基于问题扩展的散文答案候选句抽取方法研究

雷洋[1]，王素格[1,2,*]，李书琪[1]，王浩[1]
1.山西大学计算机与信息技术学院，山西省 太原 030006
2.山西大学计算智能与中文信息处理教育部重点实验室，山西省 太原 030006
wsg@sxu.edu.cn

摘要

在散文阅读理解中，一方面问题的题干通常较为简洁、用词较为抽象，机器难以直接理解问题的含义和要求；另一方面，散文文章较长，答案候选句分散在文章的多个段落，给答案候选句的抽取任务带来巨大的挑战。因此，本文提出了一种基于问题扩展的散文答案候选句抽取方法。首先，利用大语言模型抽取文章中与问题题干相关的词，构建问题词扩展库，其次，利用大语言模型强大的生成能力对原问题的题干进行重写，进一步，利用问题词扩展库对其扩展，最后，通过对散文文章分块处理，建立基于全局上下文信息、历史信息的问题和文章句子的相关性判断模型，用于抽取答案候选句。通过在散文阅读理解数据集上进行实验，实验结果表明本文提出的方法提高了散文抽取答案候选句的准确率，为散文阅读理解的生成类问题的解答提供了技术支撑。

关键词： 文本抽取 ； 问题词扩展库 ； 散文体裁

Sentiment classification method based on multitasking and multimodal interactive learning

Yang Lei[1]，Suge Wang[1,2,*]，Shuqi Li[1]，Hao Wang[1]，
1.School of Computer and Information Technology,
Shanxi University,Shanxi 030006
2.Key Laboratory Computational Intelligence and Chinese Information
Processing of Ministry of Education, Shanxi University,Shanxi 030006
wsg@sxu.edu.cn

Abstract

In prose reading comprehension, on one hand, the questions are usually concise and abstract, making it difficult for machines to directly understand the meaning and requirements of the questions. On the other hand, prose passages are lengthy, and candidate answer sentences are dispersed across multiple paragraphs in the passage, posing significant challenges for sentence extraction tasks. Therefore, this paper proposes a question-enhanced approach for extracting candidate answer sentences from prose passages. Firstly, relevant words in the passage are extracted using a large model to build a question word extension library. Subsequently, the powerful generative capability of the large model is utilized to rewrite the original question, and then the question word extension library is employed for further expansion. Lastly, by segmenting the prose passage, a model is established to assess the relevance between questions and sentences based on global context information and historical information, facilitating the extraction of candidate answer sentences. Experimental results on a prose reading comprehension dataset demonstrate that the proposed method improves the accuracy of

第二十三届中国计算语言学大会论文集，第613页-第624页，太原，中国，2024年7月25日至28日。
卷1：主会论文

extracting candidate answer sentences from prose passages, thereby providing technical support for answering generative-type questions in prose reading comprehension tasks.

Keywords: text extraction , question expansion library , prose genre

1 引言

机器阅读理解是自然语言处理领域的一个重要研究方向，其目标是让机器理解文本中的语义并回答相应的问题。现有的机器阅读理解方法大多针对文本长度较短、问题形式相对固定的数据集。散文作为作者表达自身思想情感的载体，题材广泛、结构自由，具有深邃的意境和丰富的情感，使得针对散文的机器阅读理解任务更具挑战性。以表1的《遍野荆花》为例，从问题题干可以看出，题干内容精简，用词抽象，且与文章内容的字面呼应较少，若直接将问题与文章内容进行匹配时，导致机器难以理解问题中的抽象词，如“外部环境”，且问题与文章内容之间的关联困难，从而影响问题与文章内容的匹配，降低机器抽取答案候选句的准确率。另外，散文通常较长，我们对散文数据集SWQA统计发现，每篇散文平均包含1500字左右，且包含10个段落以上，而每个问题的答案候选句平均分布在3.5个段落以上，这将导致全面获取候选答案句较为困难。

针对阅读理解任务中问题题干有效信息较少的情况，现有研究(Yasunaga et al., 2021)通常引入知识图谱或外部知识对问题扩展，但对于散文体裁的阅读理解问题，题干用词抽象和多样，难以利用已有的知识或其他外部资源直接进行扩展。因此，本文利用大语言模型对原问题进行扩展和丰富语义。首先利用大语言模型抽取文章中与问题题干相关的词，构建问题词扩展库，再利用大语言模型强大的生成能力和丰富的词汇信息对原问题的题干进行重写。例如，表1中的问题，基于对文章内容的理解，大语言模型重写后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？”。经过重写后，题干中融入了文章中的部分信息，使问题与文章的内容可以相呼应。随后调用问题词扩展库中具体的环境描绘词对“环境描写”做出解释。最终处理后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？（水源缺乏，土壤贫瘠，怪石遍布，植被稀疏）”，由此可以看出通过题干重写和问题扩展，增强了问题与文章之间的联系。

对于散文长度较长、答案候选句分散的问题，预训练模型普遍存在输入长度受限的问题，如BERT模型最大输入长度为512个token，因此无法直接使用预训练模型应用于散文文本。研究者们一般采用滑动窗口(Zhang et al., 2023)、直接截断(Zhang et al., 2023)的方法处理长文本。而直接截断的方法势必会丢失一部分文章信息，影响模型的召回率。滑动窗口虽然分段将文本输入到模型中，但是窗口分割可能导致文本中重要的语境信息被切分，降低模型对整体文本的理解能力。因此，在一次不能将文章全部输入的情况下，我们参考文献(侯祺积, 2023)，将文本分为多个子块，若文章的长度超过每个子块的长度，将剩余部分划分到下一个子块中，以此类推，直至将长文本处理结束。面对答案候选句分散的情况，常见的多片段抽取模型(Deng et al., 2021; Yang et al., 2021)存在抽出的句子重复率较高、准确率较低的问题。因此，本文提出了问题与文章句子相关性判断模型，引入了历史上下文信息编码模块，在待抽取句子和已抽取句子间使用自注意力机制，降低了模型抽取答案候选句的冗余率。最后，将问题与分块后的散文输入到相关性判断模型中，实现散文答案候选句抽取。

本文的主要贡献如下：

（1）针对散文问题题干较为抽象的特点，设计了基于大语言模型的问题重写策略，同时利用大语言模型文本生成能力，构建了问题词扩展库，增加了问题的有效信息，丰富了问题的题干。

（2）针对候选句散落在散文的多个段落中，提出了问题与文本相关性判断模型，使模型可以捕获文本语句间的长距离依赖关系，提高了与问题内容相关的答案候选句的抽取性能。

* 通讯作者 Corresponding Author.
基金项目：国家自然科学基金项目(62076158,62376143)

文章	…… ……崮崖是个小山村，全村500多口人，只有不到300亩山岭地， 却有6000多亩山场。可这么多山场有什么用? 这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！…… …… ……站在这里放眼四望，山野怪石遍布，植被稀疏。 多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。 刺槐之类的，也长得歪歪扭扭，一副苦大仇深的架势，他的心里更加迷茫了…… …… 崮崖村生产的荆花蜜，色如纯净琥珀，入口留香绵长，投放市场后供不应求。 从此，这遍野荆花成为村里永不枯竭的财富之源……（有删改）
问题	文章通过哪些外部环境描写展现了农村的贫困?
候选句	崮崖是个小山村，全村500多口人，只有不到300亩山岭地，却有6000多亩山场。 可这么多山场有什么用？这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！ 站在这里放眼四望，山野怪石遍布，植被稀疏。 多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。 刺槐之类的，也长得歪歪扭扭，一副苦大仇深的架势，他的心里更加迷茫了。

表 1: 散文《遍野荆花》节选

（3）在散文阅读理解数据集上，将本文方法与对比方法进行了实验，结果表明本文方法的性能优于其他对比方法。

2 相关工作

对于机器阅读理解任务，开放域问答是此类应用研究的重要组成部分，该任务是通过从大量文档中收集证据并回答相应的问题。针对散文阅读理解的答案候选句抽取，需要多片段文本抽取技术的支持，目前常用的机器阅读理解的方法是基于预训练模型方法以及结合外部知识的抽取技术。

在基于预训练模型的抽取式机器阅读理解方法中，Cui等人(2020)在RoBERTa的基础上，提出了MacBERT模型，该模型利用相似的单词做MASK掩码，缩小了预训练和微调阶段间的差距，且在多个中文NLP数据集上取得显著效果，因此，MacBERT可用于散文阅读理解的数据编码；Deng等人(2021)提出ReasonBERT，通过增强模型在长文本和多重上下文的推理能力，可为解决散文长文本的问题提供思路。Khattab和Zaharia (2020)提出了COLBERT模型，不同于常见的All-to-all交互方式，该模型具有延迟交互的能力，以提高问题和文本之间的交互效率。Yavuz等人(2022)提出了PATHFID模型，线性化支持段落的分层推理路径及关键句子，在提升模型性能的同时，使得到的问题候选答案句更具可解释性。Zhan等人(2021)采用对比学习的思想，提出了动态硬负采样的训练策略，在保持较好稳定性的同时，能够关注排名靠前的答案候选句。Gu等人(2022)提出了利用多步骤情节性马尔可夫决策过程抽取长文本的方法。侯祺积(2023)提出了Top-MRV模型，引入问题词扩展集来划分子句权重，提升了模型抽取答案候选句的准确率。

近年来，将外部知识引入MRC模型成为学者们的研究热点，形成了基于外部知识库的机器阅读理解（Knowledge-Based MachineReading Comprehension，KBMRC）。KBMRC与MRC的主要区别在于输入，MRC的输入是文本和问题序列，而KBMRC(Wang J et al., 2022)的输入除此之外还从知识库中提取出额外相关知识。Wang等人(2023)利用知识引导的提示作为干预，提升了模型在文本概念提取任务的准确率。

上述模型直接用于散文阅读理解任务，存在理解散文文章语义不全面，抽取出的句子准确率较低、冗余率较高等问题。而散文阅读理解中的问题与文章内容有密切的关系，若利用这些信息可以帮助模型实现答案候选句的抽取。例如，散文中对人物具体的描绘、环境的描写，这些信息可以用于丰富问题题干，提供额外的语义信息，以此提升散文答案候选句的抽取性能。

3 散文问题扩展和答案候选句抽取模型

针对散文答案候选句的抽取任务，受文献(侯祺积, 2023)构建多片段答案关键句抽取框架方法的启发，一方面，利用大语言模型和问题抽象词扩展库对问题进行重写和扩展，另一方面，将文章进行分块处理，并将分块后的文章和问题输入到相关性判断模型。在相关性判断模型中，全局上下文编码模块对输入数据进行编码，得到文章句子的全局向量表示。历史信息编码模块包含两个多头自注意力层，分别捕获待抽取句子间的交互信息和待抽取句子与已抽取句子间的交互信息。两层自注意力层的最终输出构成待抽取句子的历史信息表示。将每个待抽取句子的全局向量与历史信息向量进行连接，形成聚合向量。聚合向量经过映射后得到每个问题的答案候选句得分，选取得分最高的句子作为答案候选句。模型整体框架如图1所示。

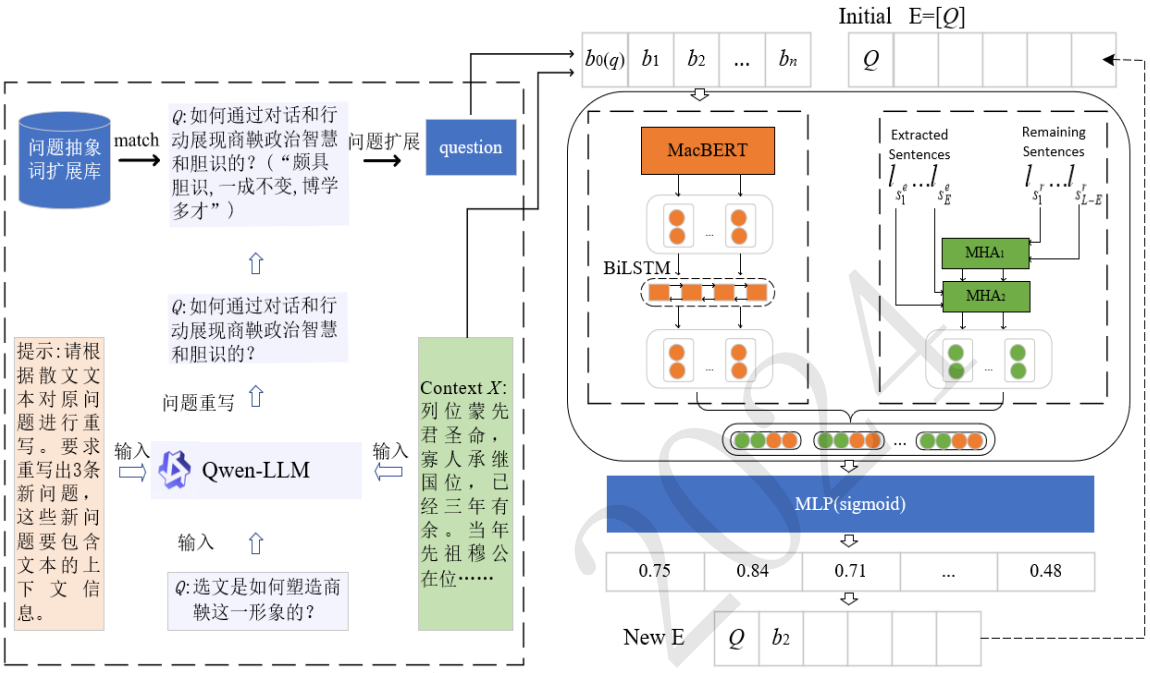

图 1: 散文问题扩展与答案候选句抽取模型框架图

3.1 基于大语言模型的问题扩展

由于散文阅读理解中的问题简洁、用词抽象，影响了答案候选句的抽取结果，为此，本文研究基于大语言模型的问题扩展。主要包括问题词扩展和问题重写。

3.1.1 问题词扩展库的建立

通过分析数据集中的各类问题，我们发现在散文人物形象类和环境描写类问题中，经常出现一些抽象词，如“形象”和“环境描写”，模型对这类抽象词的理解较为困难，进而可能会影响问题与文章内容的相似度匹配，导致散文答案候选句抽取的质量下降。为此，可以根据问题题干中提及的具体人物，从文章中找到相关的人物形象描述，将这些人物描写词扩展到问题中。类似地，针对环境描写类问题，从文中找到相关的环境描写词，扩展到问题中，以具体化抽象词，使机器更好地理解抽象概念。

本文使用千问大语言模型构建问题扩展库，具体过程如下：

（1）人物形象和环境描写词汇抽取：利用大语言模型识别出文本中出现的人名实体，然后针对各人物抽取对应的人物形象词。类似地，抽取描写环境的词语。

（2）扩展词选取：在散文文章中，一方面，对一些人物的描述较为简略，导致可供抽取的形象词数量较少；而对主要人物往往赋予较多的描写，使得可抽取的形象词数量较多。因此，在文章中不同人物的描写词数量存在差异。为了获取问题的扩展词，本文制定以下策略：选择每篇散文中人物描写或环境描写的前4个词作为扩展词。若描述词数量少于4个，则会全部选取作为问题扩展词。

例如，在《商鞅变法》一文中，问题: “选文是如何塑造商鞅这一形象的？”，该文中并未直接提及“形象”这一词汇，题干中“形象”一词与全文内容匹配度较低，但文章中出现了对商鞅具体的人物描写，若将这些形象词扩展到问题中，对抽象词起到解释的作用。扩展后的问题将变为：“选文是如何塑造商鞅这一形象的？(“颇具胆识，一成不变，博学多才，贸然行事”)”。对于环境描写类问题，采用类似的扩展方法。最终，本文共建立了7140条问题扩展词，此外，需要注意的是本文的问题词扩展库是以文章为组别的。样例如表2所示。

人物对象/环境	人物描写/环境描写
商鞅	颇具胆识，一成不变，博学多才，贸然行事
秦孝公	奇耻大辱，不悦，殷纣夏桀，激动，易弦更张
公孙大人	妄议朝政，巧舌如簧，腐朽
公子虔	博学多才，不拘古法
祝欢	早有征兆，不敢妄言
环境	暴风雪，零下三十摄氏度，温度降低，狂风

表 2: 问题抽象词库示例

3.1.2 问题重写

通过对散文数据的分析，绝大部分散文阅读理解问题的题干较为简洁，例如：“选文是如何塑造商鞅这一形象的？”，问题中的关键信息为“商鞅”和“形象”，仅依靠这两个词语，模型无法有效学习到问题题干与文章内容的联系，影响了模型对答案候选句的判断。因此，针对题干中有效信息较少的问题，本文利用大语言模型强大的文本生成能力，构建基于大语言模型的问题重写过程，以增强并扩充题干中的上下文信息。具体过程如下：

（1）提示信息建立：依据文章内容对原问题进行重写，要求对原问题重写出3条问题，这些重写的问题要求包含文本的上下文信息。

（2）问题重写：将原问题、文章内容和提示信息作为输入，利用阿里云的通义千问大语言模型，生成与原问题语义相似的问题重写，这些问题重写可能在表达方式、结构或语气上与原始问题有所不同，但在保持原问题相同的语义和意图的基础上，增加了题干中的信息量，丰富了原问题，具体形式如下所示：

$$\{Q_{\text{rewrite}}\} = \text{Qwen}\left([Q; \text{Context}; \text{Prompt}]\right)$$

其中，$\{Q_{rewrite}\}$表示重写问题集合。

（3）问题筛选：在生成新问题后，进行问题筛选。将重写后的问题与文章内容、原问题分别做相似度计算，将二者计算后的结果求和排序，选取相似度得分最高的问题作为重写后的问题。

本文使用基于TFIDF(张兆滨等, 2020)的相似度计算方法，具体计算见公式(1)-(4)：

$$tf_{S,w_i} = \frac{C_{S,w_i}}{\sum_k c_{k,w_i}} \tag{1}$$

$$idf_{w_i} = \log(\frac{|D|}{|1 + \{S : w_i \in S\}|}) \tag{2}$$

$$tfidf_{S,w_i} = tf_{S,w_i} \cdot idf_{w_i} \tag{3}$$

其中，C_{S,w_i}是词w_i在句子S中出现的次数，$|D|$是句子总数，$|\{S : w_i \in S\}|$是包含词w_i的句子数。tf_{S,w_i}和idf_{w_i}分别为词w_i的词频和逆文档频率。

重写问题Q与文章句子S_g的相似度$\cos(Q, S_g)$，具体计算见公式(4)。

$$\cos(Q, S_g) = \frac{\sum_{i=1}^{d_w} (tfidf_{Q,w_i} \times tfidf_{S_g,w_i})}{\sqrt{\sum_{i=1}^{d_w} (tfidf_{Q,w_i})^2} \times \sqrt{\sum_{i=1}^{d_w} (tfidf_{S_g,w_i})^2}} \tag{4}$$

其中$tfidf_{Q,w_i}$表示词w_i在问题Q中的TFIDF值，$tfidf_{S_g,w_i}$表示词w_i在文章句子S_g中的TFIDF值，d_w表示词向量的维度。

类似地，可以定义重写问题Q与原问题q的相似度$\cos(Q, q)$。

对两类信息的相似度得分求和，得到每个重写问题最终的得分，具体计算见公式(5)。

$$grade(Q, S_g, q) = \cos(Q, S_g) + \cos(Q, q) \tag{5}$$

对于《商鞅变法》一文，原问题为“选文是如何塑造商鞅这一形象的？”，重写后的问题为“如何通过对话和行动展现商鞅的政治智慧和胆识？”。重写后的问题相比于原问题内容更具体详尽，同时引入了文中的部分相关情节，提供给模型更多的上下文信息。

3.2 答案候选句抽取模型构建

为了实现答案候选句的抽取，设计了相关性判断模型，该模型首先利用文本分块处理散文长文本，其次，利用全局上下文信息和历史信息编码模块对问题与句子间的关系做出判定，最后根据判定分数，选取答案候选句。

（1）分块处理

预训练语言模型MacBERT在处理输入序列时，通常有长度限制，为了最大程度地输入文本内容，对文本进行了分块处理。具体地，将文本分成n个子块，将问题q记作子块b_0，文本记为子块$b_1, ..., b_n$。每个子块的长度为$l = length/n$，其中，$length$为整个文本的长度。对于每个子句s_k，若长度超过了l，则将超出部分的子句划分到下一个子块b_{i+1}中。同时，下一个子句s_{k+1}将划分到b_{i+2}中，以此类推，直到子块的个数达到n，如图2所示。

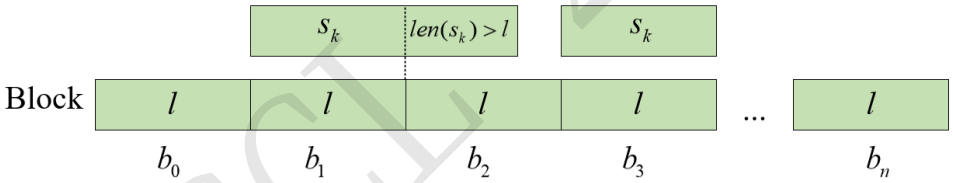

图 2: 子句超出固定长度l的情况

在数据处理中，对于文本的分块处理是关键的。考虑到问题句与散文文本之间的相关性可以为答案候选句抽取提供重要的信息。因此，在训练数据集中对答案候选句进行了标注，用于表示Block中文本与问题之间的相关性，具体形式如下所示：

$$R(Block) = [\underbrace{1, 1, ..., 1}_{b_0(q)}, \underbrace{0, 0, ..., 0}_{b_1}, \underbrace{1, 1, ..., 1}_{b_2}, ...] \in [0, 1]^{len(Block)}$$

其中，Block中的第一块b_0代表问题块，接着的n个子块则代表文本块。在b_1到b_n中，如果b_i是答案候选句，则标注为1,1,...,1；否则标注为0,0,...,0。

（2）相关性判断模型

相关性判断模型由全局上下文编码模块、历史信息编码模块和MLP三部分组成。

全局上下文编码模块：设初始文本序列为X，首先使用MacBERT(Cui et al., 2020)，将块中信息映射到一个高维向量空间中，获得包含文本信息的向量表示M，在此基础上，使用BiLSTM(Huang et al., 2015)建立各块间的前后语义关联，获得各块中句子的上下文信息表示G_{S_i}，具体表示见公式(6)-(8)：

$$X = (Q, b_1, b_2, ..., b_n) \tag{6}$$

$$M = \text{MacBERT}(X) \tag{7}$$

$$G_{s_i} = \text{BiLSTM}(M) = (l_{s_1}, ..., l_{s_n}) \tag{8}$$

历史信息编码模块：设已抽取句子序列为E，待抽取句子序列为V，首先在待抽取句子间执行多头自注意力，以使每个待抽取句子捕获其他剩余句子提供的上下文信息，记作V_{MHA}，其次在V_{MHA}和已抽取句子序列E之间执行多头自注意力，得到每个待抽取句子的历史信息表示$H_{s_i^r}$。连接待抽取句子的全局上下文表示G_{s_i}和历史信息表示$H_{s_i^r}$，得到拼接后的向量H_i^r，具体表示见公式(9)-(13)：

$$E = (l_{s_1^e}, ..., l_{s_E^e}) \tag{9}$$

$$V = (l_{s_1^r}, ..., l_{s_{n-E}^r}) \tag{10}$$

$$V_{MHA} = \text{MHA}_1(V, V) \tag{11}$$

$$H_{s_i^r} = \text{MHA}_2(V_{MHA}, E) \tag{12}$$

$$H_i^r = \text{concate}(G_{S_i}, H_{s_i^r}) \tag{13}$$

将拼接后的向量H_i^r输入到MLP中，再通过sigmoid函数得到问题和文章句子间的相关性得分$estimate(s_i)$，具体表示见公式(14)：

$$\text{estimate}(s_i) = \text{sigmoid}(W H_i^r + b) \tag{14}$$

其中,W为可训练参数，b为偏置，$estimate(s_i) \in [0, 1]$。

（3）获取答案候选句

根据判断模型计算出的问题和句子间的相关性得分$estimate(S)$，选择得分较高的前k个子句作为最终答案候选句。

3.3 模型优化

该模型的优化目标是使预测的分数$estimate(S)$与标准分数无限接近，本文采用了最小化交叉熵损失的方法，损失函数如式(15)所示。

$$\text{Loss}_{\text{estimate}} = \text{CrossEntropy}(\text{estimate}(S), R(\text{Block})) \tag{15}$$

其中,$R(Block)$表示问题与文本句子间的真实标签。

4 实验

4.1 数据集

本文采用散文阅读理解数据集SWQA，该数据集由两部分构成：它是由历年高考语文真题、模拟题、网络资料、线下纸质题集，数据集共3790篇，记作SWQA_1；科大讯飞提供的组卷网上的散文阅读理解数据集共11560篇，记为SWQA_2；将散文数据集的SWQA_1和SWQA_2进行随机排序合并，再采用json格式进行存储，合集记为SWQA。共录入数据15350份，将其按照8：1：1的比例随机划分为训练集、验证集和测试集，如表3可见。

由于散文篇幅较长，答案分散，题型类型多样。根据题型的特定关键词，将散文的阅读理解问题分为观点理解类、概括类、作用类、原因类、表现手法类、标题理解类、赏析类七类，如表4可见。

按照各类别对散文的篇幅平均长度、平均段落量、答案在段落中平均分布量进行统计，结果如表5可见。

数据集	数据量	训练集	验证集	测试集
SWQA_1	3790	3032	400	358
SWQA_2	11560	9245	1140	1175
SWQA	15350	12277	1540	1533

表 3: 数据集规模介绍

问题类别	特定关键词
概括类	概括、特点、形象、特色等
观点理解类	观点、理解、启示等
作用类	作用、妙处、好处等
原因类	原因、为什么、解释等
表现手法	表现手法、手法、描写手法等
标题理解类	标题、题目等
赏析类	赏析、鉴赏等

表 4: 题型分类关键词示例

4.2 评价指标及参数设置

本文使用精确率P（Precision）、召回率R （Recall）和F1 值（F1 Score）作为模型的评价指标。实验采用Pytorch 1.12.1 框架，GPU 选用NVIDIA TITAN RTX。具体参数设置如表6所示。

4.3 实验设计

为了验证本文中的答案关键句抽取方法的有效性，我们选择了一些片段抽取任务中的模型作为对比实验，具体模型如下：

DRhard(Zhan et al., 2021)：基于RoBERT的编码层，分别将查询和文档编码为低维嵌入，并使用查询嵌入进行高效的相似性搜索进行在线排名，此外硬负采样通过最小化前K对误差，进一步改进了排名性能。

ColBERT(Khattab and Zaharia., 2020)：将查询和文档编码为BERT上下文的嵌入，通过计算它们之间的相似性来评估相关性。此外，ColBERT的后期交互机制支持直接从大型文档集合中进行端到端的检索，进一步提高了检索效率和召回率。

MonoQA(Kongyoung et al., 2022)：查询和文档经由RoBERT构成的ConvDR检索模型编码为向量形式，这些向量经过点积运算以后得到Top-K个文档，将Top-K个文档连同查询经由Reranker再做一次排序，得到最终要选择的文档。

Block-Skim(Guan et al., 2022)：对Transformer模型进行改进，引入了一个注意力权重热图，并对其进行了处理。利用卷积神经网络来处理这个热图的对角线部分，以确定文本是否需要被保留。我们设置了一些参数，learning rate：$3e^{-5}$，batch size：8，seed：42，训练周期：50，最大序列长度：512，步幅：12。

CogLTX(Ding et al., 2020)：结合BERT（BERTbase）与多层感知机制的循环迭代模型，参数：learning_rate：$1e^{-4}$，batch size：8，weight_decay：0，epochs：70，Attention heads：12。

Top-MRV(侯祺积, 2023)：以MacBERT为编码器，引入了问题词扩展用于划分子句的初

问题类别	SWQA_1	SWQA_2	字数/篇	段落/篇	答案分布段落/篇
概括类	739	2248	1518.62	11.75	3.87
观点理解类	527	2554	1517.90	11.12	3.76
作用类	723	2071	1577.44	10.57	4.69
原因类	534	1236	1419.15	10.34	3.58
表现手法	381	938	1483.73	10.54	3.63
标题理解类	468	1102	1545.11	11.66	4.46
赏析类	418	1411	1597.54	12.09	4.73

表 5: SWQA_1和SWQA_2题型分布比例

超参数名	参数值
Dropout	0.1
Learning rate	0.0001
Batch size	8
Hidden size	768
Hidden layer	12
Max sequence length	512
Weight decay	0.0001
Attention heads	12

表 6: 超参数设置

始权重，将初始权重与预测的子句权重动态调整，增强了模型抽取答案关键句的能力，基线模型。

4.4 实验结果及分析

4.4.1 实验比较及分析

将本文的方法与第4.3节的六种方法进行对比实验，实验结果如表7所示。

Model	P	R	F1
ColBERT	35.17	39.40	30.27
DRhard	39.10	39.23	35.13
MonoQA	49.91	37.06	42.53
Block-Skim	58.79	49.84	50.04
CogLTX	59.22	56.71	55.63
Top-MRV	70.21	64.39	65.22
OURS	73.54	64.46	67.00

表 7: 本文方法与六种对比方法的实验结果

通过表7可以看出：

（1）本文提出的模型在精确率P、召回率R和F1值评价指标中均优于其他六种模型，验证了本文方法在散文中抽取答案候选句的有效性。

（2）本文提出的模型相比于检索模型ColBERT和DRhard，在F1值上分别提高了36.73和31.87。对于ColBERT模型，针对散文文本较长时未考虑文本分块，缺乏长距离语义交互，导致模型表现不佳。另外，虽然DRhard模型是对句子进行处理，在F1值上高出ColBERT模型3.93，但是所有文章的句子全都打乱混在一起，增加了搜索的难度，模型效果也较差。

（3）本文提出的模型相比于MonoQA模型在F1值上提高了24.77，主要原因是MonoQA模型是用于阅读理解生成，无法发挥多任务学习的效果。

（4）相较于Block-Skim和CogLTX模型，本文模型在F1值上分别提升了16.96和11.37。两者都对长序列进行了分块处理，其中Block-Skim使用了注意力权重热图的对角线部分判断答案是否保留，这种模型虽然提高了模型的运行速度，但是对于联系上下文更为密切的散文阅读理解任务而言，其表现并不理想。而CogLTX利用判断模型对文章选择性截取来实现缩短长文的效果，但是经过处理后的文章丢失了语句间的关联性，给阅读理解任务带来了语义理解上的困难。

（5）相比于Top-MRV模型略有提升，主要原因是本文模型利用大语言模型对原问题进行了重写，并引入了问题抽象词扩展库，具体化了问题题干的抽象词，丰富了原问题的有效信息量，此外，连同相关性判断模型的引入，共同提升了答案候选句抽取的准确率。

第二十三届中国计算语言学大会论文集，第613页-第624页，太原，中国，2024年7月25日至28日。
卷1：主会论文

4.4.2 消融实验

为了验证本文模型在各个部分的性能，本文设计了消融实验。

(1)-Block：将本文模型去除分块模块；

(2)-Rewriting Question：将本文模型去除问题重写模块；

(3)-Abstract feature：将本文模型去除问题抽象词扩展库；

(4)-Judge：将本文模型去除相关性判断模型；

(5)-History Encoder：将本文模型去除历史信息编码模块；

消融实验结果如表8所示。

Model	P	R	F1
ours	73.54	64.46	67.00
-Block	65.23	59.24	62.09
-Rewriting Question	72.63	64.02	66.20
-Abstract feature	72.46	63.40	65.93
-Judge	45.62	70.22	55.25
-History Encoder	72.56	63.70	66.08

表 8: 消融实验结果

由表8可以看出，模型去除任意子模块，其综合性能都有一定的下降，模型去除各个模块后的表现如下：

（1）-Block相比于本文方法的P、R和F1值分别下降8.31、5.22、4.91，可以看出P值下降较大。主要原因是文本分块保留了散文的上下文，去除该模块以后丢失了大量文本信息，造成模型P值大幅下降。

（2）-Rewriting Question，相比于本文方法的P、R和F1值分别下降0.91、0.44、0.8，这是因为问题题干中的有效信息量减少，模型无法有效学习到问题题干和文本之间的关系，造成模型的性能下降。

（3）-Abstract feature，相比于本文方法的P、R和F1值分别下降1.08、1.06、1.07，是由于本文模型中的问题抽象词扩展库可以帮助模型解决问题题干中难以理解的语义部分，提升了模型的语义理解能力。

（4）-Judge，相比于本文方法的P和F1值分别下降27.92和11.75，召回率达到最高，抽取出的句子的准确率较低且冗余率较高，因此，不能直接用于实际阅读理解答题中，应融合其他模块对文本进一步筛选。

（5）-History Encoder，相比于本文方法的P、R和F1值分别下降0.98、0.76、0.92，这是因为提取历史编码器可以从待抽取句子的角度对已抽取句子考虑，若去掉该模块，降低了抽取的准确性。

综合上述分析可知，将问题重写模块、问题抽象词扩展库模块、文本分块和判断模型四个部分融合起来，将有助于提高散文答案候选句的抽取能力。

4.4.3 实例分析

为了直观地了解本文方法抽取答案候选句提升的效果，利用本文模型与对比方法中性能最优的方法Top-MRV的获得实验结果进行比较分析。选取阅读理解问题为：“文章通过哪些外部环境描写展现了农村的贫困？”。该问题的标注答案的候选句如下：(1)“崮崖是个小山村，全村500多口人，只有不到300亩山岭地，却有6000多亩山场。”；(2)“可这么多山场有什么用？这是水源缺乏、土壤贫瘠、几乎连一棵大树都长不起来的山岭呀！”；(3)“站在这里放眼四望，山野怪石遍布，植被稀疏。多数植被是一种叫荆棵的低矮灌木，偶有几根针叶松。”；针对上述例子，Top-MRV模型得到的正确答案候选句为(1)和(2)。本文的方法首先会对原问题进行重写和扩展，得到扩展后的问题为“文章如何通过描绘恶劣的自然环境、贫瘠的土地资源和匮乏的农业资源等外部环境，深入展现农村的贫困？（水源缺乏，土壤贫瘠，怪石遍布，植被稀疏）”。根据问题扩展后的结果，抽取的答案候选句为(1)、(2)和(3)。由此可见，通过问题扩展，本文方法可以抽取出Top-MRV模型未能抽取出的答案句(3)，说明经过问题扩展后，可以找出外部环

境描写对贫穷的隐式刻画，而没有经过扩展的问题，仅能依据问题中的显示描写如“农村”，对文章句子进行抽取。

5 结论

针对散文阅读理解答案候选句抽取问题，提出了基于问题扩展的散文答案候选句抽取方法。该方法融合了更多文本知识，利用大语言模型实现了问题重写，建立了抽象词扩展库。针对文本较长的特性，对散文文本进行了分块处理，最终构建了候选句识别模型，通过对比实验，结果表明该方法优于目前先进方法，通过消融实验，验证了问题重写、问题抽象词扩展库和判断模型，可以更加针对性地增强模型抽取散文答案候选句的能力，为散文阅读理解的生成类问答提供技术支持。

参考文献

Yiming Cui, Wanxiang Che, Ting Liu, et al. Revisiting pre-trained models for Chinese natural language processing[C]. In *Findings of the 2020 Conference on Empirical Methods in Natural Language Processing*, 2020: 657–668.

Ming Ding, Chengyue Zhou, Haoyu Yang, et al. Cogltx: Applying BERT to long texts[J]. *Advances in Neural Information Processing Systems*, 2020, 33: 12792-12804.

Xiang Deng, Yu Su, Alyssa Lees, et al. ReasonBERT: Pretrained to reason with distant supervision[C]. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021: 6112–6127.

Ning Gu, Elliot Ash, Rüdiger Hahnloser. MemSum: Extractive summarization of long documents using multi-step episodic Markov decision processes[C]. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022: 6507-6522.

Yiyang Guan, Zhiyuan Li, Zicheng Lin, et al. Block-Skim: Efficient question answering for transformer[C]. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*, 2022, 36(10): 10710-10719.

Zhiheng Huang, Wei Xu, Kai Yu. Bidirectional LSTM-CRF models for sequence tagging[J]. *arXiv preprint arXiv:1508.01991*, 2015.

Wang J, Wang C, Qiu M, et al. KECP: Knowledge enhanced contrastive prompting for few-shot extractive question answering[C]. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022: 3152-3163.

Omar Khattab, Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over BERT[C]. In *Proceedings of the 43rd International ACM SIGIR conference on research and development in Information Retrieval*, 2020: 39-48.

Sangha Kongyoung, Craig Macdonald, Iadh Ounis. MonoQA: Multi-task learning of reranking and answer extraction for open-retrieval conversational question answering[C]. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, 2022: 7207-7218.

Yang H, Sui D, Chen Y, et al. 2021. Document-level event extraction via parallel prediction networks. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing*, pages 6298-6308.

Yuan S, Yang D, Liu J, et al. 2023. Causality-aware Concept Extraction based on Knowledge-guided Prompting. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics* , pages 9255-9272.

Masashi Yasunaga, Hongyu Ren, Antoine Bosselut, et al. QA-GNN: Reasoning with language models and knowledge graphs for question answering[C]. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2021: 535-546.

Serkan Yavuz, Kazuma Hashimoto, Yanan Zhou, et al. Modeling multi-hop question answering as single sequence prediction[C]. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022: 974-990.

Jiaxin Zhan, Jiaxin Mao, Yiqun Liu, et al. Optimizing dense retrieval model training with hard negatives[C]. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2021: 1503-1512.

Hao Zhang, Haoyu Song, Shuang Li, et al. A survey of controllable text generation using transformer-based pre-trained language models[J]. *ACM Computing Surveys*, 2023, 56(3): 1-37.

侯祺积. 散文阅读理解简答题的解答方法研究[D]. 太原: 山西大学硕士学位论文, 2023.

张兆滨, 王素格, 陈鑫, 赵琳玲, 王典. 阅读理解中观点类问题的扩展研究[J]. 中文信息学报, 2020, 34(6): 89-96, 105.