提高模型的区分能力，避免出现过拟合的现象。制作正样本是在“问句–属性”对后面加上数字“1”，表示这是正确的样本；制作负样本则是在“问句－错误属性”对后面加上数字“0”，其中错误属性是在知识图谱的所有属性集合中随机选取 5 个，从而使得一个正样本能够生成 5 个负样本。得到的属性抽取数据集规模如表 3 所示。

Table 3. The size of attribute extraction dataset
表 3. 属性抽取数据集规模

	训练集	验证集	测试集
正样本	11,869	2587	9581
负样本	59,345	12,935	47,905
总数	71,214	15,522	57,486

将属性抽取训练集数据送入 BERDAT 模型进行训练。超参数选取除了没有 LSTM 以外，与命名实体识别一致。属性抽取模型在验证集和测试集的测试结果如表 4 所示。属性抽取模型在验证集和测试集的准确率相差不大且均达到了 98%以上。同时结合 AUC 指标，可以看出模型的属性区分能力很好，为知识图谱问答模型提供了有力的保障。

Table 4. Results of attribute extraction
表 4. 属性抽取的结果(单位：%)

	验证集	测试集
Accuracy	99.12	98.61
AUC	97.94	96.55

4.4. 知识图谱问答结果

通过命名实体识别和属性抽取步骤后，选择各自表现最好的超参数，应用在知识图谱问答系统中，最终的测试结果如表 5 所示。其中验证集和测试集都来源于 NLPCC-KBQA 预处理后的原始问答对。在测试集取得了很好的实验结果，三个评价指标均超过了 96%。

Table 5. Result of knowledge-based Question Answering
表 5. 知识图谱问答的结果(单位：%)

评价指标	验证集	测试集
Precision	99.19	98.64
Recall	95.25	96.45
F1	97.18	97.54

在公开的评价指标中以测试集的F1分数为准。本文对比了与其他公开模型的测试结果，如表6所示。其中，DPQA [16]是王玥等人基于无监督方法提出的动态规划知识图谱问答，虽然无监督的方式可以减少人工标记的工作量，但其问答效果不算太好。InsunKBQA [17]是周博通等人基于知识图谱三元组中谓词的属性映射构建的问答系统。PKU [18]是 NLPCC-ICCPOL-2016KBQA 任务评测成绩的第一名，主要依靠人工规则来提高问答系统的性能，例如构造正则表达式去掉问句中的冗余信息。WHUT [19]是张芳容等人通过句法分析等方式实现的问答系统。SCU [20]使用 BERT 进行特征提取，将答案选择分解为答案匹配和阈值选择两个步骤。

Table 6. Comparison of results of different question answering models in test set
表 6. 不同问答模型在测试集的结果对比(单位：%)

问答模型	F1
DPQA	71.00
InsunKBQA	81.35
PKU	82.47
WHUT	82.94
SCU	87.05
BERDAT(本文方法)	97.54

与上述问答模型相比，本文问答模型的 F1 分数分别提升了 26.54%、16.19%、15.07%、14.6%、10.49%，取得了目前公开模型中最高的分数。主要原因是本文在 BERT 的基础上改进，提出了 BERDAT 模型，提高了对语言的理解能力，从而更好的提取问句中的特征，让命名实体识别模型和属性抽取模型的性能均得到提升。

5. 结语

本文根据 NLPCC-ICCPOL-2016 提供的开放领域的问答数据集，提出了基于 BERDAT 的知识图谱问答模型，分为命名实体识别、属性抽取和答案检索三个部分。首先通过命名实体识别模型提取问句中的实体，然后以该实体为关键词检索知识图谱，得到候选三元组集合，接着通过属性抽取模型对候选三元组集合中的每一个属性值进行判断。最后，通过筛选的三元组视为预测三元组，并将其中的第三个元素作为答案输出。

命名实体识别和属性抽取是大规模知识图谱问答的两个难点，本文提出了相应的模型解决这两个问题。在进行命名实体识别时，先使用 BERDAT 进行特征提取，然后使用 BiLSTM-CRF 模型进行实体预测。在进行属性抽取时，将其看作二分类任务，结合 BERDAT 和 Softmax 来判断属性是否正确。最后，本文方法在 NLPCC-KBQA 数据集上的 F1 分数为 97.54%，与以往已公开的方法相比取得了更好的结果。

参考文献

[1] Park, S., Kwon, S., Kim, B., *et al.* (2015) Question Answering System using Multiple Information Source and Open Type Answer Merge. *Proceedings of the* 2015 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Demonstrations*, Denver, June 2015, 111-115. https://doi.org/10.3115/v1/N15-3023

[2] Hristovski, D., Dinevski, D., Kastrin, A. and Rindflesch, T.C. (2015) Biomedical Question Answering Using Semantic Relations. *BMC Bioinformatics*, **16**, Article No. 6. https://doi.org/10.1186/s12859-014-0365-3

[3] Zhao, S., Zheng, Y., Zhu, C., *et al.* (2016) Semantic Computation in Geography Question Answering. 2016 12*th International Conference on Natural Computation*, *Fuzzy Systems and Knowledge Discovery* (*ICNC-FSKD*), Changsha, 13-15 August 2016, 1572-1576. https://doi.org/10.1109/FSKD.2016.7603410

[4] Liang, P., Jordan, M.I. and Klein, D. (2013) Lambda Dependency-Based Compositional Semantics. *Computational Linguistics*, **39**, 389-446. https://doi.org/10.1162/COLI_a_00127

[5] Yao, X. and Van Durme, B. (2014) Information Extraction over Structured Data: Question Answering with Freebase. *Proceedings of the* 52*nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, June 2014, 956-966. https://doi.org/10.3115/v1/P14-1090

[6] Dong, L., Wei, F., Zhou, M. and Xu, K. (2015) Question Answering over Freebase with Multi-Column Convolutional Neural Networks. *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics & the* 7*th International Joint Conference on Natural Language Processing*, Beijing, July 2015, 260-269. https://doi.org/10.3115/v1/P15-1026

[7] Liu, L. and Wang, D.B. (2018) A Review on Named Entity Recognition. *Journal of the China Society for Scientific and Technical Information*, **37**, 329-340.

[8] Lample, G., Ballesteros, M., Subramanian, S., *et al.* (2016) Neural Architectures for Named Entity Recognition. *Proceedings of the* 2016 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, San Diego, June 2016, 260-270. https://doi.org/10.18653/v1/N16-1030

[9] Strubell, E., Verga, P., Belanger, D., *et al.* (2017) Fast and Accurate Entity Recognition with Iterated Dilated Convolutions. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017, 2670-2680. https://doi.org/10.18653/v1/D17-1283

[10] Ma, X. and Hovy, E. (2016) End-to-End Sequence Labeling via Bi-directional LSTM-CNNs-CRF. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*), Berlin, August 2016, 1064-1074. https://doi.org/10.18653/v1/P16-1101

[11] Peters, M.E., Neumann, M., Iyyer, M., *et al.* (2018) Deep Contextualized Word Representations. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, *Volume* 1 (*Long Papers*), New Orleans, June 2018, 2227-2237. https://doi.org/10.18653/v1/N18-1202

[12] Devlin, J., Chang, M.W., Lee, K., *et al.* (2018) BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding.

[13] Vaswani, A., Shazeer, N., Parmar, N., *et al.* (2017) Attention Is All You Need. 31*st Conference on Neural Information Processing Systems* (*NIPS* 2017), Long Beach, 5998-6008.

[14] Jawahar, G., Sagot, B. and Seddah, D. (2019) What Does BERT Learn about the Structure of Language? *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 3651-3657. https://doi.org/10.18653/v1/P19-1356

[15] Loshchilov, I. and Hutter, F. (2019) Decoupled Weight Decay Regularization. *International Conference on Learning Representations*, 1-8.

[16] 王玥, 张日崇. 基于动态规划的知识库问答方法[J]. 郑州大学学报(理学版), 2019, 51(4): 37-42.

[17] 周博通, 孙承杰, 林磊, 刘秉权. 基于 LSTM 的大规模知识库自动问答[J]. 北京大学学报(自然科学版), 2018, 54(2): 286-292.

[18] Lai, Y., Jia, Y., Lin, Y., Feng, Y. and Zhao, D. (2018) A Chinese Question Answering System for Single-Relation Factoid Questions. In: Huang, X., Jiang, J., Zhao, D., Feng, Y. and Hong, Y., Eds., *Natural Language Processing and Chinese Computing*. *Lecture Notes in Computer Science*, Vol. 10619, Springer, Cham, 124-135. https://doi.org/10.1007/978-3-319-73618-1_11

[19] 张芳容, 杨青. 知识库问答系统中实体关系抽取方法研究[J]. 计算机工程与应用, 2020, 56(11): 219-224.

[20] 吴天波, 刘露平, 罗晓东, 卿粼波, 何小海. 基于弱依赖信息的知识库问答[J]. 计算机工程, 2020, 1-8. https://doi.org/10.19678/j.issn.1000-3428.0058312