计算机科学
COMPUTER SCIENCE

一种基于知识图谱的检索增强生成情报问答技术

成志宇, 陈星霖, 王菁, 周中元, 张志政

引用本文

成志宇, 陈星霖, 王菁, 周中元, 张志政. 一种基于知识图谱的检索增强生成情报问答技术[J]. 计算机科学, 2025, 52(1): 87-93.

CHENG Zhiyu, CHEN Xinglin, WANG Jing, ZHOU Zhongyuan, ZHANG Zhizheng. Retrieval-augmented Generative Intelligence Question Answering Technology Based on Knowledge Graph [J]. Computer Science, 2025, 52(1): 87-93.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.240900064

一种基于知识图谱的检索增强生成情报问答技术

成志宇[1]　陈星霖[2]　王　菁[3]　周中元[4]　张志政[5,6]
1 东南大学苏州联合研究生院　江苏 苏州 215000
2 东南大学软件学院　南京 211189
3 信息系统工程全国重点实验室　南京 210023
4 中国电子科技集团公司第二十八研究所　南京 210023
5 东南大学计算机科学与工程学院　南京 211189
6 新一代人工智能技术与交叉应用教育部重点实验室(东南大学)　南京 211189
(1103357821@qq.com)

摘　要　为实现军事情报问答,提出了一种基于知识图谱的检索增强生成框架。该框架通过问题分类、实体识别、实体链接、知识检索有效地获取了背景知识。同时考虑到情报问题多约束的特点,使用回答集编程在知识上通过约束限制减少知识数量或者直接获得答案。最后,使用大语言模型在精炼后的知识上对问题进行求解,以减少问题理解过程中的属性识别与链接。在MilRE数据集上的实验表明,所提框架能够提供基于知识图谱的增强知识检索功能,并具有较好的军事情报问题解答能力。

关键词:情报问答;回答集编程;大语言模型;检索增强生成;知识图谱
中图分类号　TP391

Retrieval-augmented Generative Intelligence Question Answering Technology Based on Knowledge Graph

CHENG Zhiyu[1],CHEN Xinglin[2],WANG Jing[3],ZHOU Zhongyuan[4] and ZHANG Zhizheng[5,6]
1 Joint Graduate School,Southeast University,Suzhou,Jiangsu 215000,China
2 College of Software Engineering,Southeast University,Nanjing 211189,China
3 Science and Technology on Information Systems Engineering Laboratory,Nanjing 210023,China
4 The 28th Research Institute of China Electronics Technology Group Corporation,Nanjing 210023,China
5 School of Computer Science and Engineering,Southeast University,Nanjing 211189,China
6 Key Laboratory of New Generation Artificial Intelligence Technology and Its Interdisciplinary Applications(Southeast University),Ministry of Education,Nanjing 211189,China

Abstract　A knowledge graph-based retrieval-augmented generation framework is proposed to achieve military intelligence question answering. The framework effectively acquires background knowledge through question classification,entity recognition,entity linking,and knowledge retrieval. Considering the multi-constraint characteristics of intelligence questions,answer set programming is used to reduce the amount of knowledge through constraints or to directly obtain answers. Finally,a large language model solves the questions based on the refined knowledge,minimizing attribute recognition and linking issues during question understanding. Experiments on the MilRE dataset demonstrate that the framework provides enhanced knowledge retrieval capabilities based on knowledge graphs and offers superior performance in answering military intelligence questions.

Keywords　Intelligence question-answering,Answer set programming,Large language models,Retrieval-augmented generation,Knowledge graph

情报是为指挥者提供决策支持服务的重要信息源。情报　　问答指基于特定的情报需求,从结构化和非结构化的数据源

到稿日期:2024-09-10　返修日期:2024-10-20
基金项目:军科委国防科技重点实验室基金(6142101210205);军科委国防科技创新特区资助项目
This work was supported by the Pre-research Key Laboratory Fund for Equipment(6142101210205) and National Defense Science and Technology Innovation Special Zone Funding Project.
通信作者:张志政(seu_zzz@seu.edu.cn)

中检索、解析并生成相关答案的过程。随着人工智能技术的发展，结合大语言模型与知识图谱的问答技术已经成为问答系统的一大发展方向[1]。

1 问题的提出

军事情报问答领域需要基于准确事实的回答。知识图谱通过结构化方式组织信息，以知识图谱为数据源的军事情报问答方法可以高效获取相关事实信息。现有基于知识图谱的军事情报问答方法大多基于深度学习的方法来实现[2-5]，基于大语言模型实现的军事情报问答方法较为罕见。基于深度学习的知识图谱问答方法主要分为语义解析方法和信息检索方法[6]，在应对复杂的、多重约束条件或需深度推理的查询时，这些方法常面临困难，并且在数据稀疏的情形下，难以准确捕捉用户意图并理解语境。大语言模型[7]通过海量数据的训练，具备处理复杂问题和模糊查询的能力，涵盖了广泛的主题与知识范围。但是，大语言模型有时会生成与现实情况不符的内容[8]，特别是在处理要求高可靠性的军事情报问题时，这种“幻觉”可能导致错误决策和评估，带来潜在风险。同时，在军事情报问答领域，知识会随着时间快速变化，微调大语言模型不仅成本高昂，且通常难以实现[9]。这些问题使得研究者在仅依靠大语言模型处理军事情报问答时面临挑战。

近年来，检索增强生成(Retrieval Augmented Generation，RAG)[10]通过结合知识库中检索到的相关事实并利用大语言模型生成回答，有效缓解了大语言模型生成幻觉[8]的问题，提升了问答的准确性。

近年来主流的 RAG 为基于文档分割的 RAG，其首先将文档分块并构建索引，再依据查询条件检索最相关的内容片段，最后将这些片段整合为上下文提供给大语言模型用于生成答案[11]。例如，WebGPT[12]结合人类反馈，以浏览器辅助问答的方式从网络检索信息；Self-RAG[13]通过‘反思令牌’机制自我检索和评估，提高了生成的连贯性和准确性；而 RAG-Fusion[14]则通过多查询策略，捕获显性和隐性信息以提高回答的相关性。基于文档分割的 RAG 方法虽然在开放域问答的任务中表现出了较高的性能，但它在特定领域应用时的有效性还面临着若干挑战：1)面对特定领域的应用，原本用于开放域场景的检索器往往难以实现高效的检索；2)在某些专业领域的知识往往以结构化的形式存储，如知识图谱或结构化数据库等。通过文档切分实现的 RAG 方法无法有效检索这些知识库，同时在匹配背景知识和覆盖复杂问题所需的全部背景知识方面面临困难。由于本研究的文本情报问答系统采用知识图谱来存储数据，因此需要研究基于知识图谱数据源的 RAG 方法来应对以上挑战。

同时，由于大语言模型如 ChatGLM3 和 Qwen1.5 在处理输入时受限于模型的输入长度上限[15]，使用 RAG 方法时，若上下文过长，不仅可能增加经济成本，还可能影响模型的推理性能[16]，甚至当输入长度超出 Token 限制时，模型可能无法处理查询。多约束问题是军事情报中常见的一类问题，约束条件代表已知信息，解答必须满足这些条件。常见的约束包括实体约束、类型约束、时间约束、地点约束等。例如，问题“空客公司生产的飞机有哪些”中含有类型约束“飞机”和实体约束“空客公司”。再如，“2021 年 4 月 16 日 RC-135 由哪个基地起飞出发前往南海开展侦察任务”中包含军事兵器约束“RC-135”、地点约束“南海”、类型约束“侦察任务”以及时间约束“2021 年 4 月 16 日”等。在查询知识图谱时，只有符合所有这些约束的事件才会被加入候选事件集合。在处理类似问题时，RAG 方法在检索时不可避免地会检索出大量相关或者冗余的知识。为了解决这一问题，需要设计一种新的 RAG 方案，以减少传递给大语言模型的背景信息长度，从而降低经济成本并提高推理性能。

总的来说，在军事情报问答领域，当前的 RAG 方法面临两大挑战。1)现有的基于文档的 RAG 主要在开放域任务上进行训练和评估。然而，军事领域的知识通常以结构化形式存储，如知识图谱和结构化知识库，这使得文档型检索器无法有效检索信息，导致 RAG 在以知识图谱为数据源的知识库中难以获得有效信息，需要探索数据源为军事领域知识图谱的 RAG 方案来解决上述问题。2)在面对军事情报领域常见的多约束问题时，现有 RAG 方法在检索阶段往往引入大量无关知识，导致上下文长度过长，不仅会增加经济成本，还会影响模型的推理性能，因此有必要对输入给大语言模型的知识进行精炼。

本文针对上述问题，提出了一种基于知识图谱的检索增强生成框架。该框架以知识图谱为数据源，通过问题分类、实体识别、实体链接、知识检索有效地获取问题的相关知识。借助回答集编程(Answer Set Programming，ASP)[17]，通过对知识施加约束条件并推理，缩减知识数量或者直接推理出答案。最后，使用大语言模型在精炼后的知识上对问题进行求解，以尽量减小属性识别与链接对问题理解的影响。

2 情报问答算法

2.1 情报问答方案设计

本文设计了一个新的框架，如图 1 所示。该框架的运作过程可以划分为 6 个步骤。1)问题分类：利用在军事问题数据集上训练的文本分类模型对问题进行分类，从而确定预定义的类别。2)实体识别：利用经过军事领域实体识别语料库中训练得到的模型对问题进行处理，提取相关的待选实体和属性，后续的实体链接步骤依赖于此过程所提供的基础。3)实体链接：基于预训练的模型，采用向量相似度等技术，将候选实体和属性与存储在知识图谱中的目标实体和属性进行对齐，以识别与问题相关的实体及其约束条件。4)知识检索：依据实体信息和兵器的类别，提取与问题求解有关的知识，通过 ASP 对事实建模，为后续步骤提供支撑。5)知识精炼：ASP 可以方便地对约束进行表示，其求解器能够通过自动推理，找到所有满足约束的解集。在知识精炼过程中，框架利用 ASP 内嵌的求解器对已经实例化的 ASP 事实和规则进行计算与推理，可以有效地压缩知识量，从而减轻大语言模型在求解问题时的推理负担。此外，ASP 还可以直接给出解答，在某些场景下减少对大语言模型的依赖。6)问题求解：结合

问题内容、提示语和使用 ASP 规则精炼后的知识，利用大语言模型生成相应的查询回答。

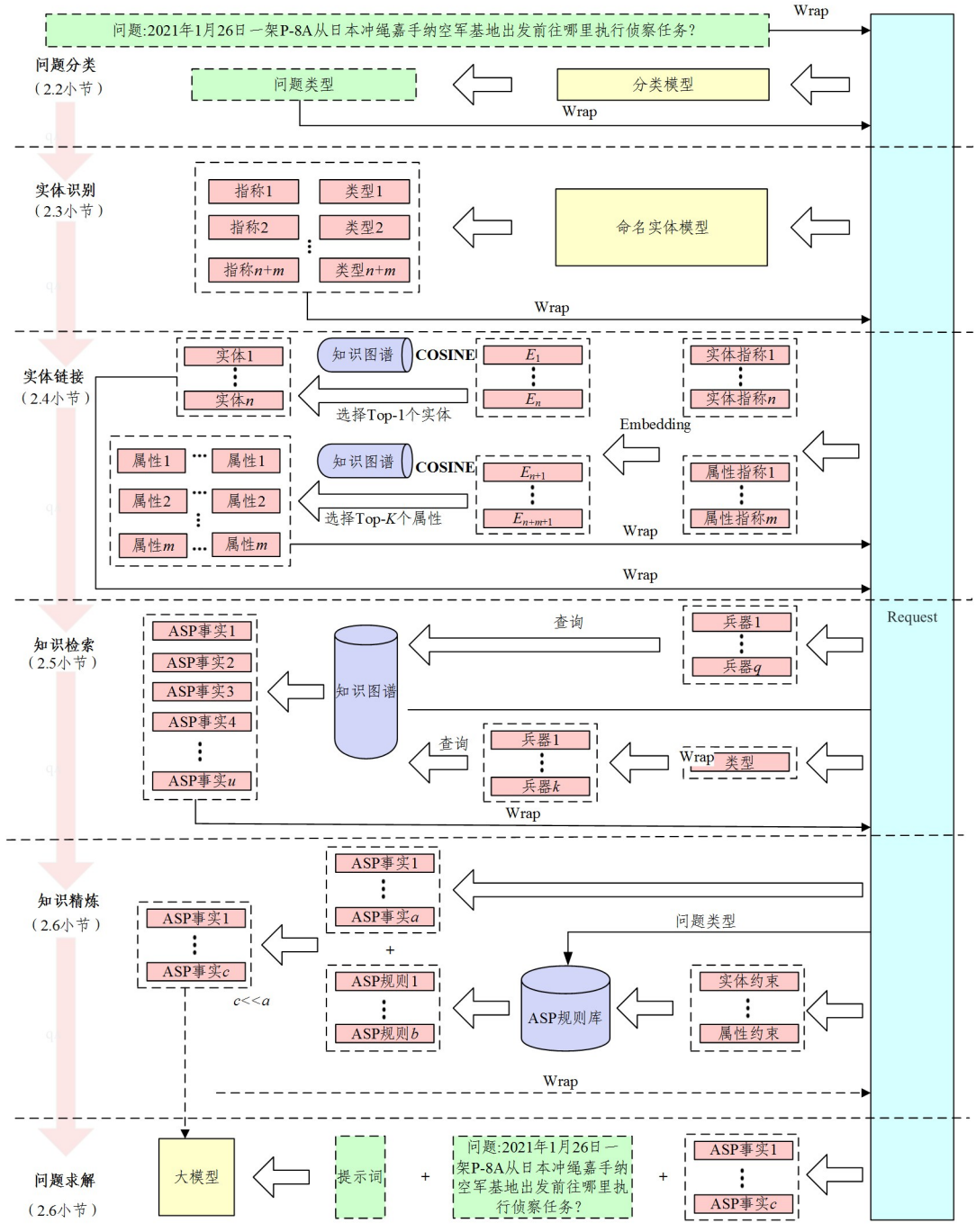

图 1　问答方案整体架构图

Fig. 1　Overall architecture of Q&A solution

2.2　问题分类

问题分类任务需要确定问题的类型。本文采用在数据集上训练好的文本分类模型对问题进行识别，确定其对应的预设类型。后续的知识检索和问题求解方法将根据问题类型的不同设定相应的策略。

在具体实施方面，为了增强模型在军事领域知识理解方面的能力并提升其整体性能，本文使用 K-BERT[18] 作为预训练模型，将结构化的军事知识图谱知识融入模型。K-BERT 通过结合知识图谱和语义上下文，增强了对实体和语义的捕捉能力，使用这些词向量初始化词节点有助于提升模型的分类效果。后文所有的预训练模型均使用 K-BERT，不再赘述。本文使用 K-BERT 模型（BERT[19] 的一种变体）将分词后的文本序列 q 嵌入到高维向量空间 hq，通过多个自注意力层计算每个词与其他词之间的依赖关系，从而捕捉全局的上下文信息，最后使用分类层将编码后的表示 hq 转换为具体的分类结果 cr。具体过程为：

$$hq = K\text{-}BERT(q) \tag{1}$$

2.3　实体识别

实体识别任务的目标是识别出问题中的候选实体。识别结果将作为输入传递给实体链接任务，以进行实体消歧以及

知识精炼任务，从而进行知识过滤。

本文定义的实体识别任务如下，针对每一个军事问句 S，生成的待选实体集为 $E=\{(ce_1,l_1),(ce_2,l_2),\cdots,(ce_n,l_n)\}$。其中，每一个元组的 ce_i 代表候选军事实体，l_i 代表识别出的兵器类型。

在具体实施方面，本节聚焦军事实体识别任务。与通用领域命名实体识别相比，此任务面临以下挑战：1）该领域实体类型丰富且数量庞大；2）军事领域实体往往伴随着复杂的修饰词修饰，这对精确界定实体边界提出了更高的要求；3）某些实体的表示形式中可能包括缩写、大小写混用或中英文混杂等情况，如装备型号等。为应对上述挑战，将实体识别模型分解为如下 3 个部分。

1）使用预训练的深度学习模型获取问题中的语义信息。BERT 等预训练模型由于主要依赖大规模通用语料库，缺乏针对特定领域如军事领域的深入训练，这限制了其在这些领域内对知识的理解。本节采用 K-BERT，该模型能够将领域知识图谱信息注入到文本中，更有效地应对军事领域中的专业词汇与特定概念。具体过程为：对于给定的长度为 N 的输入序列 $S=\{C_0,C_1,C_2,C_3,\cdots,C_{n-1}\}$，通过式(2)所示过程得到 S 的 K-BERT 的隐藏层编码序列 S_v。

$$S_v=K\text{-}BERT(S) \tag{2}$$

2）在 K-BERT 模型后增加双向长短期记忆网络(BiLSTM)，以提高处理长文本的能力。BiLSTM 的引入使得模型能够更好地捕捉词间长距离依赖，优化了文本特征表达，增强了命名实体识别的准确性，从而提升识别的准确度。具体过程为，对于通过 K-BERT 获得的隐藏层编码 S_v，使用 BiLSTM 网络处理后得到特征表示序列 X。

3）条件随机场(CRF)利用从数据中学习到的标签转移模式及约束，确保模型精确地构建合规的实体识别序列，提高标签预测的正确性。本研究在使用 BiLSTM 表示完文本特征后，整合了 CRF 层，旨在精确地识别不同的实体类型。具体过程中，假设输入 CRF 层的序列为 $X=(x_1,\cdots,x_n)$，预测的标签序列 $Y=(y_1,\cdots,y_n)$，Encoder 层得到的分数矩阵为 $\boldsymbol{P}^{n\times n}$，其中 $p_{i,j}$ 表示第 i 个输入序列 x_i 预测为第 j 个标签 y_j 的得分。由此，可以得到整体预测序列 Y 的评分函数为：

$$Score(X,Y)=\sum_{i=0}^{n}M_{y_i,y_{i+1}^*}+\sum_{i=1}^{n}P_{i,y_i} \tag{3}$$

其中，$M_{y_i,y_{i+1}}$ 表示标签 y_i 转移为 y_{i+1} 的得分，其构成了整个标签转移得分矩阵 $\boldsymbol{M}$。

在此基础上，可以计算出 Y 的产生概率：

$$P(Y|X)=\frac{e^{Score(X,Y)}}{\sum\limits_{\widetilde{Y}\in Y_X}Score(X,\widetilde{Y})} \tag{4}$$

其中，$\widetilde{Y}$ 为真实的标注序列，Y_X 为标注序列所有可能的形式。

CRF 通过对上述预测序列 Y 得分函数求最大似然函数的方式来进行模型优化，并最终得到最大得分的序列 Y^*。

$$\ln P(Y|X)=Score(X,Y)-\ln\Big(\sum_{\widetilde{Y}\in Y_X}Score(X,\widetilde{Y})\Big) \tag{5}$$

$$Y^*=\underset{\widetilde{Y}\in Y_X}{\operatorname{argmax}}\,Score(X,\widetilde{Y}) \tag{6}$$

2.4 实体链接

实体链接模块中会计算向量相似度，将候选实体与知识图谱中的对应实体匹配。这一过程生成的已链接实体将用于知识检索模块，以便从知识图谱中获取相关背景知识。

本方法基于 Neo4j 和本地存储的军事领域知识图谱实现。Neo4j 5.18 版本引入了向量索引机制，允许用户在特定列上为各类实体创建索引，通过相似度计算快速查找相似实体，降低了实体链接步骤的复杂性。本方法借助此机制，针对每个节点的属性进行了调整，新增了 Embedding 字段。该字段由预训练模型的编码器生成填充。同时，我们为不同类别的实体创建了向量索引，用于提高检索效果。

本节选用了命名实体识别任务中预训练的 K-BERT-BiLSTM-CRF 模型中的 K-BERT 编码层作为预训练模型，具体原因已在 2.3 节中说明。

2.5 知识检索

图 2 展示了本文所定义的检索任务流程。针对每个问题 S，相关实体的集合表示为 $E=\{(e_1,t_1),(e_2,t_2),\cdots,(e_n,t_n)\}$，其中 e_i 代表具体的军事实体，t_i 表示该实体所属的武器类型。同时，若待查询问题中包含兵器类型，本方法将从知识库中检索相关实体，以此补充实体集合 E，这时 $E=\{(e_1,t_1),\cdots,(e_n,t_n),\cdots,(e_{n+m},t_{n+m})\}$。随后，本文根据问题的类型制定了具体的检索策略，针对每个实体元组，利用该策略生成相应的知识集合 $K=\{k_1,k_2,\cdots,k_{n+m}\}$，其中 k_i 表示第 i 个实体的知识集合，是检索结果的一部分。为了简化系统结构，提高维护性和复用性，本研究使用 Cypher 语句设计了一些基本操作用于实现对知识的有效查询。

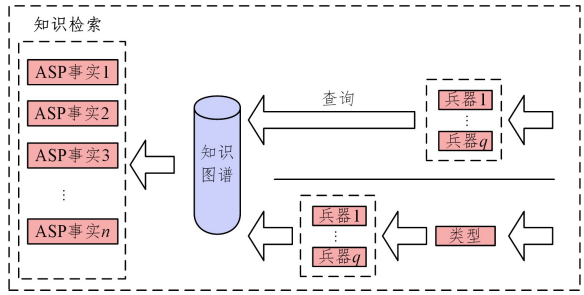

图 2 知识检索过程图

Fig. 2 Diagram of knowledge retrieval process

本文实验中采用的 MilRE 数据集中，事件通过关系表示，从而将多跳事件查询简化为单跳查询。因此，有必要将关系区分为行动关系和非行动关系。行动关系专指实体在特定地点和时间发生的事件，例如某架次 P-3C 在特定时间和地点执行的侦察任务；非行动关系则表示与行动无关的其他关联关系，像 F-22 与其制造商 L 公司之间的关系。在本研究所采用的军事知识图谱中，区分实体间关系类型依赖于连接的导向。具体来说，非行动关系表现为从当前实体节点指向其他实体节点，相对地，行动关系则由其他实体节点指向当前实体节点。

本文在实体链接环节对候选实体的非行动关系以及属性做了精简处理，目的是减轻后续流程中的大模型的计算推理

负担。在问句 S 中，其第 i 个被识别的命名实体关联到的相关知识集合表示为 $k_i=\{P_{i1},P_{i2},P_{i3},\cdots,P_{im}\}$，$k_i\in K$，$P_{ij}$ 代表第 j 个属性，其属于第 i 个实体。通过计算 S 和 P_{ij} 之间的向量余弦相似度获得属性得分，用于评估属性的相关性，随后排序，最后选出得分前 K 名的属性用于后续流程。这样不仅提升了推理效率，还显著缩短了处理时间。

2.6 基于回答集编程知识精炼和大语言模型问题的求解方法

尽管检索到了与问题相关的知识，但知识的范围过广，需要对其进行精简以缩短上下文长度，加快大语言模型的推理速度。因此，本节提出了一种将回答集编程与大语言模型相结合的方案，用于解决军事问题。其具体流程如图 3 所示。

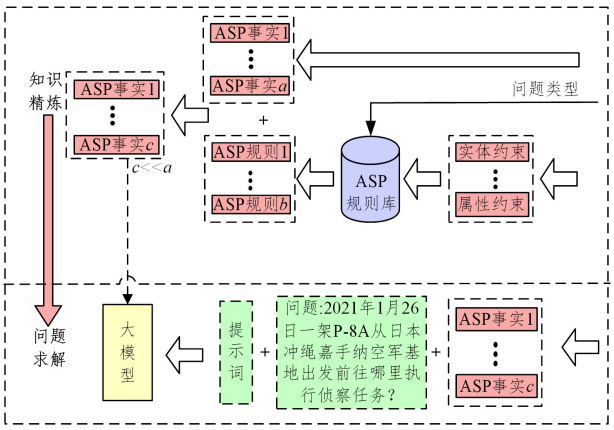

图 3 知识精炼与问题求解过程图

Fig. 3 Diagram of knowledge refinement and problem-solving process

在军事情报问答领域，识别和链接属性时常面对挑战。由于数据中涉及属性的部分常存在指称具有多义性、含义不明确、指称模糊、上下文不清等多重问题，因此确定并链接属性显得尤为困难。通常，属性的识别受到数据集完整性与质量的影响，而一些属性相关的训练数据可能较少，这让依靠数据驱动的深度学习模型在区分这些属性时面临挑战。比如，面对“哪种运输机飞行距离最远”这一问题，属性识别可能只能捕捉到“最远”，而无法链接到更具体的“最大航程”属性。因此在这一情况下，减少对属性链接的依赖和频率变得尤为关键。为应对上述挑战，我们引入了大语言模型并结合相关知识和提示进行问题求解，利用大模型的强大常识储备和自然语言理解能力，在减少对属性识别的依赖的同时，提升问题求解的效率。然而，大语言模型在处理大量上下文时，性能会受到输入长度的限制，同时带来更高的计算成本。因此，为了提高模型效率，我们通过设计 ASP 规则对相关知识进行精炼处理，减少模型所需处理的信息量，从而在保持问题求解准确性的同时提升效率。

ASP 是一种声明式逻辑编程方法，以出色的建模能力著称，尤其在表达复杂约束条件时表现优异。在处理军事领域中的多约束问题时，ASP 具备强大的描述和求解能力。ASP 内嵌的搜索和推理功能可以在给定约束条件下搜索所有可能的解集，非常适用于处理多重约束问题。同时，ASP 能够在满足硬性约束的前提下找到符合软约束要求的最优解，这一特点能够在多重约束问题求解过程中减少文本理解时产生的误差累积。接下来将详细介绍每种问题需要的 ASP 规则以及对应的求解方法。

在处理数量统计类问题时，即计算符合指定约束的实体数量时，为提升求解过程的可解释性，本节采用了 ASP 规则(式(7))来定位满足指定约束的所有实体并列出名称，而非直接计算数量。这种设计使得求解过程更加透明，便于理解和验证。

$$constraint(Entity):\text{-}entity(Entity,\_,\text{“}\{entity\_constraint\}\text{”}) \tag{7}$$

针对最大值类问题，鉴于其多样的表达方式，本文使用了 ASP 规则(式(8))进行处理。我们并不急于判断具体是哪种属性，而是计算所有相关属性的最大值并以 ASP 三元组的形式表示，最后将三元组与问题一并输入到大语言模型中以生成答案。通过这种设计实现了延迟属性链接。求最小值的步骤与式(8)所描述的类似，在此不做叙述。

$$\begin{aligned}&max\_constraint(Entity,\text{“}\{property\_constraint\}\text{”},MaxValue):\text{-}entity(Entity,\{property\_constraint\},Value),\\&Value=\#max\{\{W:entity(\_,\text{“}\{property\_constraint\}\text{”},W)\}\},MaxValue=Value\end{aligned} \tag{8}$$

针对枚举类问题，即符合特定约束条件的实体类型，本文直接应用式(7)来快速获取答案。

在处理含有多重约束的交集问题时，即需要同时满足两个约束条件的实体，本研究在式(7)的基础上增设了一个相同的约束条件，以满足双重条件要求，其结果由 ASP 直接计算得出。

差集类问题中，实体需要符合第一个约束并排除第二个。ASP 规则如式(9)所示，其中使用了 not 关键字来排除不符合第二个约束的实体。

$$\begin{aligned}&constraint(Entity):\text{-}entity(Entity,\_,\text{“}\{entity\_constraints[0]\}\text{”}),notentity(Entity,\_,\text{“}\{entity\_constraints[1]\}\text{”})\end{aligned} \tag{9}$$

针对单实体问题和有限实体限制下的多实体问题，包括选择、判断、比较以及基础四则运算问题等，其通常涉及有限属性/实体的检索和分析。通过对相关知识的 ASP 精炼，大模型可以对这些问题实现较好的解答效果。本文采用表 1 所列的提示词模板生成答案，其中{context_str}和{query_str}为动态填充项，分别用于填充背景知识和具体问题。特别地，在处理计算类问题时发现，大模型可能倾向于根据文本模式生成预测结果，导致直接的计算往往不准确。为此，我们通过在问题前添加“请编写程序进行计算”的提示，来引导大语言模型生成 Python 代码进行计算，以提高计算结果的准确性。

表 1 简单问题查询模板

Table 1 Query template of simple question

提示词模板
你需要基于我提供的上下文信息而非先验知识，回答以下问题，并在 explanation 中明确说明答案的知识来源。如果上下文中未包含相关知识，请在回答中指出，而不是凭空编造。 当前提供的上下文信息如下，每个元组代表一条事实：{context_str}，问题为：{query_str}。格式：请以 answer 和 explanation 为 key 的 JSON 形式提交回答，请严格遵循我指定的格式。

针对单事件类问题，本研究利用多种 ASP 规则对背景知识进行过滤与精炼，当对齐过程失败时，通过放宽约束条件来获取潜在的背景知识，以确保精炼后的数据保持较高的准确性。本文定义了如式(10)所示的 ASP 规则，用于表示硬约束。规则中定义了 3 个可填充的字段，这些字段可以是实体、实体类型、日期、起点、终点和任务类型。我们定义了以式(11)形式呈现的软约束，其适用于日期、起点和终点，以上软约束的惩罚值均设定为 1。通过筛选惩罚总值最小的事件，确定最终的输出结果。

$$constraint(EntityType, Entity, TaskType, Date, Departure, Destination, ID)$$
$$EntityType = \{entity\_type\_constraint\},$$
$$TaskType = \{task\_constraint\}. \tag{10}$$

$$date\_soft\_constraint(ID,1) :\text{-} constraint(\_,\_,\_, Date,\_,\_, ID),$$
$$Date! = \{date\_constraint\} \tag{11}$$

针对多事件问题，本文将实体链接分析得到的主体类型、主体、起点和终点作为约束条件进行处理，应用多种 ASP 规则对知识进行精炼。由于多事件问题的约束条件较为宽松，因此会产生大量符合条件的事件。为此，本文将所有约束条件设定为硬约束，以缩小符合条件的事件范围。这些约束均采用通配符作为默认值。在完成知识精炼后，本文将其与问题文本和预设提示词相结合，通过大语言模型生成查询结果。

3 实验

3.1 数据准备

鉴于军事领域的敏感性，公开的相关数据集非常有限，且多数仅涉及实体问答数据。经过调研，本文使用 MilKB[4]，这是军事领域的一个中文开源知识库，除了实体知识以外，还包含了丰富的事件相关知识。该知识库将问题划分为多种类别，共收录了 2829 个问题及其答案，其中包括 600 余个针对军事领域事件的实例。为了扩展对军事相关问题的覆盖范围，并提高分类的准确性，本文依托 MilKB 对军事问题类别进行了重新设计，新增了实体问题大类中的“数量比较”类型和事件问题大类下的“事件判断”类型。同时参考了军事领域专家的意见，补充了相应的问题数据。通过这些调整和补充，本文构建了多样性更强的 MilRE 数据集，该数据集相比 MilKB 能够更全面地覆盖军事领域问题并反映其知识结构。

在接下来的测试中，各测试者需要使用 MilRE 数据集作为测试数据来源，从每一类问题中随机选取 5 个样本，总共抽取 115 个样本用于后续评估。MilRE 数据集包含 23 个不同的问题类别，因此 115 个问题覆盖了数据集中的所有类别。每个类别抽取 5 个样本，涵盖了每个类别的数据多样性和特征。基于以上方法抽取的样本，测试人员使用本文设计的框架调用 GLM4[20] 进行实验。测试流程包含两个步骤：1)衡量检索结果的正确率，对应框架中的知识精炼结果；2)评估框架的总体效能，对应框架的求解结果。

3.2 框架测试效果评估

在对问答框架进行评估的过程中，由于用户的主观期待和理解上存在差异，各用户对答案的满意度可能会有所不同。此外，大语言模型的输出往往表现出一定程度的不一致性或不确定性。为确保评估的客观性和全面性，本研究聘请了两名专业评测人员参与了测试工作。在测试过程中，评测人员会对答案进行标记：若判断答案为正确，标记为“TRUE”；若判断为不正确，标记为“FALSE”。在测试中，我们特别关注以下两种特殊情况：1)在评估知识检索正确率时，若查询结果显示知识库中缺少或只有部分与问题直接相关的信息，则此类输出仍然被判断为正确检索；2)在评估问答准确率时，在知识库的内容不足以全面回答问题的情况下，如果模型仍能明确表示出无法求解的原因，例如通过声明“无法解答”或“比较困难”并附带正确的理由，则认为问题解答正确。

3.3 实验结果与分析

实验结果如表 2 所列。

表 2 框架测试结果

Table 2 Test results of framework

测试人员	测试条数		检索正确率/%	问答正确率/%
A	实体问题	80	93.75	92.50
	事件问题	35	88.57	85.71
	问题总和	115	92.17	90.43
B	实体问题	80	95.00	92.50
	事件问题	35	91.43	91.43
	问题总和	115	93.91	92.17

从框架的测试结果来看，无论是 A 还是 B 测试人员，其在实体问题处理上的检索和回答正确率均超过 90%，表明了本框架在此类问题上的可靠性。在处理事件问题时，尽管 A 和 B 两位测试人员所测得的检索正确率都保持在 88%以上，但 A 所得的问答准确率有所下降，为 85.71%，表明框架在事件问题上的问答正确率略低于在实体问题上的表现。整体而言，在两位测试人员的测试中，框架在所有问题类型上的表现均保持高水平，检索和问答的正确率均超过 90%。以上评估显示，本文提出的框架对于解决实体和事件的问题均有高正确率。

结束语 本文提出了一种基于知识图谱的检索增强生成框架，以实现军事情报问答。该框架通过问题分类、实体识别、实体链接、知识检索有效获取问题背景知识。同时，考虑到情报问题多约束的特点，通过回答集编程，表示并利用约束条件减少知识数量或者直接推理获得答案。最终，大语言在精炼后的背景知识基础上求解并回答问题，减少了在问题理解过程中对于属性识别与链接的依赖。在 MilRE 数据集上的实验表明，框架能够提供基于知识图谱的增强知识检索功能，并提供了较好的军事情报问题解答能力。但文中工作仍存在以下不足需要改进。

1)意图识别的优化：为了降低知识检索过程的复杂性，可以摆脱传统依赖问题模板设计和问题分类模型的方法，转而探索约束识别的方法，通过设定约束类型并生成相应的 ASP 规则实现高效的知识获取。

2)ASP 事实的动态化表达改进：由于当前的事件问题处理受限于背景知识的固定表示形式，在应对以非七元组的形式表示的约束时存在一定困难。后续的研究将探索采用新的数据结构，以期实现对 ASP 事实和规则更加灵活的表示。

参 考 文 献

[1] PAN S, LUO L, WANG Y, et al. Unifying large language models and knowledge graphs: A roadmap[J]. arXiv. 2306. 08302, 2023.

[2] PENG H. Design and Implementation of Military Knowledge Q&A System Based on Knowledge Graph [D]. Beijing: Beijing University of Posts and Telecommunications, 2022: 25-45.

[3] GUO A B. Research on Key Technologies and Systems of Intelligent Intelligence Q&A [D]. Changsha: National University of Defense Technology, 2021: 8-11.

[4] XU P K. Research on Key Technologies for Semantic Analysis of Questions in Military Knowledge Q&A [D]. Changsha: National University of Defense Technology, 2021: 34-46.

[5] FAN J J, MA H Q, LIU X L. Research on Military Knowledge Graph Q&A Intelligent Service for Open Source Intelligence in the Digital Intelligence Era [J/OL]. Data Analysis and Knowledge Discovery: 1-15. [2024-06-30]. http://kns. cnki. net/kcms/detail/10. 1478. G2. 20231026. 1305. 002. html.

[6] LAN Y, HE G, JIANG J, et al. Complex knowledge base question answering: A survey[J]. IEEE Transactions on Knowledge and Data Engineering, 2022, 35(11): 11196-11215.

[7] HUANG L, YU W, MA W, et al. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions[J]. arXiv: 2311. 05232, 2023.

[8] TONMOY S M, ZAMAN S M, JAIN V, et al. A comprehensive survey of hallucination mitigation techniques in large language models[J]. (arXiv: 2401. 01313, 2024.

[9] SCHLAG I, SUKHBAATAR S, CELIKYILMAZ A, et al. Large language model programs[J]. arXiv: 2305. 05364, 2023.

[10] LEWIS P, PEREZ E, PIKTUS A, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks[J]. Advances in Neural Information Processing Systems, 2020, 33: 9459-9474.

[11] GAO Y, XIONG Y, GAO X, et al. Retrieval-augmented generation for large language models: A survey [J]. arXiv: 2312. 10997, 2023.

[12] NAKANO R, HILTON J, BALAJI S, et al. Webgpt: Browser-assisted question-answering with human feedback [J]. arXiv: 2112. 09332, 2021.

[13] ASAI A, WU Z, WANG Y, et al. Self-rag: Learning to retrieve, generate, and critique through self-reflection [J]. arXiv: 2310. 11511, 2023.

[14] RACKAUCKAS Z. Rag-fusion: a new take on retrieval-augmented generation[J]. arXiv: 2402. 03367, 2024.

[15] PAWAR S, TONMOY S M, ZAMAN S M, et al. The What, Why, and How of Context Length Extension Techniques in Large Language Models--A Detailed Survey [J]. arXiv: 2401. 07872, 2024.

[16] WANG X, SALMANI M, OMIDI P, et al. Beyond the limits: A survey of techniques to extend the context length in large language models[J]. arXiv: 2402. 02244, 2024.

[17] GEBSER M, KAMINSKI R, KAUFMANN B, et al. Answer set solving in practice[M]. Morgan and Claypool Publishers, 2022.

[18] LIU W, ZHOU P, ZHAO Z, et al. K-bert: Enabling language representation with knowledge graph[C]// Proceedings of the AAAI Conference on Artificial Intelligence. 2020: 2901-2908.

[19] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv: 1810. 04805, 2018.

[20] TEAM G L M, ZENG A, XU B, et al. ChatGLM: A Family of Large Language Models from GLM-130B to GLM-4 All Tools [J]. arXiv: 2406. 12793, 2024.

CHENG Zhiyu, born in 1999, postgraduate. His main research interests include knowledge graph and question answering systems.

ZHANG Zhizheng, born in 1980, Ph.D, associate professor, is a member of CCF (No. 32012M). His main research interests include knowledge representation and reasoning, and knowledge agents.

（责任编辑：柯颖）