

知识图谱在智能问答中的应用研究

卢 琪, 谢艺菲, 谢 钧, 潘志松

(中国人民解放军陆军工程大学 指挥控制工程学院, 江苏 南京 210000)

摘 要:智能问答是自然语言处理领域的研究热点之一,近年来受到广泛的关注。随着知识图谱的不断发展,越来越多的知识图谱出现在人们视野当中,这些知识图谱在各个领域中得到广泛应用,能够作为智能问答的高质量数据源。知识图谱用于智能问答领域主要有两个方面:一方面,问答系统可以直接从知识图谱中检索答案;另一方面,知识图谱能够给智能问答系统提供外部信息,从而提高问答系统的性能,这两个方面都是研究的热点。文中简要介绍知识图谱及智能问答的基本概念和起源发展,综述当下知识图谱用于智能问答研究综述最新研究进展,总结了知识图谱用于智能问答在理论、算法和应用方面存在的挑战,讨论了显著的进步并概述了研究趋势,展望了知识图谱在智能问答领域的应用前景,以期推动知识图谱在智能问答领域的进一步研究。

关键词:知识图谱;智能问答;自然语言处理;语义解析;深度学习

中图分类号:TP183

文献标识码:A

文章编号:1673-629X(2021)07-0013-08

doi:10.3969/j.issn.1673-629X.2021.07.003

Research on Application of Knowledge Graphs in Intelligent Question Answering

LU Qi, XIE Yi-fei, XIE Jun, PAN Zhi-song

(School of Command & Control Engineering, Army Engineering University of PLA, Nanjing 210000, China)

Abstract: Intelligent question answering is one of the research hotspots in the field of natural language processing, which has attracted extensive attention in recent years. With the continuous development of knowledge graphs, more and more knowledge graphs appear in people's vision. These knowledge graphs are widely used in various fields and can be used as a high-quality data source for intelligent question answering. There are two main aspects of knowledge graph used in intelligent question answering system. On the one hand, the question answering system can directly retrieve answers from the knowledge graph. On the other hand, the knowledge graph can provide external information to the intelligent question answering system, thereby improving the performance of the question answering system. These two aspects are both research hotspots. The origin, development and basic concepts of knowledge graph and intelligent question answering are introduced, the latest research progresses of the technology of knowledge graph used in intelligent question answering are reviewed. The existing challenges of knowledge graph used in intelligent question answering in aspect of theory algorithm and application are summarized, the significant progress and outline research trends are discussed, and the application prospect of knowledge graphs in the field of intelligent question answering is prospected, in order to promote the further development of knowledge graphs for intelligent question answering.

Key words: knowledge graph; question answering; natural language processing; semantic parsing; deep learning

0 引 言

智能问答(question answering, QA)是自然语言处理(natural language processing, NLP)领域一个重要分支,经过多年的发展,已逐渐成为NLP中最为火热的研究方向之一。智能问答最早可以追溯到二十世纪五六十年代的图灵测试,要求计算机在有限时间内回答

用户给出的一系列问题,并且要让用户做出超过30%的误判。第一个问答系统^[1]一般认为是诞生于20世纪60年代的“Eliza”,用于对精神病人进行心理治疗。

按照数据可以把问答分为三种类型:

(1)基于文本问答:也称为机器阅读理解式(machine reading comprehension, MRC)问答,每个问

收稿日期:2020-08-25

修回日期:2020-12-28

基金项目:国家自然科学基金(62076251)

作者简介:卢 琪(1996-),男,硕士研究生,研究方向为知识图谱与智能问答;通信作者:谢 钧(1973-),男,博士,教授,研究方向为计算机网络与智能信息处理;潘志松,博士,教授,研究方向为模式识别。

题对应若干篇非结构化文本数据,然后从文本数据中检索和抽取答案;

(2) 基于知识库问答:也称为知识图谱问答(knowledge graph question answering, KGQA),即直接从构建好的结构化知识库中检索答案;

(3) 基于社区的问答:用户生成的问答对组成了社区问答的数据,例如百度知道、搜狗问答、知乎等论坛。

该文在简单介绍知识图谱和智能问答的基础上,总结归纳了知识图谱用于智能问答系统的研究进展和挑战,并讨论了新兴的研究趋势。

1 知识图谱

知识表示是知识组织的前提和基础。语义网便是早期知识表示的代表,它通过万维网联盟(world wide web consortium, W3C)标准来扩展万维网,使之变成一个数据网。

1.1 知识图谱问答

知识图谱问答的关键在于把用户的自然语言问题转化为机器可以理解的形式查询。传统的知识图谱问答方法主要包括三种:

(1) 基于语义解析的方法:把用户给出的自然语言问题转化成逻辑形式,在知识图谱上查询;

(2) 基于模板的方法:根据模板提取问题中的信息表示成特征向量,用分类器对问题特征向量进行筛选,得到答案;

(3) 基于向量建模的方法:把问题和候选答案用分布式表示,用分布式表示训练模型,使问题和正确答案的分数尽可能高。

随着深度学习领域不断发展,神经网络在 KGQA 中也取得了非常优秀的性能,得到广泛关注。

1.2 知识图谱和 KGQA 数据集

近年来出现众多大型的开源知识图谱,推动了知识图谱领域的快速发展。利用这些开源的知识图谱构建出各种大规模 KGQA 的数据集,也极大地促进了知识图谱问答的发展。该文将知识图谱数据集根据规模、语言、来源等特征进行整理,如表 1 所示。将知识图谱问答数据集根据规模、有无形式查询、来源等特征进行整理,如表 2 所示。这些数据集弥补过去数据集缺陷的同时,也提出新的挑战,为 KGQA 发展提供了研究基础。

表 1 知识图谱数据集汇总

知识图谱	规模	语言	来源	时间
ConceptNet ^[2]	390 万概念,2 800 万声明	跨语言	众包,文本抽取,游戏	2004
DBpedia ^[3]	2 800 万实体,95 亿三元组	跨语言	Wikipedia	2007
YAGO ^[4]	1 000 万实体,1.2 亿事实	跨语言	Wikipedia, WordNet, GeoNames	2007
Freebase ^[5]	230 万实体,1.25 亿关系	跨语言	Wikipedia	2008
BabelNet ^[6]	1 400 万词目,36.4 万词语关系	跨语言	Wikipedia, WordNet	2010
Wikidata ^[7]	8 000 万实体,10 亿关系	跨语言	Wikipedia	2014
Zhishi. me ^[8]	1 000 万实体,1.2 亿三元组	中	百度百科,互动百科,中文维基百科	2010
XLore ^[9]	1 000 万实体,66 万概念,5 万属性	中/英	中英文维基百科,百度百科	2013
Zhishi. schema ^[10]	40 万概念,150 万三元组	中	社交站点	2014
CN_DBpedia ^[11]	1 600 万实体,2 亿关系	中	百度百科,互动百科,中文维基百科	2017

表 2 知识图谱问答数据集汇总

数据集	规模	形式查询	来源	时间
QALD-1 (http://qald.aksu.org/)	100	有	DBpedia	2011
QALD-2 *	199	有	DBpedia, YAGO	2012
QALD-3 *	199	有	DBpedia, YAGO	2013
Free917 ^[19]	917	有	Freebase	2013
WebQuestions ^[20]	5 810	无	Freebase	2013
QALD-4 *	150	有	DBpedia, YAGO	2014
QALD-5 *	229	有	DBpedia, YAGO	2015
SimpleQuestions	108 442	无	Freebase, Reverb	2015
QALD-6 *	350	有	DBpedia, YAGO	2016

续表 2

数据集	规模	形式查询	来源	时间
WebQuestionsSP ^[21]	4 737	有	Freebase	2016
GraphQuestions ^[22]	5 166	有	Freebase	2016
ComplexQuestions ^[23]	2 100	有	Freebase	2016
QALD-7 *	250	有	DBpedia, Wikidata	2017
LC-QuAD ^[24]	5 000	有	DBpedia	2017
LC-QuAD 2 ^[25]	30 000	有	DBpedia, Wikidata	2019

2 智能问答

近年来智能问答取得了很大的发展,很多智能问答系统走进了人们的生活,为人们带来了极大的便利。苹果公司研发的智能语音助手 Siri 不仅能智能问答还可以对手机进行语音控制等,之后各大公司也推出了自己的语音助手或者问答系统。

2.1 机器阅读理解问答

机器阅读理解式问答又称为基于文本的问答系统,给定问题 $Q_i = (x_1^Q, x_2^Q, \dots, x_m^Q)$ 和对应的文章 $P_i = (x_1^P, x_2^P, \dots, x_n^P)$, 然后通过检索文章得到答案 $A_i = (a_1, a_2, \dots, a_j)$ 并抽取出来,其中 m, n, j 分别代表问题、文章和答案中的单词个数。

2.2 机器阅读理解问答数据集

2016 年斯坦福大学公布了 SQuAD^[12] 数据集,包含了涉及 500 篇文章的超过 10 万条问题-答案对, SQuAD 是一个抽取式问答数据集,答案被限定为段落中的一个连续子片段;同年,微软发布了 MARCO 数据集,包含 10 万个问题,每个问题有 10 个左右的相关段落, MARCO 是一个生成式问答数据集,答案词汇不一定来自于段落本身,而是由人工编写的。然而也有检索完文档发现无法回答问题的情况,这时候就希望模型能给出“Unanswerable”的答案,所以斯坦福大学又公布了 SQuAD2.0^[13] 数据集。TriviaQA^[14] 数据集包含超过 65 万个问题-答案-证据三元组,该数据集包含相对复杂的问题,简单的文本匹配方法无法适应该数据集。QAngaroo^[15] 数据集考虑到有些问题需要考虑多个文档才能回答,利用知识图谱技术构造了两个多跳阅读理解数据集,要求模型不仅需要正确回答出答案,还需要提供支持答案的证据。HotpotQA^[16] 数据集包含了 11.3 万个基于维基百科的问题-答案对,能够训练可执行多跳推理并提供答案支持的问答系统。中文方面,百度发布了大规模开放域数据集 DuReader^[17],包含 20 万个问题、100 万篇文章和 42 万个答案,这些问题和文章都来源于百度搜索引擎数据和问答社区。哈工大讯飞联合实验室发布的 CMRC 数据集^[18],该数据集包含专家在 Wikipedia 段落上标

注的近 2 万个真实问题。这些数据集的提出为问答系统的研究提供了极大的便利。

2.3 机器阅读理解问答系统

把近年机器阅读理解模型根据特征提取器、是否引入知识等特征进行整理,如表 3 所示。

神经网络由于其强大的表示能力,如今在问答系统中已取得很大的进展。基于神经网络的问答模型常用框架为“编码-交互-预测”,如图 1 所示。

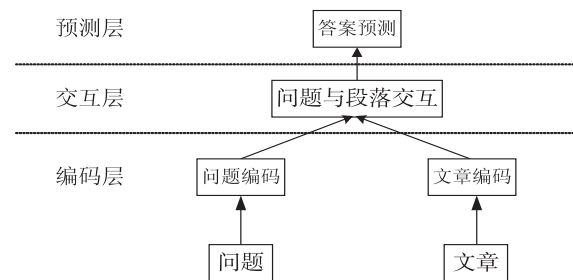


图 1 问答系统常用框架

编码层的作用是将问题和文章用固定维度的向量表示,对于问题 $Q_i = (x_1^Q, x_2^Q, \dots, x_m^Q)$ 和对应的文章 $P_i = (x_1^P, x_2^P, \dots, x_n^P)$, 在编码层中首先得到问题和文章的向量表示 $Q \in \mathbb{R}^{d \times m}$, $P \in \mathbb{R}^{d \times n}$, 其中 d 代表词向量的维度,常用的词向量有 Word2Vec 和 GloVe^[26] 等。接着经过如 LSTM、BiLSTM 或 CNN 等神经网络分别得到问题和文章的表示 $H^q = \text{LSTM}(Q)$, $H^p = \text{LSTM}(P)$, 其中 $H^p \in \mathbb{R}^{h \times n}$, $H^q \in \mathbb{R}^{h \times m}$, h 表示隐层向量的维数。

交互层的作用是把问题和文章的编码进行交互,首先计算文章和问题中每对单词之间的相似度:

$$S = \alpha (H^p)^T H^q \quad (1)$$

然后通过注意力机制或者 LSTM 得到交互后的表示。

预测层的作用是根据相似度预测答案,通过把候选答案在文章中的开始位置和结束位置的概率相乘得到该候选答案的分数。训练过程中的目标函数为:

$$L(\theta) = -\frac{1}{N} \sum_i [\log(p_{s_i}^1) + \log(p_{e_i}^2)] \quad (2)$$

其中, p^1, p^2 为开始和结束位置的概率, s_i, e_i 分别表示起始位置和结束位置的 groundtruth。

Vaswani 等人提出了 Transformer^[27] 架构,其中自注意力机制(self-attention)使得每个词都有全局的语义信息,长距离依赖关系的提取能力要强于 RNN。使用多头注意力机制在多个语义空间上进行表示,性能远好于 RNN 和 CNN。

3 知识图谱问答

KGQA 主流方法有基于语义解析(semantic parsing)的方法和基于信息检索这两种。下面分别介绍这两个主流方法的具体算法和发展。

3.1 基于语义解析的 KGQA

基于语义解析的 KGQA 方法把要解决的问题看作语义解析问题,即把自然语言问题转化成语义表示,再映射成逻辑形式。基于语义解析的 KGQA 可以看作答案能否回答问题的二分类任务,也可以看作候选答案的排序问题,随着编码-解码模型在翻译领域的广泛使用,也有学者使用翻译中的模型解决 KGQA 问题。

3.1.1 基于分类的 KGQA

分类任务即预测问题 q 中的关系属于 n 种关系 r_1, \dots, r_n 中的哪一类,可以分为三个步骤:

(1) 使用编码器把可变长度的输入 q 映射成固定维度的向量 $q \in \mathbf{R}^d$;

(2) 问题编码经过映射计算得到分数向量:

$$s(q) = W_o q + b_o$$

其中, $W_o \in \mathbf{R}^{n \times d}$, $b_o \in \mathbf{R}^n$ 。

(3) 输出层中,模型会基于 softmax 函数把分数向量转化成条件概率分布:

$$p(r_k | q) = \frac{e^{s_k(q)}}{\sum_{j=0}^n e^{s_j(q)}}$$

其中, $k = 1, 2, \dots, n$, 选择给定 q 时概率最高的关系。

对于有 N 个自然语言问题和形式查询对的数据集: $D = \{(q^{(i)}, f^{(i)})\}_{i=1}^N$, 其中 f 是实体-关系元组:

$$f^{(i)} = (e^{(i)}, r^{(i)})$$

使用最大化对数似然函数来训练模型:

$$\sum_{i=1}^N \log p_{\theta}(r^{(i)} | q^{(i)})$$

Mohammed 等人^[28]提出:对于简单问题,基于分类的系统使用最基本的神经网络结构(CNN、LSTM)加上一些简单的规则就能够达到 SOTA 性能。

3.1.2 基于排序的 KGQA

给定自然语言问题 q 和知识图谱时,基于排序的 KGQA 通常会使用一些搜索策略来寻找合适的候选形式查询集合: $C(q) = \{f_1, f_2, \dots, f_N\}$, 然后使用基于神经网络的排序模型选择最匹配的形式查询。一般有两个步骤:

(1) 问题 q 和候选形式查询 f 通过编码器映射到向量空间;

(2) 形式查询向量 f 和编码后的问题 q 一起送入评分函数,返回的分数 $s(q, f)$ 表示形式查询 f 和问题 q 的匹配程度。得分最高的形式查询作为模型的预测:

$$f^* = \operatorname{argmax}_{f \in C(q)} s_{\theta}(q, f)$$

Li 等人^[29]提出使用多列卷积神经网络(multi-column CNN)从答案路径、答案类型和答案上下文三个角度来表示问题,提取了更丰富的信息并且不依赖手工特征和规则。

3.1.3 基于编码-解码模型的 KGQA

序列到序列(sequence to sequence, Seq2Seq)模型^[30]主要由编码-解码架构和注意力机制组成,一些学者对模型进行改进,应用到基于语义解析的 KGQA 中。Seq2Seq 架构如图 2 所示,在编码器中,问题的词序列经过 embedding 层映射为向量,然后经过一个 RNN 结构,最后得到编码向量。在解码器中,每时刻的输入为“context vector”以及前一时刻预测词的向量 \hat{y}_{n-1} , 输出为当前时刻的预测词的向量,其中 <start> 和 <end> 为句子起始和终止标识符。

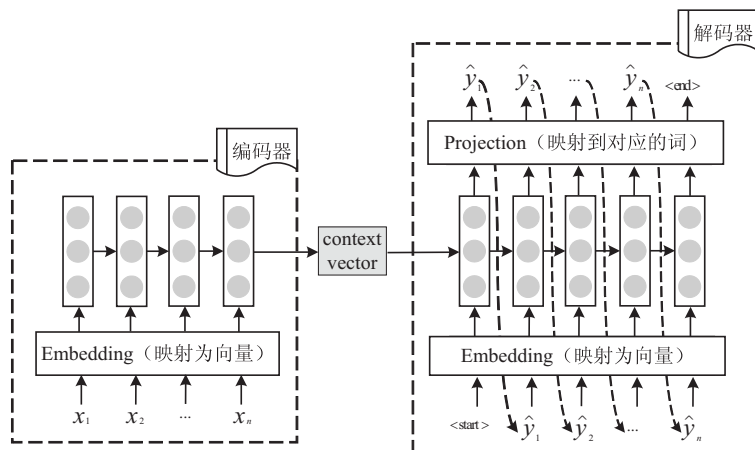


图 2 Seq2Seq 架构

Alvarezmelis 等人^[31]提出了一种改进型树状解码器,使用两个独立的 RNN 分别对父-子和兄弟-兄弟节点之间的信息流建模,提供了使用 RNN 从自然语言查询生成可执行的查询语言这一思路。

3.2 基于信息检索的 KGQA

基于信息检索的 KGQA 通常先确定用户查询中的中心实体,然后链接到知识图谱中确定相关实体得到候选答案集合,之后通过评分或者排序的方式找出最可能的答案。该方法不需要大量手工特征或者规则,将复杂语义问题转化为大规模可学习问题。基于信息检索的 KGQA 根据使用的方法可以分为基于特征工程的方法和基于深度学习的方法。

3.2.1 基于特征工程的方法

Yao 等人^[32]率先提出了该类方法的通用模型,作者首先对用户查询进行句法分析,然后把依存关系转化成更通用的问题特征图。接着利用问题主题词在知识图谱中提取相关的节点,构成主题图。最后把问题中的特征与主题图中的特征进行组合,通过分类器学习特征的权重。

Bast 等人^[33]提出了一种端到端的系统可以自动将自然语言问题转化成 SPARQL 查询语言。以交互方式回答问题,要求用户反馈关键决策,大大提高准确性。

3.2.2 基于深度学习的方法

基于特征工程的方法需要预先定义并抽取特征,受主观因素限制,并且难以处理复杂问题。而深度学习可以很好地解决这些问题,通过神经网络把问题和候选答案映射为空间向量,然后进行匹配。

Yih 等人^[34]使用卷积神经网络解决单关系问答。通过 CNN 构建两个不同的匹配模型,分别用来识别问题中出现的实体和匹配实体与 KG 中实体的相似度,最后给所有关系三元组打分,分数最高的三元组作为问题的答案。但是模型难以处理复杂的多关系情况。Hao 等人^[35]更关注问题的表示,提出了一种新的基于 cross-attention 的模型,根据不同的答案类型赋予问题中不同单词的权重,这种动态表示不仅精确而且更加灵活。

4 知识图谱用于机器阅读理解问答

除了知识图谱问答,知识图谱还能用于机器阅读理解问答。机器阅读理解要求从给定的文章中提取信息回答问题,而当人类在做阅读理解任务时,利用给定上下文回答问题的同时,也会利用一些先验知识。

百度 Sun 等人发布的 ERNIE (enhanced representation from knowledge integration) 模型,在预训练时引入了多源数据知识:百科类、新闻资讯类以及论坛对话类数据,通过建模这些海量数据中的词、实体以及实体之间的关系,把知识编码到预训练模型中,增强了模型的语义表示能力。

清华大学 Zhang 等人提出 ERNIE (enhanced language representation with informative entities) 模型,不同于百度通过 MASK 的方法隐式地引入知识,清华的 ERNIE 通过改进 BERT 模型结构,将知识和语义信息显式地在预训练时进行编码学习。模型在编码过程中引入一个聚合器 (aggregator) 用于知识编码以及知识融合,有效解决了结构化的知识编码和异质信息融合问题。

Liu 等人提出了 K-BERT (knowledge-enabled bi-directional encoder representation from transformers) 模型,在模型中引入了软定位和可见矩阵来限制知识的影响。首先把句子中提到所有的命名实体提取出来去知识图谱中查询对应的三元组,然后把提取的三元组引入句子中生成句子树,以此来引入知识,提高模型的表达能力。

上面的这三种模型都是在模型预训练阶段引入知识图谱,Yang 等人提出了 KT-Net (knowledge and text fusion net),模型的知识整合 (knowledge integration) 模块在面向下游任务的微调阶段引入知识图谱中的信息。

其中知识表示是在整个知识图谱上学习的嵌入表示,能够捕获整个知识图谱的全局信息,并且知识融合也易于扩展至融合多个知识图谱的信息。上述提到的把知识图谱信息引入机器阅读理解的模型也汇总到表 3 中。

表 3 机器阅读理解问答模型对比

模型	特征抽取器	引入知识	贡献	时间
GPT1.0 ^[36]	单向 Transformer	否	首个基于 Transformer 的生成式预训练模型	2018
ELMo	BiLSTM	否	使用两个独立的单向 LSTM 获取双向语义信息	2018
BERT ^[37]	双向 Transformer	否	首次使用双向 Transformer 编码,增加 MLM 和 NSP 任务提高模型预训练效果	2018
GPT2.0	单向 Transformer	否	使用更大的模型和更大规模的预训练数据,采用无监督多任务联合训练	2019

续表 3

模型	特征抽取器	引入知识	贡献	时间
XLNet ^[38]	双向 Transformer-XL	否	提出排列语言模型, 基于 Transformer-XL 使用片段递归机制和相对位置编码获得长距离依赖	2019
BERT-WWM	双向 Transformer	否	引入全词掩码, 有效改善中文数据集上的性能	2019
ALBERT ^[39]	双向 Transformer	否	采用嵌入层因式分解和跨层参数共享大幅减少参数量	2019
RoBERTa	双向 Transformer	否	引入动态掩码, 摒弃 NSP 任务, 采用更大规模数据集	2019
ELECTRA ^[40]	双向 Transformer	否	引入对抗学习思想, 采用 RTD 任务, 计算效率大幅提升	2020
Longformer	双向 Transformer	否	改进 Transformer 机制, 只针对固定窗口大小附近的内容计算局部注意力	2020
KBLSTM ^[41]	BiLSTM	是	使用外部知识改进 RNN 性能, 提出带有哨兵的注意力机制, 适应性引入知识	2017
ERNIE1.0	双向 Transformer	是	采用单字级别、短语级别和实体级别的掩码, 引入外部知识	2019
ERNIE2.0 ^[42]	双向 Transformer	是	采用多任务学习和持续学习机制, 包括词级别、结构级别和语义级别的任务	2019
ERNIE ^[43]	双向 Transformer	是	显式的引入聚合器用于编码和知识融合, 有效解决结构化知识编码和异质信息融合	2019
K-BERT ^[44]	双向 Transformer	是	把知识信息引入句子中并完成句子树转换, 使用软定位和可见矩阵限制知识的影响	2019
KT-Net ^[45]	双向 Transformer	是	在整个知识图谱上学习知识嵌入表示, 捕获全局信息, 微调阶段引入知识	2019

5 挑战与研究方向

近年来国内外涌现出多种把知识图谱应用于问答的方法, 在取得很大进展的同时仍然存在不同方面的挑战。本章简单介绍当前研究的几个瓶颈问题, 并提出下一步的研究方向。

5.1 存在的挑战

(1) 知识图谱的数据问题。知识图谱数据的质量显著影响着问答系统的性能, 如何确保知识图谱中数据广而准确非常关键。目前知识图谱大多存在噪声、数据稀疏、数据冗余等问题, 这些问题会给问答系统带来影响。

(2) 用户复杂查询转化成逻辑形式的问题。KGQA 中的一大难点便是如何把用户提出的自然语言查询转化成机器可以理解的逻辑形式。实际应用中用户提出的往往是复杂问题, 如何得到复杂问题的通用解决范式也是一大挑战。

(3) 机器阅读理解问答引入知识的问题。在长尾问题、少样本问题、样本不均衡问题等背景下, 引入哪些知识, 如何引入知识是需要探索的方向, 更具体的, 使用何种方法把知识和文本语义信息融合起来, 判断哪些知识是相关的都是值得探索的。

5.2 下一步研究方向

(1) 尝试多个知识图谱信息融合。针对知识图谱数据质量低下的问题, 考虑结合多个知识图谱, 保留不冲突、不重复的信息, 可以一定程度上解决数据稀疏问

题。同时可以检测知识图谱之间的冲突信息, 并进行消解和避免, 以此解决噪声问题。

(2) 借助于不断涌现的语言模型强大的表示能力, 从语义层面解析问题。近年随着语言模型的飞速发展, 使用语言模型得到问题的表示, 再到知识图谱中查询会是未来的研究方向。

(3) 知识图谱用于多跳问答推理。多跳问答是机器阅读理解问答的一大研究热点, 回答多跳问答需要结合多篇文章甚至外部知识, 如何把知识图谱中的信息有效利用起来会是接下来的研究方向。

(4) 基于迁移学习的知识图谱问答系统。由于某些领域没有知识图谱或者知识图谱规模小, 大量有标注的样本难以获得, 考虑使用迁移学习解决这一问题。

6 结束语

该文从两个方面介绍了知识图谱在智能问答中的应用, 对比了多个常用的数据集, 概括了主流的方法以及存在的问题。智能问答作为自然语言处理的一个重要分支, 能够从海量数据中简明扼要地给出用户需要的答案, 便于用户获得精准信息。而知识图谱作为自然语言处理的另一个重要分支, 被广泛地应用于智能问答、推荐系统、搜索引擎以及辅助决策等领域, 也是实现自然语言理解的重要一环。实现知识图谱和智能问答的有机结合, 则有望让机器像人类一样去理解和回答问题, 是使机器实现知识应用并能够与真实世界交互的关键环节。因此, 在自然语言处理领域飞速发

展的过程中,把知识图谱应用于智能问答研究意义重大。

参考文献:

- [1] 王树西. 问答系统:核心技术,发展趋势[J]. 计算机工程与应用,2005,41(18):1-3.
- [2] LIU H, SINGH P. Commonsense reasoning in and over natural language[C]//Knowledge-based and intelligent information and engineering systems. Wellington, New Zealand: Springer,2004:293-306.
- [3] AUER S, BIZER C, KOBILAROV G, et al. Dbpedia: a nucleus for a web of open data[C]//The semantic web. Busan, Korea: Springer,2007:722-735.
- [4] SUCHANEK F M, KASNECI G, WEIKUM G. Yago: a core of semantic knowledge[C]//Proceedings of the 16th international conference on World Wide Web. Banff, Alberta, Canada: ACM,2007:697-706.
- [5] BOLLACKER K, EVANS C, PARITOSH P, et al. Freebase: a collaboratively created graph database for structuring human knowledge[C]//Sigmod conference. Vancouver, Canada: ACM,2008:1247-1250.
- [6] NAVIGLI R, PONZETTO S P. BabelNet: building a very large multilingual semantic network[C]//Meeting of the association for computational linguistics. Stroudsburg, PA: ACL,2010:216-225.
- [7] VRANDECIC D, KROTZSCH M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM,2014, 57(10):78-85.
- [8] NIU X, SUN X, WANG H, et al. Zhishi. me: weaving chinese linking open data[C]//The semantic web. Bonn, Germany: Springer,2011:205-220.
- [9] WANG Z, LI J, WANG Z, et al. XLORE: a large-scale English-Chinese bilingual knowledge graph[C]//International semantic web conference. Sydney, Australia: CEUR - WS. org,2013:121-124.
- [10] WANG H, WU T, QI G, et al. On publishing Chinese linked open schema[C]//The semantic web. Riva del Garda, Italy: Springer,2014:293-308.
- [11] XU B, XU Y, LIANG J, et al. CN-DBpedia: a never-ending Chinese knowledge extraction system[C]//Advances in artificial intelligence: from theory to practice. Arras, France: Springer,2017:428-438.
- [12] RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100,000+ questions for machine comprehension of text[C]//Proceedings of the 2016 conference on empirical methods in natural language processing. Stroudsburg, PA: ACL, 2016: 2383-2392.
- [13] JIA R, RAJPURKAR P, LIANG P. Know what you don't know: unanswerable questions for SQuAD[C]//Proceedings of the 56th annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL,2018:784-789.
- [14] JOSHI M, CHOI E, WELD D S, et al. Triviaqa: a large scale distantly supervised challenge dataset for reading comprehension[C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL,2017:1601-1611.
- [15] WELBL J, STENETORP P, RIEDEL S. Constructing datasets for multi-hop reading comprehension across documents[J]. Transactions of the Association for Computational Linguistics,2018,6:287-302.
- [16] YANG Z, QI P, ZHANG S, et al. HotpotQA: a dataset for diverse, explainable multi-hop question answering[C]//Proceedings of the 2018 conference on empirical methods in natural language processing. Stroudsburg, PA: ACL,2018:2369-2380.
- [17] HE W, LIU K, LYU Y, et al. DuReader: a Chinese machine reading comprehension dataset from real-world applications[C]//Proceedings of the workshop on machine reading for question answering. Stroudsburg, PA: ACL,2017:37-46.
- [18] CUI Y, LIU T, CHE W, et al. A span-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of the 2019 conference on empirical methods in natural language processing. Stroudsburg, PA: ACL,2018:5882-5888.
- [19] CAI Q, YATES A. Large-scale semantic parsing via schema matching and lexicon extension[C]//Meeting of the association for computational linguistics. Stroudsburg, PA: ACL, 2013:423-433.
- [20] BERANT J, CHOU A K, FROSTIG R, et al. Semantic parsing on freebase from question-answer pairs[C]//Empirical methods in natural language processing. Stroudsburg, PA: ACL,2013:1533-1544.
- [21] YIH W, RICHARDSON M, MEEK C, et al. The value of semantic parse labeling for knowledge base question answering[C]//Meeting of the association for computational linguistics. Stroudsburg, PA: ACL,2016:201-206.
- [22] SU Y, SUN H, SADLER B M, et al. On generating characteristic-rich question sets for QA evaluation[C]//Empirical methods in natural language processing. Stroudsburg, PA: ACL,2016:562-572.
- [23] BAO J, DUAN N, YAN Z, et al. Constraint-based question answering with knowledge graph[C]//International conference on computational linguistics. Osaka: The COLING 2016 Organizing Committee,2016:2503-2514.
- [24] TRIVEDI P, MAHESHWARI G, DUBEY M, et al. LC-QuAD: a corpus for complex question answering over knowledge graphs[C]//The semantic web - ISWC 2017. Vienna, Austria: Springer,2017:210-218.
- [25] DUBEY M, BANERJEE D, ABDELKAWI A, et al. LC-QuAD 2.0: a large dataset for complex question answering over Wikidata and DBpedia[C]//The semantic web. Auckland, New Zealand: Springer,2019:69-78.
- [26] PENNINGTON J, SOCHER R, MANNING C. Glove: global

- vectors for word representation [C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Stroudsburg, PA: ACL, 2014: 1532–1543.
- [27] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need [C]//Advances in neural information processing systems 30. Red Hook, NY: Curran Associates Inc., 2017: 5998–6008.
- [28] MOHAMMED S, PENG S, LIN J, et al. Strong baselines for simple question answering over knowledge graphs with and without neural networks [C]//Proceedings of the 2018 conference of the north american chapter of the association for computational linguistics: human language technologies. New Orleans, Louisiana: ACL, 2018: 291–296.
- [29] LI D, WEI F, MING Z, et al. Question answering over freebase with multi-column convolutional neural networks [C]//Meeting of the association for computational linguistics & international joint conference on natural language processing. Stroudsburg, PA: ACL, 2015: 260–269.
- [30] SUTSKEVER I, VINYALS O, LE Q V. Sequence to sequence learning with neural networks [C]//Advances in neural information processing systems 27. Cambridge, MA: MIT Press, 2014: 3104–3112.
- [31] ALVAREZMELIS D, JAAKKOLA T S. Tree-structured decoding with doubly-recurrent neural networks [C]//5th international conference on learning representations. Toulon, France: OpenReview. net, 2017.
- [32] YAO X, VAN DURME B. Information extraction over structured data: question answering with freebase [C]//Proceedings of the 52nd annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL, 2014: 956–966.
- [33] BAST H, HAUSMANN E. More accurate question answering on freebase [C]//Proceedings of the 24th ACM international conference on information and knowledge management. Melbourne, Australia: ACM, 2015: 1431–1440.
- [34] YIH W, HE X, MEEK C. Semantic parsing for single-relation question answering [C]//Proceedings of the 52nd annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL, 2014: 643–648.
- [35] HAO Y, ZHANG Y, LIU K, et al. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge [C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL, 2017: 221–231.
- [36] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training [EB/OL]. 2018. <https://www.cs.ubc.ca/~amuham01/LING530/papers/radford2018improving.pdf>.
- [37] DEVLIN J, CHANG M W, 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. Minneapolis, MN: ACL, 2018: 4171–4186.
- [38] YANG Z, DAI Z, YANG Y, et al. XLNet: generalized autoregressive pretraining for language understanding [C]//Advances in neural information processing systems 32: annual conference on neural information processing systems 2019. Vancouver, BC: Curran Associates Inc., 2019: 5754–5764.
- [39] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite bert for self-supervised learning of language representations [C]//8th international conference on learning representations. Addis Ababa, Ethiopia: OpenReview. net, 2019.
- [40] CLARK K, LUONG M, LE Q V, et al. ELECTRA: pre-training text encoders as discriminators rather than generators [C]//8th international conference on learning representations. Addis Ababa, Ethiopia: OpenReview. net, 2020.
- [41] YANG B, MITCHELL T M. Leveraging knowledge bases in LSTMs for improving machine reading [C]//Proceedings of the 55th annual meeting of the association for computational linguistics. Stroudsburg, PA: ACL, 2017: 1436–1446.
- [42] SUN Y, WANG S, LI Y, et al. ERNIE 2.0: a continual pre-training framework for language understanding [C]//The thirty-fourth AAAI conference on artificial intelligence. Menlo Park, CA: AAAI Press, 2019: 8968–8975.
- [43] ZHANG Z, HAN X, LIU Z, et al. ERNIE: enhanced language representation with informative entities [C]//Proceedings of the 57th conference of the association for computational linguistics. Stroudsburg, PA: ACL, 2019: 1441–1451.
- [44] LIU W, ZHOU P, ZHAO Z, et al. K-BERT: enabling language representation with knowledge graph [C]//The thirty-fourth AAAI conference on artificial intelligence. Menlo Park, CA: AAAI Press, 2019: 2901–2908.
- [45] YANG A, WANG Q, LIU J, et al. Enhancing pre-trained language representations with rich knowledge for machine reading comprehension [C]//Proceedings of the 57th conference of the association for computational linguistics. Stroudsburg, PA: ACL, 2019: 2346–2357.