智能系统学报
CAAI TRANSACTIONS ON INTELLIGENT SYSTEMS

基于知识图谱、TF-IDF和BERT模型的冬奥知识问答系统

罗玲， 李硕凯， 何清， 杨骋骐， 王宇洋恒， 陈天宇

引用本文:
罗玲, 李硕凯, 何清, 等. 基于知识图谱、TF-IDF和BERT模型的冬奥知识问答系统[J]. 智能系统学报, 2021, 16(4): 819-826.
LUO Ling, LI Shuokai, HE Qing, et al. Winter Olympic Q & A system based on knowledge map, TF-IDF and BERT model[J]. *CAAI Transactions on Intelligent Systems*, 2021, 16(4): 819-826.

在线阅读 View online: https://dx.doi.org/10.11992/tis.202105047

DOI: 10.11992/tis.202105047

基于知识图谱、TF-IDF 和 BERT 模型的冬奥知识问答系统

罗玲[1,2], 李硕凯[1,2], 何清[1,2], 杨骋骐[2], 王宇洋恒[2], 陈天宇[2]

(1. 中国科学院计算技术研究所 智能信息处理重点实验室, 北京 100190; 2. 中国科学院大学, 北京 100049)

摘 要: 传统信息检索技术已经不能满足人们对信息获取效率的要求, 智能问答系统应运而生, 并成为自然语言处理领域一个非常重要的研究热点。本文针对中文的冬奥问答领域, 提出了基于知识图谱、词频-逆文本频率指数 (term frequency-inverse document frequency,TF-IDF) 和自注意力机制的双向编码表示 (bidirectional encoder representation from transformers,BERT) 的 3 种冬奥问答系统模型。本文首次构建了冬奥问答数据集, 并将上述 3 种方法集成在一起, 应用于冬奥问答领域, 用户可以使用本系统来快速准确地获取冬奥内容相关的问答知识。进一步, 对 3 种模型的效果进行了测评, 测量了 3 种模型各自的回答可接受率。实验结果显示 BERT 模型的整体效果略优于知识图谱和 TDIDF 模型, BERT 模型对 3 类问题的回答可接受率都超过了 96%, 知识图谱和 TDIDF 模型对于复合统计问答对的回答效果不如 BERT 模型。

关键词: 智能问答; 冬奥问答; 对话模型; 知识图谱; TF-IDF; BERT

中图分类号: TP391 **文献标志码:** A **文章编号:** 1673−4785(2021)04−0819−08

中文引用格式: 罗玲, 李硕凯, 何清, 等. 基于知识图谱、TF-IDF 和 BERT 模型的冬奥知识问答系统 [J]. 智能系统学报, 2021, 16(4): 819–826.

英文引用格式: LUO Ling, LI Shuokai, HE Qing, et al. Winter Olympic Q & A system based on knowledge map, TF-IDF and BERT model[J]. CAAI transactions on intelligent systems, 2021, 16(4): 819–826.

Winter Olympic Q & A system based on knowledge map, TF-IDF and BERT model

LUO Ling[1,2], LI Shuokai[1,2], HE Qing[1,2], YANG Chengqi[2], WANG Yuyangheng[2], CHEN Tianyu[2]

(1. Key Lab of Intelligent Information Processing, Institute of Computing Technology of Chinese Academy of Sciences, Beijing 100190, China; 2. University of Chinese Academy of Sciences, Beijing 100049, China)

Abstract: With the advent of the information age, traditional information retrieval technology can no longer meet people's requirements for the efficiency in information acquisition, so intelligent question answering systems are proposed and have become a very important research hotspot in natural language processing. This paper proposes three Winter Olympics Q&A system models based on knowledge graph, TFIDF and BERT for the Chinese Winter Olympics Q&A, constructing the Winter Olympics Q&A data set for the first time and integrating the above three methods into the Winter Olympics Q&A. Users can use this system to quickly and accurately obtain the Q&A knowledge related to the Winter Olympics content. Furthermore, this paper evaluates the effects of the three models and measures the acceptance rate of each model. The experimental results show that overall the BERT model is slightly better than the knowledge graph and TDIDF model. The acceptance rate of the BERT model for each of the three types of questions exceeds 96%. The knowledge graph and TDIDF model are not so effective as the BERT model for the answer to the composite statistical question and answer pair.

Keywords: Intelligent Q & A; Winter Olympics Q & A; dialogue model; knowledge map; TF-IDF; BERT

网络是当今世界人们获取信息的一个重要途径。随着信息化时代的到来, 网络信息变得更庞大且杂乱无章, 传统的搜索引擎已经难以满足人们的需要。得益于人工智能技术的飞速发展, 智能问答系统应运而生[1]。智能问答系统旨在帮助人们在海量信息中快速而准确地找到自己需要的信息。智能问答系统响应用户提问的效果要明显好于当今流行的基于布尔检索技术的搜索引擎, 提供的回答也更加快捷准确。智能问答是指通过

收稿日期: 2021−05−31.
基金项目: 国家重点研发计划项目 (2017YFB1002104).
通信作者: 何清. E-mail: heqing@ict.ac.cn.

人工智能、知识图谱等技术建立的以一问一答形式精确定位网站用户所需要的知识，通过与网站用户进行交互，为网站用户提供个性化的信息服务的问答系统[2]。

自第一届冬季奥林匹克运动会于 1924 年在法国的夏慕尼举行以来，冬奥会至今已有近一百年的历史。2022 北京冬奥会即将来临，人们也都很想了解这一百年来冬奥会的举办和获奖情况。但是由于参加冬奥会的运动员数量大、比赛项目多导致人们获取准确答案变得不容易。为了解决这个问题，帮助人们快速高效地获取需要的冬奥会的有关信息，本文提出了冬奥问答系统。该系统首先收录了从第一届冬奥会至今所有冬奥会的举办信息和获奖信息，然后对比了 3 种问答方法，并对其优劣进行了比较。本文的贡献主要分为如下 3 个部分：

1) 利用爬虫爬取了有关冬奥会问答的事实性句子，包含运动员的姓名、性别、年龄、身高、体重、国家、参与奥运会的年份、参加奥运会的项目、获奖情况等信息。并自己定义模板，将事实性信息转化为问答对，整合后提出了冬奥问答数据集，包含了冬奥会的事实性问题。

2) 将智能问答系统应用在冬奥项目数据上，分别基于知识图谱、基于自注意力机制的双向编码表示 (bidirectional encoder representation from transformers,BERT) 模型和词频−逆文本频率指数 (term frequency-inverse document frequency,TF-IDF) 模型建立了冬奥问答系统，根据用户提供的自然语言形式的问题可以给出准确的回答。

3) 本文使用 3 种模型建立冬奥问答系统，并比较了 3 种模型的正确率和优缺点，实验结果达到了预期目标。

1　相关工作

1.1　对话系统

智能对话系统是人工智能领域的核心技术，是人机交互的重要研究方向，对话系统的基础是问答系统。问答系统是信息检索系统的高级形式之一，采用自然语言处理技术，可以实现对用户疑问的理解，进而完成答案的生成[3]。问答系统首先接受自然语言的问句输入进行问句分析，这一阶段的主要任务是完成对问句语义的理解，将自然语言转为逻辑语言，在问句分析后进行信息检索和直接答案输出。问答系统主要分为信息检索式问答系统和生成式的问答系统[4]，前者通过判断输入问句在知识库中匹配对应答案，后者则基于模型训练生成答句。Yao 等[5] 实现了一种实际工程应用中的基于深度学习模型的任务导向型对话系统的通用框架。Feng 等[6] 实现了一个不依赖语言的基于卷积神经网络的口语问答系统，基于问题和训练集中距离的度量，返回度量值最高的问答对。Zhang 等[7] 实现了一个基于知识库的开放领域问答系统，该系统采用自定义词典分词和条件随机场模型 CRF 相结合的方法识别问句中的主体，采用模糊匹配方法将问句中的主体和知识库中的实体建立连接，系统平均 F-Measure (F 值) 达到 0.695 6。Noraset 等[8] 实现了一个基于泰语的问答系统。Höffner 等[9] 调查分析了 62 个不同的语义问答 (SQA) 系统。基于他们的分析，本文选出 3 种方法，在中文冬奥问答数据集上进行比较。

1.2　知识图谱

1977 年，知识工程概念在第五届国际人工智能大会上被提出，随即知识库系统的研究开始进入人们的视野[10]。Google 公司于 2012 年 11 月提出了知识图谱概念，并表示在其搜索结果中加入知识图谱功能。知识图谱概念一经提出，就逐渐成为热门，进入蓬勃发展阶段。知识图谱模型基于符号化通过三元组表达具体知识，并且以有向图的形式进行存储链接，在问答系统、搜索、推荐等领域有着广泛的应用。Liu 等[11] 从知识图谱的定义和技术架构出发,对构建知识图谱涉及的关键技术进行了自底向上的全面解析。Xu 等[12] 探索了一种基于知识图谱的多轮问答系统可实现方案。Chen 等[13] 应用知识图谱通过结合其领域词表、规范等内容实现了一个可用于查询数据、进行知识问答的智能系统。Piotr[14] 提出了一个开放域因子式问答系统，引入并实现了深度实体识别。它允许基于先前组装的实体库，全面搜索与给定词网 WordNet 语法集匹配的所有形式的实体引用。Yih 等[15] 提出了一个三阶段的搜素方法，提高了知识图谱搜素问题的准确度。Jia 等[16] 引用知识图谱模型和用户长短期偏好提出了一种个性化景点推荐方法，能够预测并返回用户可能感兴趣的推荐列表。

1.3　TF-IDF 模型

TF-IDF 是一种针对关键词的统计分析方法，旨在通过判断某一词汇在测试语句和语料库中的出现次数来判断词的重要程度。

TF-IDF 是一种常用于信息检索的加权技术，有着广泛的应用：在对话领域，Lu 等[17] 基于向量空间的 TF-IDF 相似度模型,实现了面向服务机器人的口语对话系统；在知识图谱领域，Zhang 等[18] 基于 TF-IDF 模型通过关键词特征分析和共现矩阵分析,从而更好地构建全面从严治党知识图谱，加强对习近平新时代中国特色社会主义思想的学习与理解；在 SQL 注入检测领域，Su 等[19] 基于 TF-IDF 和 N-Gram 提出了一种 SQL 注入检测方

法, 在保证召回率的基础上,可接受率有所提高; 在舆情挖掘领域, Liu 等[20] 提出了基于 TF-IDF 权重分析法的网络舆情信息挖掘技术, 提高了网络舆情信息挖掘效果, 增强舆情引导的准确性; 在垃圾短信分类领域, Wu 等[21] 提出了一种结合 TF-IDF 的基于自注意力机制的双向长短期记忆网络模型来进行垃圾短信识别, 这种模型相比于传统分类模型的短信文本识别可接受率提高了 2.1%~4.6%,运行时间减少了 0.6~10.2 s; 在文本相似度度量领域, Li 等等[22] 提出了一种运用 TF-IDF 方法提取文本关键词的文本相似性度度量方法, 准确率高, 且时间效率比其他方法更高。

1.4　BERT 模型

BERT 是一种预训练语言表示的方法, 在大量文本语料上训练了一个通用的“语言理解”模型, 然后用这个模型去执行具体的自然语言处理任务。BERT 可以通过加深网络的方式增强对文本信息的挖掘能力。另外, BERT 基于无监督的语料数据进行学习的, 可以减少数据搜集和人工标注的成本。

BERT 在专利匹配、中文语义匹配、金融文本情感分析、中文地址分词、问答等领域有着广泛的应用: 在专利匹配领域, Cao 等[23] 提出了一种基于 BERT+注意力机制和基于密度聚类 (DBSCAN) 的长三角专利匹配算法, 有助于分析和研究长三角地区的专利情况; 在中文语义匹配领域, Wu 等[24] 提出一种基于 BERT 模型的用于问答系统的中文语义匹配算法, 实现了高效准确的语义匹配, 显著提高文本搜索、问答匹配的效率; 在金融文本情感分析领域, Zhu 等[25] 提出基于金融领域的全词覆盖与特征增强的 BERT 预处理模型, 显著提高了金融文本情感分析的可接受率和召回率; 在中文地址分词领域, Sun 等[26] 提出了一种基于 BERT 的中文地址分词方法, 将非行政级别的地址标签进行重新设计, 将中文地址分词任务转换为命名实体识别任务, 提取出正确的地址级别; 在问答领域, Peng 等[27] 提出了基于 BERT 的三阶段式问答模型, 该模型相较于同类基准模型, 抽取出的答案片段更加准确。

不难看出, BERT 在自然语言处理领域有着十分广泛的应用, 在各种任务上都有着不错的表现。

2　3 种模型的结构

2.1　BERT 模型

BERT 模型由 Google 公司在 2018 年发布。BERT 的网络架构使用的是《Attention is all you need》中提出的多层自注意力机制 Transformer 结构, 其最大的特点是抛弃了传统的循环神经网络和卷积神经网络, 通过注意力机制将任意位置的两个单词的距离转换成 1, 有效地解决了自然语言处理中棘手的长期依赖问题。多层自注意力机制 Transformer 的网络架构如图 1[28] 所示, 它是一个序列到序列的结构, 由若干个编码器和解码器堆叠形成。

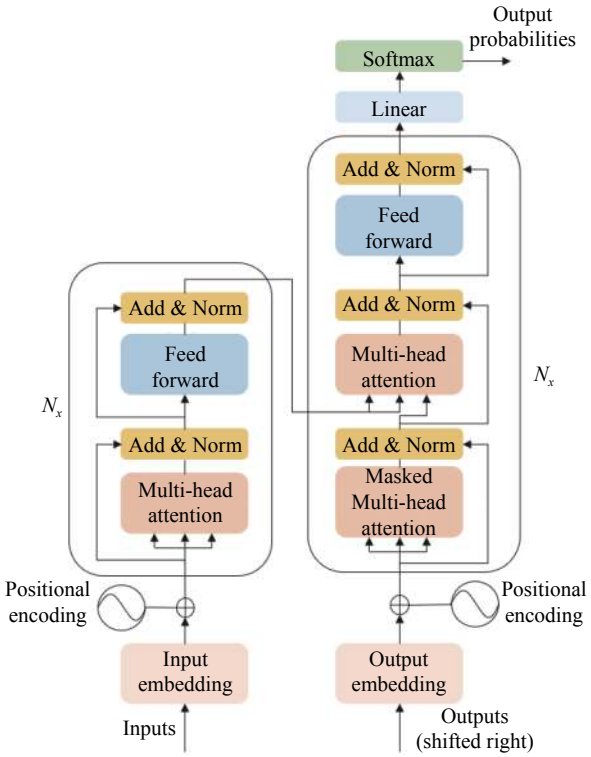

图 1　BERT 模型的网络架构

Fig. 1　Encoder-diverter-decoder with auxiliary information

模型的主要创新点在 pre-train 方法上, 使用了 Masked LM 和 Next sentence prediction 两种方法分别捕捉词语和句子级别的 representation。图 2~4 给出了其模型结构与另外两种著名的模型: 生成式预训练 (Generative pre-training, GPT) 和从语言模型中学得词嵌入 (Embeddings from language models, ELMO) 对比效果。其中 Trm 表示 Transformer, LSTM 表示长短期记忆网络, Ei 表示输入, Ti 表示输出。对比 OpenAI GPT, BERT 是双向的多层自注意力机制连接; 就像单向循环神经网络和双向循环神经网络的区别, 直觉上来讲效果会好一些。对比 ELMo, 虽然都是双向, 但是目标函数是不同的。

本文的模型整体使用“检索式对话系统”的思路, 以关键词和句子的相似度为指标将测试集的问句与训练集的问句进行匹配, 并取和测试集问句“最相似”的问句的答案作为最终答案输出。模型设计要点如下:

1) 生成句向量

生成句向量由很多种方式, 如使用 sklearn 词袋模型和 word2vec 模型等, 本文选择了 BERT 的

简单模型，调用 BERT_serving.client 库中的 BERTClient 函数对每个句子进行特征提取，维度为 768。

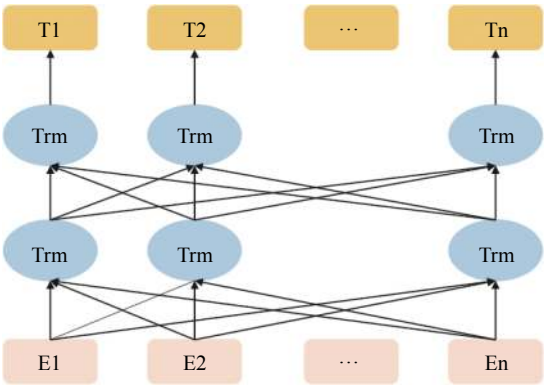

图 2　BERT 模型结构
Fig. 2　The architecture of BERT

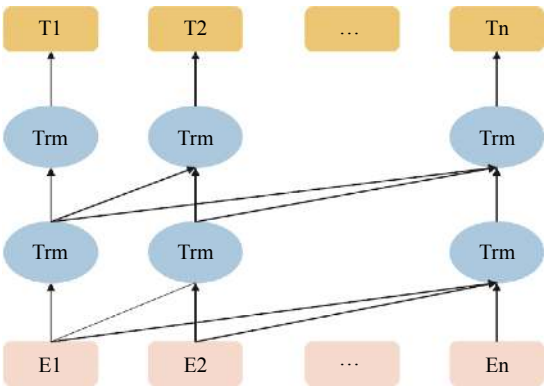

图 3　GPT 模型结构
Fig. 3　The architecture of GPT

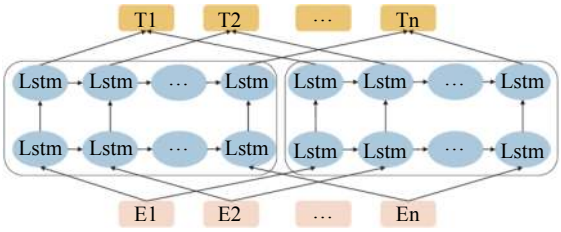

图 4　ELMo 模型结构
Fig. 4　The architecture of ELMo

2) 匹配特殊关键词的数量

由于 BERT 模型直接将输入的句子进行特征提取生成句向量，采用的语料库不完全是冬奥问答领域的，因此在冬奥会的专业领域特征提取的结果可能不够准确。为了弥补这一缺陷，实验在匹配含有测试问句中关键词的训练问题库中的问题时加入了优先特殊关键词 (人名、地名和时间)。

为了减少运算量并减少问题句式的不同对结果造成的影响，实验时手动挑选了若干个特殊关键词存在文件里，这些特殊关键词包括人名、地名和时间。在匹配的最开始，对于问句中的每一个关键词，查询包含它的训练集句子，比对它和问句之间相同的特殊关键词，并维护一个变量来表示当前和问句的最大特殊关键词匹配数，同时将对应的训练集问句存储在列表 L 中。如果存在多个和问句的特殊关键词匹配数一样的句子，那么就会进入下一步的精细选择：首先根据提问句子中的每一个关键词 item，查询训练集中包含这个关键词 item 的句子的 ID，再获取这个句子的关键字数组，然后将这个 item 的关键词匹配数置为 0。如果这个 item 在对应的训练问句的关键字数组里，就继续判断该 item 是否是特殊关键词，如果是就将它的关键词匹配数加 1。统计完之后判断当前的关键词匹配数是否大于先前维护的最大关键词匹配数，如果的确更大那么就将最大匹配数更新并且清空存储列表 L，将当前句子的 ID 存储到其中。这样得到的存储列表 L 就是初次选择的和问句“最像”的句子的集合。

3) 取句向量相似度最大的句子作为结果

在初步选择了特殊关键词匹配度最高的一些句子之后，需要在这些句子中选出和问句“更像”的句子来提供最终的答案。所以需要将问句都转化为特征向量，然后对特征向量采取一些运算并比较后得到最终的结果，具体步骤为：对于存储列表 L 中每一个句子的 ID，得到这个 ID 对应的句子的特征向量，与提问句的特征向量一起输入得分函数计算两个句子的相似度得分，之后循环统计最大的得分和对应的训练集句子的 ID，然后返回这个句子的 ID。在这里本文设计了一个“门槛”来过滤掉无关的句子：如果最大的相似度计算出来都低于 70%，那么就认为这个句子不存在答案，返回“暂时没有合适的答案”。

需要说明的是本文利用余弦函数计算句向量之间的相似度。对于向量来说，最大的相似度就是两个向量是平行向量，所以整体来说，两个向量的夹角越小，它就越相似。所以直接使用两个向量的积除以它们的模长的积来计算它们夹角的余弦值，并以这个余弦值作为两个句向量之间的相似度。得分函数值也取这个余弦值。

2.2　TF-IDF 模型

gensim 库用于从原始的非结构化的文本中，无监督地学习到文本隐层的向量表达。gensim 库有语料、向量和模型三大概念。语料是 gensim 库输入一组语句的集合，即为问答对序列。由于汉语自身语言特点，中文句子是由连续的词和词组组成，中间没有空格，为了使机器理解，需要对中文句子进行切分处理，因此在收集语料之后，把中文词语切分成词组，本文使用 jieba 库实现此功能。如果要对语句的隐含结构进行推断，就需要使用适当的数学模型：本文在使用的方法是 doc2bow，也就是将语句转化为词袋。在词袋向量中，每个句子被表示成一个向量，代表字典中

每个词出现的次数。例如，给定一个包含[‘2020’,‘冬奥会’,‘举办地’, ‘金牌’] 的字典，语句[‘2020’, ‘冬奥会’, ‘2020’] 字可以表示成向量[(0,2),(1,1)], 表示‘2020’(编号 0) 出现了 2 次、‘冬奥会’(编号 1) 出现了 1 次。在向量化语料后，可以使用各种模型进行训练，使用模型实质上在两个向量空间中进行转换。

TF-IDF 模型中需要计算词频 (TF) 和逆向文档频率 (IDF), 计算方式分别为

$$\mathrm{TF}=\frac{\text{某词在语句中出现的次数}}{\text{语句的总词量}}$$

$$\mathrm{IDF}=\ln\left(\frac{\text{语料库中语句总数}}{\text{包含该词的语句数}}+1\right)$$

在得到 TF 与 IDF 之后将两个值相乘就得到了一个词的 TF-IDF 值。某个词对测试语句的重要性越高，TF-IDF 值就越大。TF-IDF 模型能有效避免常用词对关键词的影响，提高了关键词与测试语句之间的相关性。

文章使用余弦相似度进行 TF-IDF 值的比较。对于训练集中的每一个问题，可以得到一个 TF-IDF 向量，表示为

$$[(\text{词语1的id},\text{词语1的TF}-\text{IDF值}),(\text{词语2的id},\text{词语2的TF}-\text{IDF值}),\cdots]$$

通过计算两个向量的夹角余弦值来评估他们的相似度。余弦值的范围为 [−1,1], 值越趋近于 1, 代表两个向量的方向越接近，也就是相似度越高。之后利用相似度值来检索最适合的答案。

2.3　知识图谱

知识图谱 (Knowledge graph, KG) 是一种有向图。图中的点代表实体，边代表关系。一个边 (通常为谓词) 连接两个实体，分别为头部实体和尾部实体。这样一个头部实体、一个关系、一个尾部实体构成了一个三元组，也被称为事实。知识图谱通常由数千万乃至数十亿个事实构成。为了在如此庞大的数据中快速且准确地获取目标信息，提出了知识图谱问答。它的目标是把自然语言转换为结构化查询，且返回知识图谱中的实体或谓词为答案。

本文的基于知识图谱嵌入的问答系统主要面向于简单问题。对于简单问题的定义是：如果一个自然语言问题只涉及知识图谱中的一个头部实体和一个谓词，并以它们的尾部实体作为答案，那么这个问题称为简单问题。对于一个简单问题，首先需要预测他的头部实体和谓词，即关系，再在知识图谱中找到对应的尾部实体，将之做为答案返回。

知识图谱的嵌入：用 (h, l, t) 代表一个三元组，用 eh 代表预测头实体表示，类似的，用 et 代表尾实体表示。用 Pl 代表一个预测谓词表示。生成知识图谱的时候，用 (h, l, t) 构成高维知识图谱，为了便于查询，需要把这个高维知识图谱转换为形如 (eh, Pl, et) 的低维向量。这样就把高维的知识图谱嵌入到低维空间，而不会损失信息。

预测关系　首先把句子进行分词，得到标志 (token) 串，用 LSTM 捕捉词与词之间的关系，并转化为向量表达。对每个向量加权后，形成带权向量。最后把一个标志串的向量相加，形成预测关系表达。

预测实体　对于预测实体，需要提前标出哪个位置的词是实体。然后类似于预测关系，先生成标志串，输入 LSTM，形成向量表达，输入全连接层，最终输出每个标志的向量。再根据提前的标注区分是否为实体。

嵌入空间的联合搜索　对于每一个问题，现已经预测了它的谓词表示和头部实体表示。如果一个预测头部实体表示在知识图谱的三元组中，称之为候选头部实体；这个三元组称之为候选事实。度量这个候选事实的谓词与谓词表示的距离，就可以搜素到尾部实体，并将尾部实体作为答案返回。

3　数据集、评估方法与实验结果

3.1　实验数据

为了得到问答数据集，我们上网爬取了 1924 年以来的冬奥会赛事记录 48 563 条，数据中包含运动员的姓名、性别、年龄、身高、体重、国家、参与奥运会的年份、参加奥运会的项目、获奖情况等信息，然后通过脚本生成了对应的“提问−回答对”，具体包括：

1) 针对每一条冬奥会记录，生成关于人物、获奖时间、地点、获奖项目、获奖届别、所属参赛队、奖牌情况的单项问答句。冬奥会纪录共 48 563 条，生成单项信息问答句共 265 757 对。

2) 关于运动员个人情况的所有单项信息问句，如谁是男是女？身高多少？体重多少？是哪个国家的？获奖牌那年多大年龄？参加过哪届运动会？获得过什么奖牌？这些问题还更换了不同的问法，比如询问运动员的国籍有“运动员是哪个国家的？”和“运动员来自哪里？”两种不同的问法。

3) 生成 2 000 对复合统计问答句，如问某国家获得金牌总数、银牌总数、铜牌总数是多少？问某人获得金牌总数、银牌总数、铜牌总数是多少？问某一地区获得金牌总数、银牌总数、铜牌总数是多少？问冬奥会举行过多少届?总共产生多少金牌、多少银牌、多少铜牌？哪个国家获得奖牌数最多，哪个国家获得金牌数最多？哪位运动员个人奖牌数最多？哪位运动员个人金牌数最多？问某个国家在某一个项目上获奖成绩是不断

上升还是不断下降，还是有升有降？

得到的问答对按照 9∶1 的比例划分训练集和测试集。

3.2　实验方法

考虑到 BERT 模型和 TDIDF 模型实现的时候是在所有可能的句子中寻找“最像”的句子并且输出其对应的答句，所以生成的答句应全包含在训练集中，不会产生语义相同但表达方式不同的答句。因此，只需要将答句输出与正确答句直接进行对比，如果相同则可接受。为了进行评估，在模型实现后添加对答句的判断，输出可接受率。在此理论基础上，本文实现了测试 BERT 模型和 TF-IDF 模型实验结果的脚本。

对于知识图谱模型，基于标注好的问答对来构建知识图谱，用其中一部分作为 test 和 valid 集用于训练。最后用所有问答对来检测知识图谱模型的正确率。

3.3　实验结果

实验的最后结果统计如表 1 所示，这里分别列出每一类问答对的可接受率。

表 1　3 种模型对于 3 类问题的回答可接受率

Table 1　Accuracy of three models for three types of questions　%

对话数据集	平均可接受率		
	冬奥会单项记录	运动员个人情况	复合统计问答对
TF-IDF	96.1	99.2	91.5
BERT	98.9	98.4	96.8
知识图谱	99.7	98.9	95.5

上述实验结果表明 BERT 模型的整体效果略优于知识图谱和 TD-IDF 模型，BERT 模型对于 3 类问题的回答可接受率都超过了 96%，知识图谱和 TD-IDF 模型对于复合统计问答对的回答效果不如 BERT 模型。

4　分析与对比

BERT 模型包含两个预训练任务：遮盖部分词的语言模型 (masked language modeling) 和下一个句子预测 (next sentence prediction)。Masked language modeling 预训练过程可以看作是完形填空过程，这个过程使得模型在预测一个词汇时，模型并不知道输入对应位置的词汇是不是正确的词汇，所以模型会更多地根据上下文的信息来预测词汇，并且具有一定的纠错能力；Next sentence prediction 与训练过程可以看作段落重排过程，这使得模型能够更准确地刻画语句的语义信息。BERT 模型的这两个预训练过程使得模型对于问答对特征向量的提取能够做得很优秀，提取出的向量都尽可能全面准确地刻画了输入文本的整体信息。例如，对于问句“Christine Jacoba Aaftink 的身高是多少？”和问句“Christine Jacoba Aaftink 有多高？”，这两个问句虽然提问方式不同，但是使用 BERT 模型提取出的特征向量差别很小，也就是说 BERT 模型对于提问方式的转换能够处理得不错。再比如对于问句“Christine Jacoba Aaftink 参加过哪一年的冬奥会？”，BERT 模型会根据问题中的关键词：“Christine Jacoba Aaftink”“参加”“年”“冬奥会”快速提取出正确的 3 个答案：1988 冬奥会，1992 冬奥会和 1994 冬奥会。由于关键词和特征向量的提取具有代表性，因此对于 3 类问答对 BERT 模型的实现可接受率都整体较高。但是 BERT 模型在训练中的 mask 标记也可能会影响模型表现，因为这个标记在实际预测中不会出现，所以对于个别问题 BERT 模型回答的答案可能会有较大的失误，甚至出现问答不符的现象。另外，BERT 模型结构复杂，复现开销较大。

TF-IDF 模型是建立在区别文档有意义的词语出现频率一定高的假设上的，但显然此理论并不是完全正确的，IDF 的简单结构并不一定能有效地反映单词的重要程度以及特征词的分布情况。因此 TF-IDF 模型的精度比较受限，在实验中基于 TF-IDF 的问答系统可接受率在很大程度上受限于输入数据的影响。在输入数据集有相似问答句时，随机算法生成的测试集与训练集可能有很大部分是重叠的，此时进行测试时 TF-IDF 模型往往能检索到训练集中类似的问句并进行正确输出，可接受率最高可以达到 99.8%。而在没有重复问答句的数据集中进行测试，训练集与测试集往往没有共性，此时可接受率较低，是不可接受的。以实际数据集为例：在生成关于冬奥会记录的问答对时，对于同一语义的问题生成了许多替换类似表达法的问句，比如对于“Christine Jacoba Aaftink 的身高是多少厘米？”这个问句进行表达法上的替换，生成了“Christine Jacoba Aaftink 的身高是多少？”“Christine Jacoba Aaftink 有多高？”等许多同义问句；而这些同义问句都对应着同样的答案，使用这种问答对数据进行测试就会得到较好的效果。反之，在对于“Christine Jacoba Aaftink 参加过哪一年的冬奥会？”这个问题上我们并没有进行替换表达法的完善，在这种每个问句语义都独立的数据集上进行测试，效果就较差。当用于训练的数据趋于完备时，生成的模型才是可接受的。

相比于上面两种模型而言，三元组知识图谱更加贴近实际语言的理解，回答方式有更多的调整空间。比如对于问句“Christine Jacoba Aaftink 参加过哪一年的冬奥会？”，在知识图谱模型中，搜索时会分 3 次搜索到 3 个实体：1988 冬奥会，

1992 冬奥会, 1994 冬奥会。我们可以在输出答案的时候处理为: 若识别到关系是 Athlete/参加/Game, 则返回的答案形如: “Christine Jacoba Aaftink 参加过:”+“1988 冬奥会”+“1992 冬奥会”+“1994 冬奥会”这样的结果, 更加贴近于理解而不是机械返回已有答案。另外, 三元组知识图谱模型对于同一答案的不同提问方式能够很方便地复用, 例如问句“Christine Jacoba Aaftink 的身高是多少？”和“Christine Jacoba Aaftink 有多高？”, 这两个例子在询问身高。在三元组知识图谱模型中, 仅需把这两种问法对应一个 relation, 就可以复用同一个三元组, 节省空间, 且能降低不同提问方式带来的噪声。但是目前来看, 三元组知识图谱仍然不能摆脱人工标注数据。每个问答对的实体, 关系需要人工标注。在预训练时, 想复用同一关系, 需要人工把多个关系合并为一个 ID。与之相比, TF-IDF 模型只需要问答对而不需要标注, BERT 模型更是想把每一篇文章, 作为天生的训练语料, 彻底摆脱人工标注。而且, 三元组知识图谱面对多实体, 多关系的问答对和推理类的问答对回答能力弱。如提问形如“xxx 的身高和体重”这样的问题, 往往只能搜索到身高或体重。提问形如“xxx 在某一届冬奥会上是否获得奖牌”。尽管知识图谱中保存了运动员获得奖品的信息, 但很难通过这些信息推理出答案。

5　结束语

本研究旨在形成冬奥会信息问答系统。我们在网上爬取、收集冬奥会相关信息, 并处理成冬奥问答数据集。之后基于知识图谱、TF-IDF 和 BERT 模型分别训练得到 3 种问答系统。本文在设计对话系统的基础上, 针对系统回答的可接受度进行实验, 将自 1924 年以来的冬奥会数据转化为问答对形式, 在模型上进行训练, 验证了这些回答的可接受率, 对比了 3 种方法的特点以及适用场景。总体来讲, BERT 模型的整体效果略优于知识图谱和 TDIDF 模型, BERT 模型对于 3 类问题的回答可接受率都超过了 96%, 知识图谱和 TD-IDF 模型对于复合统计问答对的回答效果不如 BERT 模型。

参考文献:

[1] 鞠晓峰, 都军, 覃军, 等. 人工智能在智能问答系统中的应用 [J]. 智能建筑与智慧城市, 2021(3): 36–37.
JU Xiaofeng, DU Jun, QIN Jun, et al. Application of artificial intelligence in intelligent question answering system[J]. Smart building and smart city, 2021(3): 36–37.

[2] 王银丽. 限定领域内智能问答系统的研究与实现 [D]. 内蒙古: 内蒙古大学, 2008.
WANG Yinli. Research and implementation of intelligent question answering system in limited domain[D]. Inner Mongolia: Inner Mongolia University, 2008.

[3] 张宁, 朱礼军. 中文问答系统问句分析研究综述 [J]. 情报工程, 2016, 2(1): 32–42.
ZHANG Ning, ZHU Lijun. A survey of the research on question and answer system in Chinese[J]. Technology intelligence engineering, 2016, 2(1): 32–42.

[4] MISHRA A, JAIN S K. A survey on question answering systems with classification[J]. Journal of king saud university-computer and information sciences, 2016, 28(3): 345–361.

[5] 姚冬, 李舟军, 陈舒玮, 等. 面向任务的基于深度学习的多轮对话系统与技术 [J]. 计算机科学, 2021, 48(5): 232–238.
YAO Dong, LI Danjun, CHEN Shuwei, et al. Task oriented multi round dialogue system and technology based on deep learning[J]. Computer science, 2021, 48(5): 232–238.

[6] FENG Minwei, XIANG Bing, GLASS M R, et. al. Applying deep learning to answer selection: a study and an open task[C]// 2015 IEEE Workshop on Automatic Speech Recognition and Understanding. IEEE, Piscataway, 2015: 813−820.

[7] 张涛, 贾真, 李天瑞. 等. 基于知识库的开放领域问答系统 [J]. 智能系统学报, 2018, 13(4): 557–563.
ZHANG Tao, JIA Zhen, LI Tianrui, et al. Open-domain question-answering system based on large-knowledge base[J]. CAAI transactions on intelligent systems, 2018, 13(4): 557–563.

[8] NORASET T, LOWPHANSIRIKUL L, TUAROB S. Wabiqa: A wikipedia-based thai question-answering system[J]. Information processing & management, 2021, 58(1): 102431.

[9] HÖFFNER K, WALTER S, MARX E, et al. Survey on challenges of question answering in the semantic web[J]. Semantic web, 2017, 8(6): 895–920.

[10] 李涛, 王次臣, 李华康. 知识图谱的发展与构建 [J]. 南京理工大学学报(自然科学版), 2017, 41(1): 22–34.
LI Tao, WANG Cichen, LI Huakang. Development and construction of knowledge map[J]. Journal of Nanjing University of Science and Technology (natural science edition), 2017, 41(1): 22–34.

[11] 刘峤, 李杨, 段宏, 等. 知识图谱构建技术综述 [J]. 计算机研究与发展, 2016, 53(3): 582–600.
LIU Qiao, LI Yang, DUAN Hong, et al. Overview of knowledge mapping technology[J]. Journal of computer research and developmen, 2016, 53(3): 582–600.

[12] 徐梦婷. 基于知识图谱的多轮问答系统 [D]. 南京: 南京邮电大学, 2020.
XU Mengting. Multi round question answering system based on knowledge map[D]. Nanjing: Nanjing University of Posts and Telecommunications, 2020.

[13] 陈勇. 基于知识图谱的智能系统在电力行业的应用 [D]. 南京: 南京师范大学, 2020.
CHEN Yong. Application of intelligent system based on knowledge map in power industry[D]. Nanjing: Nanjing Normal University, 2020.

[14] PRZYBYŁA P. Boosting question answering by deep entity recognition[J]. arXiv preprint arXiv: 1605.08675, 2016.

[15] YIH Wentau, CHANG Mingwei. Semantic parsing via staged query graph generation: question answering with knowledge base[C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015: 1321−1331.
[16] 贾中浩, 宾辰忠, 古天龙, 等. 基于知识图谱和用户长短期偏好的个性化景点推荐 [J]. 智能系统学报, 2020, 15(5): 990–997.
JIA Zhonghao, BIN Chenzhong, GU Tianlong, et al. Personalized attraction recommendation based on the knowledge graph and users' long-term and short-term preferences[J]. CAAI transactions on intelligent systems, 2020, 15(5): 990–997.
[17] 陆亚辉. 面向服务机器人的口语对话系统研究与实现 [D]. 哈尔滨: 哈尔滨工业大学, 2017.
LU Yahui. Research and implementation of oral dialogue system for service robot[D]. Harbin: Harbin Institute of Technology, 2017.
[18] 张辛. 基于 TFIDF 算法的全面从严治党重要论述关键词共现分析 [J]. 现代盐化工, 2019(7): 150–152.
ZHANG Xin. Key words co-occurrence analysis of comprehensive and strict party governance based on TFIDF algorithm[J]. Modern salt chemical industry, 2019(7): 150–152.
[19] 苏林萍, 林小倩, 陈飞, 等. 基于 N-Gram 和 TFIDF 的 SQL 注入检测方法 [J]. 计算机与数字工程, 2021(6): 1177–1181.
SU Linping, LIN Xiaoqian, CHEN Fei, et al. SQL injection detection method based on N-gram and TFIDF[J]. Computer and digital engineerin, 2021(6): 1177–1181.
[20] 刘娟, 郝云强. 尹雪雪 网络舆情信息挖掘关键技术分析 [J]. 信息科技, 2021(3): 94–95.
LIU Juan, HAO Yunqiang, YIN Xuexue. Analysis on key technologies of network public opinion information mining[J]. Information technology, 2021(3): 94–95.
[21] 吴思慧, 陈世平. 结合 TFIDF 的 Self-Attention-Based Bi-LSTM 的垃圾短信识别 [J]. 计算机系统应用, 2020, 29(9): 171–177.
WU Sihui, CHEN Shiping. Spam message recognition based on self attention based Bi LSTM combined with TFIDF[J]. Computer systems & applications, 2020, 29(9): 171–177.
[22] 李海林, 邹金串. 基于分类词典的文本相似性度量方法 [J]. 智能系统学报, 2017, 12(4): 556–562.
LI Hailin, ZOU Jinchuan. Text similarity measure method based on classified dictionary[J]. CAAI transactions on intelligent systems, 2017, 12(4): 556–562.
[23] 曹旭友, 周志平, 王利, 等. 基于 BERT+ATT 和 DBSCAN 的长三角专利匹配算法 [J]. 信息技术, 2020, 44(3): 1–5, 12.
CAO Xuyou, ZHOU Zhiping, WANG Zhao, et al. Patent matching algorithm in Yangtze River Delta Based on Bert + ATT and DBSCAN[J]. Information technology, 2020, 44(3): 1–5, 12.
[24] 吴炎, 王儒敬. 基于 BERT 的语义匹配算法在问答系统中的应用 [J]. 仪表技术, 2020(6): 19–22, 30.
WU Yan, WANG Rujing. Application of semantic matching algorithm based on Bert in question answering system[J]. Instrumentation technology, 2020(6): 19–22, 30.
[25] 朱鹤, 陆小锋, 薛雷. 基于 BERT 的金融文本情感分析模型 [J], 上海大学学报:自然科学版. https://kns.cnki.net/kcms/detail/31.1718.n.20210616.1757.002.html.
ZHU He, LU Xiaofeng, XUE Lei. Financial text sentiment analysis model based on BERT[J]. Journal of Shanghai University (natural science edition). https://kns.cnki.net/kcms/detail/31.1718.n.20210616.1757.002.html.
[26] 孙士琦, 汤鲲. 基于 BERT 的中文地址分词方法 [J]. 信息科技, 2021(9): 155–159.
SUN Shiqi, TANG Kun. Chinese address segmentation method based on Bert[J]. Information technology, 2021(9): 155–159.
[27] 彭宇, 李晓瑜, 胡世杰, 等. 基于 BERT 的三阶段式问答模型 [J]. 计算机应用, 2021(8): 1–8.
PENG Yu, LI Xiaoyu, HU Shijie, et al. Three stage question answering model based on Bert[J]. Journal of computer applications, 2021(8): 1–8.
[28] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Advances in neural information processing systems. 2017: 5998-6008.

作者简介:

罗玲, 女, 硕士, 主要研究方向为自然语言处理与强化学习。

李硕凯, 博士研究生, 主要研究方向为数据挖掘、推荐系统与元学习。

何清, 研究员, 博士生导师, 中国人工智能学会副秘书长、常务理事、知识工程与分布智能专业委员会秘书长、机器学习专业委员会常务委员, 中国计算机学会高级会员、人工智能与模式识别专业委员会委员, 中国电子学会云计算专家委员会委员. 主要研究方向为机器学习、数据挖掘、文本挖掘、基于云计算的分布式并行数据挖掘。主持和参与国家“863”和“973”计划、国家自然科学基金等科研项目多项, 2008 年底, 何清研究员带领他的中科院计算所数据挖掘团队, 受中国移动研究院委托, 合作开发完成了基于云计算的并行数据挖掘平台, 用于 TB 级实际数据的挖掘, 实现了高性能、低成本的数据挖掘。发表学术论文近百篇。

[责任编辑: 李雪莲]