浙江科技学院学报，第 34 卷第 5 期，2022 年 10 月
Journal of Zhejiang University of Science and Technology
Vol. 34 No. 5, Oct. 2022

doi：10.3969/j.issn.1671-8798.2022.05.005

针对机器问答中多跳问题的深度学习网络模型

邵　霭[a]，许彩娥[a]，万　健[a]，张　蕾[a]，郑慧琳[b]
（浙江科技学院 a.信息与电子工程学院；b.生物与化学工程学院，杭州 310023）

摘　要：多跳问答（multi-hop question answering，multi-hop QA）是文本问答的一项重要且具有挑战性的任务。针对现有方法在解决多跳问题时答案推理能力弱、答案寻找的准确率低等问题提出一种多跳问题的深度学习网络模型 AGTNet（albert graph attention network，轻量双向编码图注意力网络）。首先在神经网络隐藏层使用参数共享和矩阵分解技术，然后使用点积计算方式进行答案预测，最后使用已标注的数据集对 AGTNet 模型进行训练验证。试验结果表明，本模型经过训练后在测试集上的 F_1 值达到 70.4；与现有的多跳问答推理模型相比，本模型拥有较优的实体级推理能力，能够有效提高多跳问答推理能力，从而提升了问答系统的响应速度和准确率。本研究结果为问答系统和多轮对话机器人的研发提供了理论依据。

关键词：多跳问答；深度学习；表征提取；问答推理
中图分类号：TP183　　**文献标志码**：A　　**文章编号**：1671-8798(2022)05-0419-07

Deep learning network model for multi-hop problems in machine question answering

SHAO Ai[a], XU Caie[a], WAN Jian[a], ZHANG Lei[a], ZHENG Huilin[b]
(a. School of Information and Electronic Engineering; b. School of Biological and Chemical Engineering, Zhejiang University of Science and Technology, Hangzhou 310023, Zhejiang, China)

Abstract: Multi-hop question answering (multi-hop QA) is an important and challenging task in text question answering. Aiming at the problems that the existing methods are afflicted by weak reasoning ability and low accuracy in answer finding when solving multi-hop problems, a deep learning network model was proposed in the name of AGTNet (albert graph attention network) for multi-hop problems in the field of question answering. Firstly, parameter sharing and matrix factorization techniques were employed in the neural network hidden layer, and then

收稿日期：2021-08-14
基金项目：国家自然科学基金项目(61972358)
通信作者：万　健（1969—　），男，福建省泉州人，教授，博士，主要从事云计算及大数据研究。E-mail：wanjian@zust.edu.cn。

the dot product calculation method was used to predict the answer. Finally, the labeled data sets were applied to train and verify the AGTNet model. The experimental results show that the F_1 value of the model on the test set reaches 70.4 after training. Compared with the existing multi-hop question answering reasoning model, the proposed model boasts better entity-level reasoning ability, which can effectively improve the multi-hop question answering reasoning ability, so as to improve the response speed and accuracy of the question answering system. The results of this study provide a theoretical basis for research and development of question answering system and multi-round dialogue robot.

Keywords: multi-hop QA; deep learning; representation extraction; question answering reasoning

自然语言处理是人工智能的重要分支,自动问答系统为评估自然语言处理能力提供了一种可以量化的客观方法,可以用来测试人工智能系统的推理能力[1-2],这正逐渐成为一种人与机器进行自然交互的新趋势。问答系统能够更准确地理解以自然语言描述的用户问题,并依据用户的真实意图返回给用户更精准的答案,它将成为下一代搜索引擎的新形态。近年来,深度学习模型的发展使得机器问答取得了长足的进步,针对机器问答的深度学习模型层出不穷,机器问答的研究进入了一个全新的阶段。然而当前大多数问答工作都集中在从单一段落中寻找问题和答案[3-4],很少测试底层模型的深层推理能力,且单段问答模型在与问题匹配的句子中寻找答案,不涉及复杂的推理,因此多跳问答模型成为下一个需要攻克的前沿课题。近年来研究者们提出了一些专门用于评估问答模型多跳推理能力的数据集,如 WikiHop[5]、ComplexWebQuestions[6]和 HotpotQA[7]等。在此基础之上,越来越多的研究者根据多个句子或段落中的实体之间的共现关系来构建图网络结构。Song 等[8]设计了一个 DAG(database availability group,数据实体组)样式的递归层来对实体之间的关系进行建模,有效提升了实体级信息传递能力。Dhingra 等[9]使用 GCN(graph convolutional network,图神经网络)处理实体图,将神经网络[10]引入图中实体后极大提升了节点间的信息互联。Xiao 等[11]提出了一种基于 GFN(dynamic fusion entity graph,动态实体图)的多跳问答模型,通过实体信息动态传递的方式进行问答的推理和预测,提升了问答的准确率。Tu 等[12]通过引入文档节点和查询节点将实体图扩展为异构图,提升了图中节点信息查询的速度,为图神经网络提供了新模式。但是这些基于深度学习神经网络模型的问答系统在进行文本的特征提取时无法保证其质量,同时在推理计算相似度层面的运算能力不足,所以在面对复杂问句及长难问句时仍然存在问句解析难度大,实体级别模型推理能力弱,问答匹配准确率较低等问题。

基于上述研究,本研究提出了 AGTNet(albert graph attention network,轻量双向编码图注意力网络)模型,本模型包含了表征提取、推理计算、结果预测 3 个模块。模型在表征抽取层的神经网络隐藏部分使用参数共享和矩阵分解技术来降低模型的空间复杂度,同时使用点积计算方式的图注意力机制进行答案预测,从而提升了字词级别的表征提取质量,提高了模型实体级推理能力和问答匹配的准确率。

1 神经网络注意力机制算法

1.1 预训练模型

BERT(bidirectional encoder representation from transformers,双向表示编码器)是谷歌团队于 2018 年发布的预训练模型,在实际应用中往往受到硬件内存的限制,而增加模型的隐藏层大小也会导致性能下降。2019 年 ALBERT(a lite bidirectional encoder representation from transformers,轻量双向表示编码器)预训练模型的发布改善了 BERT 参数大、资源消耗多的缺点。BERT 的隐藏层单元数为

2 048，相比 ALBERT 翻了一倍；然而，该模型在 RACE（reading comprehension dataset collected from english examinations，英语测验阅读理解数据集）上的准确率却较低，而 ALBERT 不但参数量比 BERT 少得多，而且准确率与 BERT 相比较高，达到 70.2%。

1.2　图神经网络

图神经网络是神经网络的一个分支，最早于 2005 年被提出。其基本思路是，图中节点的性质由其自身的属性和邻居节点的属性共同决定[13]，网络为图中每个节点分配向量 $\boldsymbol{h}$，该向量包含了自身节点和邻节点的信息，可以用于节点预测、图分类等任务。GAT（graph attention networks，图注意力机制网络）[14]在图神经网络中引入注意力机制[15]，赋予邻节点不同的重要性。图注意力机制模型如图 1 所示，点 $\boldsymbol{h}_1$ 为研究对象，把节点自身也视为一阶邻居，根据节点的特征向量 $\boldsymbol{h}_1$ 计算节点词向量权重 α。在更新节点的特征向量时，根据权重对邻居信息进行聚合，得到包含邻居信息的节点向量[16]。

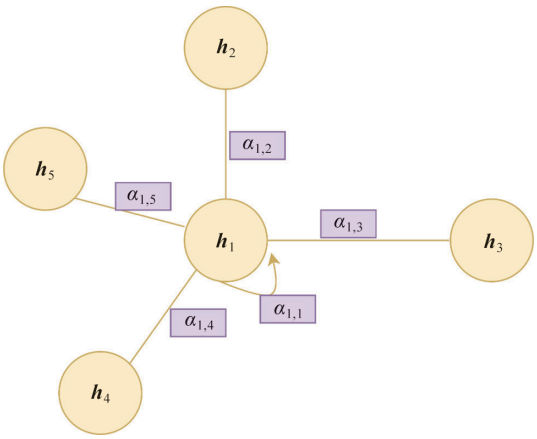

图 1　图注意力机制模型

Fig. 1　Graph attention mechanism model

1.3　点积注意力机制

注意力机制模型为了获得足够的表达能力，需要采用线性转换来提升输入特征的维度[17]，为此，模型使用权重矩阵 $\boldsymbol{W}$ 对每个节点进行线性变换，然后在每个节点上计算注意力机制权重系数。$\boldsymbol{h}_i$ 和 $\boldsymbol{h}_j$ 为点积注意力机制节点的实体向量，采用点积算法计算节点中词向量的权重 α，通过激活函数 σ 计算实体节点间的注意力权重 β。点积注意力机制在点积算法的基础上进一步融合了注意力权重，并且对注意力权重的计算采用了相同数学表达式[18]。

2　模型结构详细设计

本研究提出的 AGTNet 模型结构如图 2 所示，模型包括表征提取和推理计算模块。$[\boldsymbol{P}_1,\boldsymbol{P}_2,\cdots,\boldsymbol{P}_n]$ 与 $[\boldsymbol{Q}_1,\boldsymbol{Q}_2,\cdots,\boldsymbol{Q}_n]$ 为输入的问题和段落向量，$[\boldsymbol{E}_1,\boldsymbol{E}_2,\cdots,\boldsymbol{E}_n]$ 与 $[\boldsymbol{T}_1,\boldsymbol{T}_2,\cdots,\boldsymbol{T}_n]$ 为表征提取后的特征向量，$\boldsymbol{W}$ 为权重矩阵，$[\boldsymbol{R}_1,\boldsymbol{R}_2,\cdots,\boldsymbol{R}_n]$ 为图中节点。本模型在表征抽取模块的神经网络隐藏部分使用参数共享和矩阵分解技术，有效降低了模型的空间复杂度；同时使用点积计算方式的图注意力机制进行答案预测，有效提升了模型问答推理能力和问答准确率，解决了用户问句理解难、模型推理预测计算能力不足等问题。

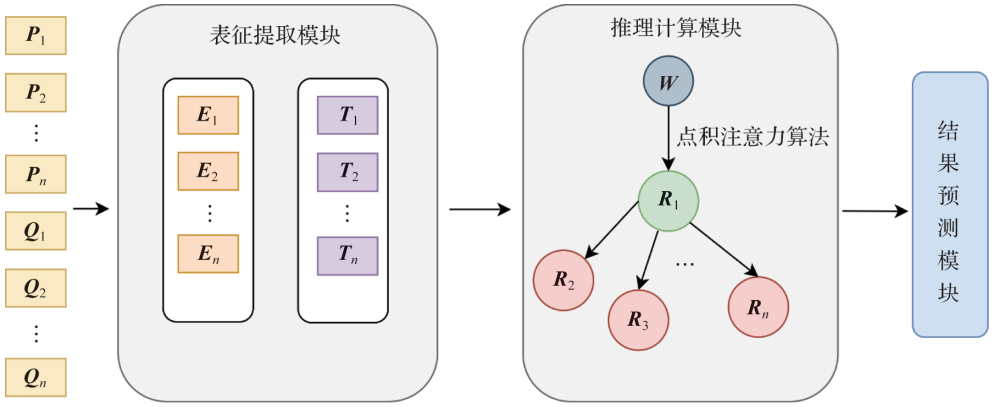

图 2　AGTNet 模型结构

Fig. 2　AGTNet model structure

2.1 表征提取模块

本研究的表征提取模块采用了 ALBERT 模型[19]，将问答语料输入模型，输出对应问题 $\boldsymbol{Q}$ 和段落的词向量 $\boldsymbol{P}$，以及从中提取出的语义向量 $\boldsymbol{s}$。表征提取流程如图 3 所示，首先将问题[$\boldsymbol{Q}_1$，$\boldsymbol{Q}_2$，…，$\boldsymbol{Q}_n$]与段落[$\boldsymbol{P}_1$，$\boldsymbol{P}_2$，…，$\boldsymbol{P}_n$]打包成一个序列共同输入；其次 ALBERT 通过词嵌入方式生成表征向量[$\boldsymbol{E}_1$，$\boldsymbol{E}_2$，…，$\boldsymbol{E}_n$]；最后将表征向量输入编码层，输出编码向量[$\boldsymbol{T}_1$，$\boldsymbol{T}_2$，…，$\boldsymbol{T}_n$]，从编码层随之输出的还有语义向量 $\boldsymbol{s}$。

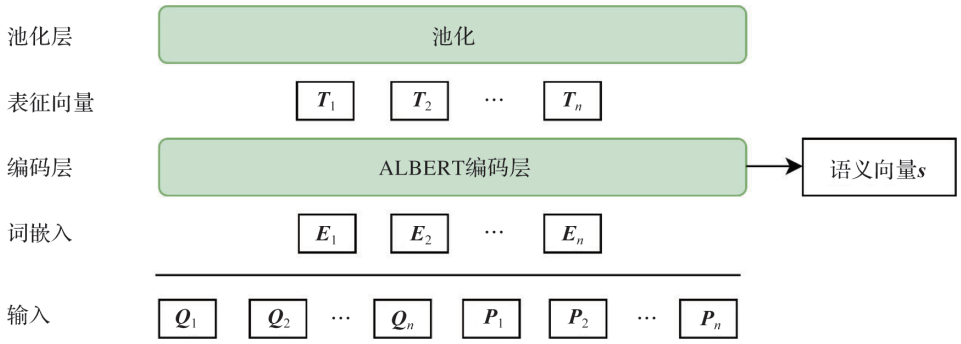

图 3 表征提取流程

Fig. 3 Representation extraction processing

2.2 推理计算模块

2.2.1 实体计算

推理计算模块实体间注意力权重计算过程如图 4 所示，用点积注意力机制来模仿人类逻辑探索和推理过程。$\boldsymbol{R}_1'$ 为融合邻节点信息后的实体。

将图中的节点信息传播到每个邻节点进行融合。在每个推理步骤中，假设每个节点都有信息要传播到它的邻节点，与问题和段落相关性越大的邻节点从附近接收的信息越多。通过在实体上关联问题来查询相关的节点，把问题表示和实体表示相结合，再乘以抽取出的语义向量，计算得出融合邻节点信息后的实体

$$\boldsymbol{E}=\sum_{i=1}^{n}\boldsymbol{Q}\boldsymbol{e}_i\boldsymbol{s}。\tag{1}$$

式(1)中：$\boldsymbol{Q}$ 为问题向量；$\boldsymbol{e}_i$ 为第 i 个词向量；$\boldsymbol{s}$ 为语义向量。

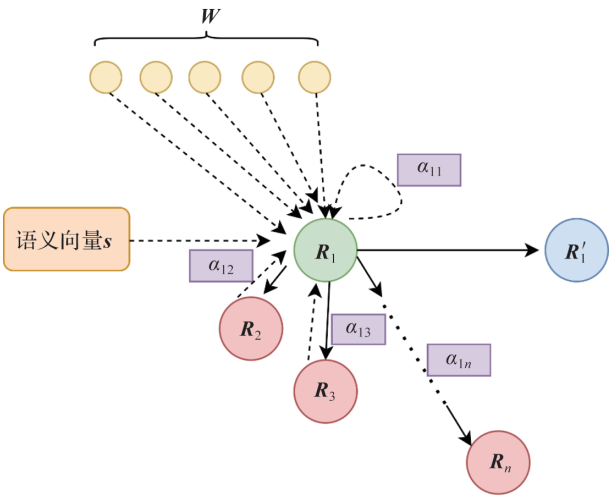

图 4 实体间注意力权重计算过程

Fig. 4 Process of calculating attention weights among entities

2.2.2 推理过程

通过以下方法计算两个实体之间的注意力权重：

$$\beta=\frac{\alpha}{\sum_{i=1}^{k}\alpha_i}。\tag{2}$$

其中

$$\alpha=(\boldsymbol{WE})^{\mathrm{T}}\boldsymbol{WE}。\tag{3}$$

本研究使用的点积图注意力算法与传统 GAT 的不同之处在于每个节点对其邻节点加权求和，从而形成一个新的实体节点状态

$$\boldsymbol{E}'=\mathrm{ReLU}(\beta\boldsymbol{E})。\tag{4}$$

式(4)中：ReLU 为激活函数，用来计算新的实体节点 $\boldsymbol{E}'$ 的状态信息。

2.3　结果预测模块

2.3.1　图转表征

本研究设计了图转表征功能来将实体转化为向量。模块将文本向量 $\boldsymbol{C}$ 与关联实体向量串联，二进制矩阵 $\boldsymbol{M}$ 中的每行对应一个向量，从 $\boldsymbol{E}'$ 中选择一个实体嵌入，使用 LSTM(long-short-term memory network，长短期记忆网络)[20] 进一步处理该信息，以产生下一级上下文向量 $\boldsymbol{C}'$，并且将 $\boldsymbol{C}'$ 作为下一个网络的输入。

$$\boldsymbol{C}'=L(\boldsymbol{C},\boldsymbol{M}\boldsymbol{E}')。\tag{5}$$

式(5)中：$\boldsymbol{C}'$ 为下一级的上下文向量；L 为长短期记忆网络函数；$\boldsymbol{C}$ 为文本向量；$\boldsymbol{M}$ 为二进制矩阵。

2.3.2　预测器

本研究使用的预测器采取级联结构来解决输出依赖，其中 3 个同构的 LSTM 逐层堆叠，预测器有 3 个输出维度，包括答案的开始位置、答案的结束位置及答案的类型。将下一级上下文文本向量 $\boldsymbol{C}'$ 输入预测器，每个预测器输出一个对数，并在这些对数上计算出交叉熵损失，这些交叉熵损失在预测器中都进行了优化，每个损失项都有对应的系数加权。预测器输出的损失函数如下：

$$O_{\text{start}}=F_0(\boldsymbol{C}')；\tag{6}$$

$$O_{\text{end}}=F_1(\boldsymbol{C}',O_{\text{start}})；\tag{7}$$

$$O_{\text{type}}=F_2(\boldsymbol{C}',O_{\text{start}},O_{\text{end}})。\tag{8}$$

式(6)～(8)中：F_0、F_1、F_2 分别为预测器函数；O_{start}、O_{end}、O_{type} 分别为答案开始位置、结束位置、类型的对数。

预测器输出损失函数后将进行损失值的计算，计算过程如下：

$$\boldsymbol{L}=\boldsymbol{L}_{\text{start}}+\boldsymbol{\lambda}\boldsymbol{L}_{\text{end}}+\boldsymbol{\lambda}\boldsymbol{L}_{\text{type}}。\tag{9}$$

式(9)中：$\boldsymbol{\lambda}$ 为相应损失的系数矩阵；$\boldsymbol{L}_{\text{start}}$、$\boldsymbol{L}_{\text{end}}$、$\boldsymbol{L}_{\text{type}}$ 分别为答案开始位置、结束位置、类型的损失值。

3　试验设计与结果分析

3.1　数据集

本研究使用 HotpotQA 数据集[7] 进行试验，数据集由 11.3 万个人工设计的问题组成，分为干扰项和完整项。在干扰项中 84% 的问题需要多跳推理，这些数据被分成训练集、验证集和测试集。本研究分别在 HotpotQA 数据集的干扰项和完整项设置下进行了试验。

3.2　试验细节

表征提取模块使用了 ALBERT 预训练模型。在图的构建阶段，使用斯坦福大学预训练的 NER(named entity recognition，命名实体识别)模型[21] 来提取命名实体，图中实体的最大数量被设定为 40，实体图中每个实体节点的平均度为 3.52；在推理阶段，图注意力模型中每个节点都融合了其余节点的信息。

3.3　试验结果

不同模型在 HotpotQA 测试集干扰项设置下的试验结果见表 1，试验评估指标为 EM(expectation maximum，最大期望)值及综合指标 F_1 值。本研究提出的 AGTNet 模型的 EM 值达到 45.53，F_1 值达到 65.25，与其他对比模型相比，AGTNet 在 HotpotQA 测试集干扰项设置下取得了较优的结果。

不同模型在 HotpotQA 测试集完整项设置下的试验结果见表 2。AGTNet 模型在每个指标上的表现均优于其他模型，EM 值达到 69.66，F_1 值达到 81.15。QFE(query-focused extractor，聚焦查询器)[8]、DFGN(dynamically fused graph network，动态融合图网络)[11] 等基于深度学习的模型在进行文本特征提取时无法保证其质量，同时在相似度计算层面的运算能力不足，所以在面对复杂问句时仍然存

在问句解析难度大、模型推理能力弱等问题，AGTNet 针对这些问题，使用参数共享、矩阵分解技术及点积注意力算法，与其他对比模型相比，AGTNet 在 HotpotQA 测试集完整项设置下取得了较优的结果。

表 1 不同模型在 HotpotQA 测试集干扰项设置下的试验结果

Table 1 Experimental results of different models in distractor items of the HotpotQA test set

模型	EM	F_1
Baseline[5]	23.90	32.90
QFE[6]	34.63	59.61
DFGN[11]	33.62	59.82
TAP2[12]	39.77	69.12
ALBERT+GAT	42.13	63.08
AGTNet	45.53	65.25

表 2 不同模型在 HotpotQA 测试集完整项设置下的试验结果

Table 2 Experimental results of different models in complete items of the HotpotQA test set

模型	EM	F_1
Baseline[5]	45.61	59.05
QFE[8]	63.88	68.26
DFGN[11]	56.35	69.77
LQR-Net	60.22	73.85
C2F Reader	67.82	81.25
ALBERT+GAT	65.73	79.09
AGTNet	69.66	81.15

3.3.1 评估依据

为了评估问答系统的逻辑严谨性，本研究使用联合答案/正确答案，即正确答案中联合答案的比例来评估 AGTNet 的有效性。联合答案指那些从支持性事实中推导出来的答案，因此，这个比例代表了推理的逻辑严谨性。本研究的联合答案/正确答案比例达到 59.4%，超过文献[5]的 10.9%和 QFE 的 34.6%。

3.3.2 结果分析

本研究提出的 AGTNet 模型在 EM 值及 F_1 值上都超过比较的其他模型。由于多跳问答的难点在于问题与段落在字面上只有很少的共同词汇，甚至与问题没有语义关系，因此，传统的检索提取模型很难找到问题与段落之间的关联。然而，本研究提出的模型 AGTNet 会根据线索逐渐发现相关实体。

4 结 语

本研究提出了一个新模型来解决大规模的多跳问答问题。模型的表征抽取模块使用 ALBERT 做实体抽取，推理计算模块使用了结合语义向量的点积图注意力机制算法，从而达到较优的实体级推理水平。本研究提出的 AGTNet 模型在 HotpotQA 数据集上进行训练，取得了良好的训练结果，表明了 AGTNet 的有效性。但是在应对复杂问句和深层推理时，AGTNet 的计算逻辑性还有待提高，根据图注意力机制的特性，后续的研究中可以通过加入逻辑性变量来提高可靠性。此外，通过优化系统之间的交互，结合微调和基于特征的表征抽取将提高 ALBERT 的容量。

参考文献：

[1] 王瑛，何启涛. 智能问答系统研究[J]. 电子技术与软件工程，2019(5)：174.

[2] RAIPURKAR P, ZHANG J, LOPYREY K, et al. SQuAD: 100,000+ questions for machine comprehension of text [J]. Advances in Neural Information Processing Systems, 2017, 1(8): 5999.

[3] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension[J]. High Technology Letters, 2017, 23(2): 179.

[4] 姚智. 基于深度学习的医疗问答系统的开发[J]. 中国医疗设备，2019，34(12)：88.

[5] WELBL J, STENETORP P, RIEDEL S. Constructing datasets for multi-hop reading comprehension across documents[J]. Transactions of the Association for Computational Linguistics, 2018, 6(8): 287.

[6] TALMOR A, BERANT J. The web as a knowledge-base for answering complex questions[C]//Proceedings of the conference of 2018 the North American Chapter of the Association for Computational Linguistics. San Francisco: Margan Kaufmann,2018:269.

[7] YANG Z, QI P, ZHANG S, et al. Hotpotqa: a dataset for diverse, explainable multi-hop question answering[J]. Semantic Scholar,2016,2(4):4.

[8] SONG L, WANG Z, YU M, et al. Exploringgraph-structured passage representation for multi-hop reading comprehension with graph neural networks[J]. Communications of the ACM,2020,63(11):139.

[9] DHINGRA B, JIN Q, YANG Z, et al. Neural models for reasoning over multiple mentions using coreference[J]. International Journal of Computational Intelligence Systems,2020,28(12):2704.

[10] KIP F, WELLING M. Semi-supervised classification with graph convolutional networks[J]. Computer Science, 2015,9(13):4.

[11] QIU L, XIAO Y, QU Y, et al. Dynamically fused graph network for multi-hop reasoning[C]//2019 Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Boston: Kluwer,2019:58.

[12] TU M, WANG G, HUANG J, et al. Multi-hop reading comprehension across multiple documents by reasoning over heterogeneous graphs[C]//NAACL HLT 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Washington: IEEE Computer Society,2019:13.

[13] SCARSELLI F, GORI M, TSOI A C, et al. The graph neural network model[J]. IEEE Transactions on Neural Networks,2009,20(1):61.

[14] VELIKOVI P, CUCURULL G, CASANOVA A, et al. Graph attention networks[C]//52nd Annual Meeting of the Association for Computational Linguistics. Piscataway: IEEE,2017:125.

[15] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//2017 Advances in neural information processing systems. New York: IEEE Communications Society,2017:257.

[16] LUONG M T, PHAM H, MANNING C D. Effective approaches to attention-based neural machine translation[C]//5th International Conference on Learning Representations ICLR 2017 Conference Track Proceedings. Columbus: McGraw-Hill,2017:48.

[17] 陈雨龙,付乾坤,张岳.图神经网络在自然语言处理中的应用[J].中文信息学报,2021,35(3):1.

[18] KIPF T N, WELLING M. Variational graph auto-encoders[J]. Computer Science,2014,7(11):5.

[19] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: a lite bert for self-supervised learning of language representations[C]//Proceedings of the IEEE International Conference on Computer Vision. New York: IEEE, 2020:166.

[20] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural computation,1997,9(8):1735.

[21] MANNING C D, SURDEANU M, BAUER J, et al. The stanford corenlp natural language processing toolkit[C]// Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics. New York: IEEE,2014:98.