Artificial Intelligence and Robotics Research 人工智能与机器人研究, 2024, 13(3), 673-683
Published Online August 2024 in Hans. https://www.hanspub.org/journal/airr
https://doi.org/10.12677/airr.2024.133068

Hans汉斯

基于多特征融合的中医药问题生成模型

酒敬天，李红莲

北京信息科技大学信息与通信工程学院，北京

收稿日期：2024年7月16日；录用日期：2024年8月21日；发布日期：2024年8月30日

摘　要

目的：提出一种基于多特征融合的中医药问题生成模型(MFFQG)，以改善现有的自动生成技术在处理特定领域时存在的领域关键词信息缺失和生成问题表达不规范问题。方法：利用RoBERTa向量和五笔向量捕捉输入序列的语义特征和字形特征，同时融合句法信息和所构建的中医药领域主副关键词信息，将得到的多特征向量信息送入UniLM生成模型得到生成结果，实现对中医药领域问题的自动生成。结果：MFFQG模型融合多种特征，在Rouge-1、Rouge-2、Rouge-L评价指标上分别达到64.93%、34.57%、63.05%。局限：数据主要来源于中医药领域，在其他领域中的效果有待验证。结论：MFFQG模型相较于对比模型，可以显著提升中医药问题的生成质量。

关键词

中医药，问题生成，句法分析，五笔特征

A Traditional Chinese Medicine Question Generation Model Based on Multi-Feature Fusion

Jingtian Jiu, Honglian Li

School of Information and Communication Engineering, Beijing Information Science and Technology University, Beijing

Received: Jul. 16th, 2024; accepted: Aug. 21st, 2024; published: Aug. 30th, 2024

Abstract

Objective: To propose a traditional Chinese medicine problem generation model (MFFQG) based on multi feature fusion, in order to improve the problems of missing domain keyword information and non-standard expression of generation problems in existing automatic generation technolo-

文章引用: 酒敬天, 李红莲. 基于多特征融合的中医药问题生成模型[J]. 人工智能与机器人研究, 2024, 13(3): 673-683.
DOI: 10.12677/airr.2024.133068

gies when dealing with specific fields. Method: Using RoBERTa vectors and Wubi vectors to capture the semantic and glyph features of the input sequence, while integrating syntactic information and the constructed main and auxiliary keyword information in the field of traditional Chinese medicine, the obtained multi feature vector information is fed into the UniLM generation model to obtain the generated results, achieving automatic generation of problems in the field of traditional Chinese medicine. Result: The MFFQG model integrates multiple features and achieves 64.93%, 34.57%, and 63.05% in Rouge-1, Rouge-2, and Rouge-L evaluation indicators, respectively. Limitation: The data mainly comes from the field of traditional Chinese medicine, and its effectiveness in other fields needs to be verified. Conclusion: Compared to the comparative model, the MFFQG model can significantly improve the quality of generating traditional Chinese medicine problems.

Keywords

Traditional Chinese Medicine, Problem Generation, Syntactic Analysis, Five Stroke Characteristics

1. 引言

中医是中国古代劳动人民智慧的结晶，到今天流传下来的中医药学著作汗牛充栋，对于中医药领域浩如烟海烟海的数据，如何提取高质量的问题用来辅助医生和患者将变得十分重要。问题生成(Question Generation, QG)是自然语言处理(Nature Language Process, NLP)的一个非常重要的分支，旨在通过给定一些文章和相应的答案生成与之对应的自然语言问题。通过自动生成的问题可以取代人工提问，让机器自动提问。随着人工智能技术的不断发展，基于端到端的问题生成技术逐渐成为主流。

然而，随着文本长度的不断增大，序列到序列的问题生成模型不可避免地存在语义信息和句法结构信息丢失的问题。为了改善以上问题，本文以“中医药”问题生成数据集为研究对象，提出一种基于多特征融合的中医药问题生成模型。人工构建的中医药领域主副关键词，基于哈工大 LTD 工具对输入序列进行句法分析以提取句法信息，利用 RoBERTa 向量和五笔向量捕捉输入序列的语义特征和字形特征，融合关键词信息和句法信息后，送入端到端的序列生成模型 UniLM 得到生成问题，实现对中医药领域问题的自动生成。主要包括以下三个创新点：

1) 提出一种基于多特征融合的中医药问题生成模型，融合五笔字型信息、关键词信息和句法信息，经过 UniLM 生成模型得到模型的最终输出结果，在“中医药”领域上得到更优的效果。

2) 通过人工标注获取“中医药”领域的主副关键词信息。通过专业人员的手动标注，捕捉问题中最重要的语义单元。将主副关键词信息融入嵌入层，使得模型能够更加精确地学习问题的关键语义特征，以提高对问题语义的理解和表达。

3) 将句法依存信息融入输入特征，使得模型更好地理解问题中的语法关系，有助于提高模型对问题结构的理解和把握。

2. 相关研究

现有的问题生成方法主要分为传统的基于规则的问题生成方法和基于神经网络的问题生成方法。传统的问题生成方法直接利用启发式规则将陈述句转换为疑问句，其严重依赖手工制作的规则和模板，导

致生成的问题具有大的同限性。

Dhole 等[1]利用通用依赖关系、浅层语义分析、词汇资源和自定义规则，将陈述句转化为问答对，并通过反向翻译在显著提高语法正确性。Labutov 等[2]将原始文本表示为低维本体，然后众包获取与该空间对齐的候选问题模板，最后为小说文本的一个区域对潜在相关的模板进行排名，可以在保持 70%的召回率的同时生成精度超过 85%的相关问题。Mazidi 等[3]构建了一个利用语义模式识别来自动生成各种深度和类型问题的自学或辅导用的自动问题生成器，可以显著降低生成句子的错误率。Heilman 等[4]使用手动编写的规则执行通用的句法转换将陈述句转化为问题，并将这些问题通过逻辑回归模型进行排名。Chali [5]等提出使用句法树核来计算问题的句法正确性，通过考虑问题在给定文本语境中的重要性和句法正确性以对问题进行排名。基于规则的方法需要专家创建规则和模板，这非常耗时而目昂贵。此外，规则和模板缺乏多样性，难以适应不同文本领域，无法实现大规模应用。

随着自然语言处理技术的不断发展，各种预训练语言模型对语义特征的捕捉能力不断增强，将其应用到问题生成领域以改善基于规则的方法的不足，成为一种可行的方案。Duan 等[6]通过在嵌入层中整合义原的外部知识，增强了文本本身的语义特征，同时在编码层后加入双向注意力流，增强了文本和答案之间的语义表示。Zeng 等[7]使用基于单词的覆盖机制进行训练，并使用不确定性感知波束搜索进行解码，提高了生成问题的质量。Hu 等[8]在解码端利用 Bert 等预训练语言模型增强文本语义表示，增强模型对语义的理解，生成更加高质量的问题。Fei 等[9]设计了基于迭代图网络解码器，在每个解码步骤中使用图神经网络[10]对先前生成的问题进行建模，并通过图模型捕捉段落中的依赖关系，提升了生成效果。Ma 等[11]设计了一个答案感知的段落表示模块，将答案信息整合到段落中，并利用基于注意力的长短时记忆解码器生成问题。Li 等[12]利用基于图交互的知识增强神经网络用于问题生成，取得了不错的效果。

基于神经网络的问题生成方法借助预训练语言模型，可以学到丰富的词汇和语义信息，使其在处理歧义性和复杂语境中表现优异，但中医药领域具有大量专有名词，且为方便医护人员和患者的理解，其生成问题有特定的表达格式，而现有方法难以准确处理专有名词并生成特定格式的问题。为了改善以上问题，本文提出一种基于多特征融合的中医药问题生成模型 MFFQG，首先通过汉字的字形差异，利用 Word2vec 词向量模型，采用五笔的方式训练大量中医药领域的词向量信息，融合 Roberta 语义特征使得模型可以更好的理解中医药词汇信息；而在针对于生成问题的关键词的缺失或者部分缺失问题，本文通过人工标注的形式，标注了大量的问题的主副关键词，加强模型对于关键信息的识别；最后为了加强生成问题的语法格式的问题，采取对生成的问句进行词性标注和句法依存度分析，使之生成的句子更加具备完整的语法特征，提升了中医药问题的生成质量。

3. 基于多特征融合的中医药问题生成模型

在本研究中，我们开发了一个基于多特征融合的中医药问题生成模型(MFFQG)，其设计过程严格遵循人类提问的逻辑流程。我们的目标是确保生成的问题不仅在语义上准确，而且符合自然语言的表达习惯。

模型的第一步是预处理阶段，涉及对输入数据的整理和格式化。在这个阶段，我们对原始文本进行必要的清洗和标准化处理，包括去除无关字符、文本分词等。预处理的目的是为 RoBERTa 模型的输入准备格式化的数据，确保其能够有效解读和处理。

预处理完成后，模型采用 RoBERTa 预训练模型来处理输入数据。在这一步，输入的句子结构被定型为“[CLS]原文[SEP]问题[SEP]答案[SEP]”。这种特定的结构使得 RoBERTa 能够清楚地区分原文、问题和答案的不同部分，并有效地提取每部分的语义信息。“[CLS]”标记用于表示句子的开始，而“[SEP]”标记用于分隔原文、问题和答案，确保模型能够理解不同部分之间的关系。

接下来，模型通过人工标注方法提取问题文本中的主副关键词信息，并利用 BERT 模型进行深入处

理。这一步骤关键在于将 RoBERTa 产生的高质量词向量与通过人工标注获得的关键词信息相结合，从而形成具有丰富语义的表示。这样的结合有助于模型更准确地识别和理解问题的关键要素。

此外，模型还整合了哈工大 LTP 工具提取的词性和句法依存度信息和 Word2vec 五笔字型信息，这些信息同样被转换为向量形式，并与先前的词向量结合。这一步骤使得模型能够从词汇层面扩展到句法结构层面，为生成结构合理的问题提供了必要的信息。

最终，经过这些阶段的处理和信息叠加，所得数据被输入到 UniLM 模型中。UniLM 模型利用前面各阶段的输出进行综合分析和处理，最终完成问题的生成。这一流程确保了从文本分析到问题生成各个环节的顺畅衔接和信息的有效传递。

模型的整体架构如图 1 所示，清晰地展示了从预处理到问题生成的完整流程。

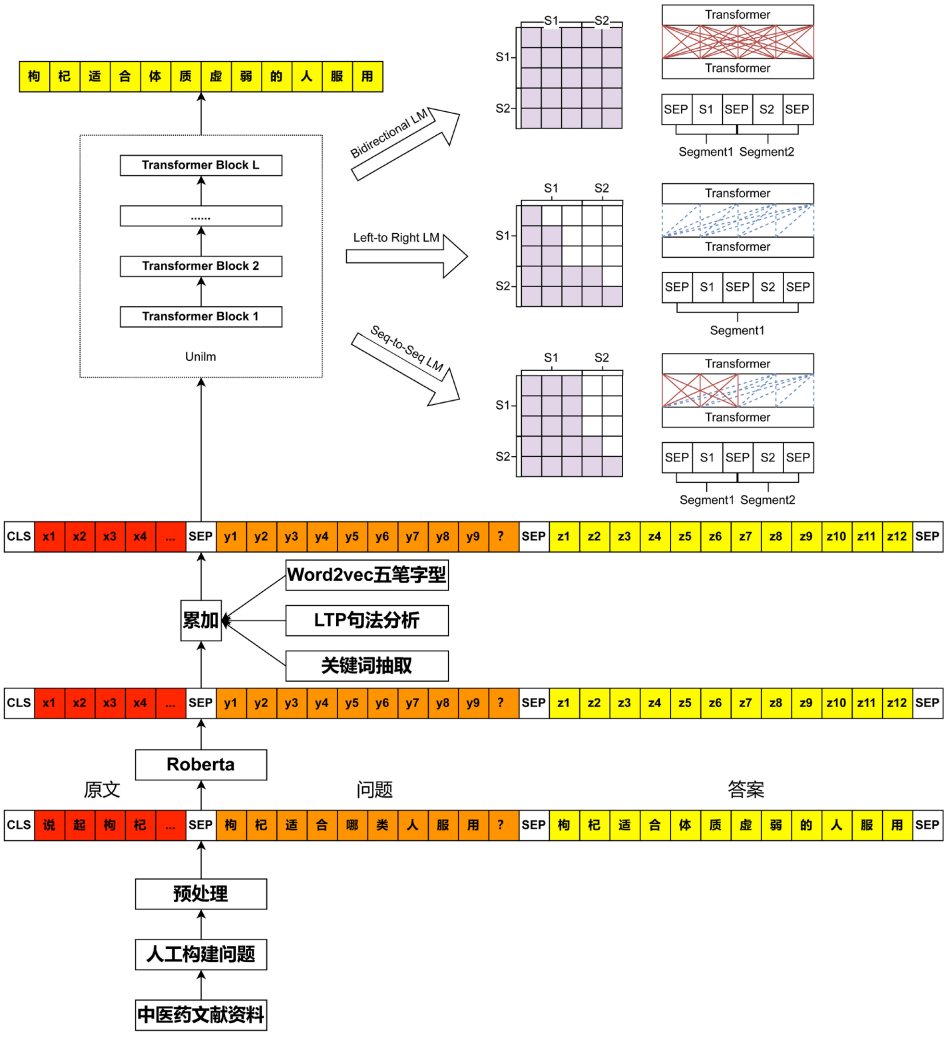

Figure 1. Overall architecture diagram
图 1. 整体架构图

3.1. RoBERTa

在中医药问题生成模型的开发中，RoBERTa 模型的应用阶段扮演着核心角色。RoBERTa (Robustly Optimized BERT Pretraining Approach)是一种基于 BERT (Bidirectional Encoder Representations from

Transformers)架构的进阶模型，旨在提供更加深入和精确的语言理解能力。它通过对大量文本数据进行深度学习，能够有效地捕获和理解语言中的复杂模式。

在处理流程中，RoBERTa 模型接收预处理后按照特定格式组织的文本数据。这些数据被格式化为“[CLS]原文[SEP]问题[SEP]答案[SEP]”的形式，其中“[CLS]”用于标记序列的开始，而“[SEP]”则用于区分原文、问题和答案的不同部分。这种结构的设计关键在于使模型能够清晰地识别和分析原文、问题和答案之间的语义关系。

在 RoBERTa 模型的处理过程中，它不仅解析原文的内容和上下文，还分析问题的结构和答案的相关性。模型通过深入理解文本中的每个词汇以及它们之间的相互关系，提取出有助于问题生成的关键信息。RoBERTa 的强大之处在于它能够理解文本的细微差异和隐含的语义，这对于处理中医药领域中常见的复杂和专业化文本尤为重要。

3.2. Word2vec

在中医药问题生成模型中融合 Word2vec 和五笔字型的训练方法，首先需要理解 Word2vec 的基本原理以及如何通过它来捕捉词语的语义和结构特征。

Word2vec 是一种有效的词嵌入方法，用于将词语转化为向量形式。它基于以下假设：在相似的上下文中出现的词语在语义上是相近的。Word2vec 通过训练神经网络模型学习每个词语的向量表示，使得语义相近的词语在多维空间中的向量距离更近。

Word2vec 主要有两种模型架构：

连续词袋(CBOW)：预测目标词基于其上下文。其基本形式的数学表示为：

$$\hat{v} = \frac{1}{C}\sum_{i=1}^{C} v\left(w_i\right)$$

其中，$\hat{v}$是上下文词的向量的平均，C 是上下文词的数量，$v\left(w_i\right)$ 是上下文中第 i 个词的向量表示。

跳跃式 Gram (Skip-Gram)：通过目标词来预测其上下文。其基本形式的数学表示为：

$$\text{maximize}\sum_{t=1}^{T}\sum_{\substack{-C\le j\le C,\\ j\neq 0}} \log p\left(w_{t+j}\middle|w_t\right)$$

其中，T 是文本中的词汇总数，C 是上下文窗口大小，$w_{(t+j)}$和 w_t 分别是目标词和上下文中的词。

在中医药领域的应用中，结合 Word2vec 和五笔字型意味着不仅捕捉汉字的语义信息，还要捕捉其基于五笔输入法的结构特征。五笔字型通过汉字结构进行编码，反映了汉字的形状和构成。将这种结构信息融合到 Word2vec 模型中，可以使得模型不仅理解词汇的语义含义，还能感知其结构特征。

例如，对于中医药术语“气虚”，其五笔编码提供了关于汉字“气”和“虚”结构的信息。将这些结构信息与 Word2vec 生成的语义向量结合，可以形成一个更全面的词表示，它不仅包含了“气虚”的语义含义，还包括了汉字的结构特性。这种结合可以用以下形式简单表示：

$$v_{\text{combined}} = \text{Word2vec}\left(\text{term}\right) + v_{\text{wubi}}\left(\text{term}\right)$$

其中，v_{combined}是融合了语义和结构信息的词向量，Word2vec (term)是通过 Word2vec 获得的词语的语义向量，而 $v_{\text{wubi}}\left(\text{term}\right)$是基于五笔字型的结构向量。

通过这种方式，中医药问题生成模型能够在生成问题时，更准确地涵盖和反映中医药专业知识的细致差别，同时保留了语义的准确性和表达的自然性。这种方法特别适合处理中医药文本这类需要捕捉细腻语义和复杂结构的领域。

3.3. 关键词抽取

在中医药问题生成模型的构建中，关键词提取与处理阶段对模型性能的影响尤为显著。这一阶段涉及到繁重的人工标注工作，其中专业人员需对中医药文献进行细致阅读，从而识别出文本中的主要和次要关键词。这些关键词是理解和回答问题的基石，因此它们的准确识别对于训练有效的问题生成模型至关重要。为了将这些关键词有效地融合到模型的词向量中，我们采用了一种二值化的方法。这个过程中，每个单词都会被赋予一个二值化标签，关键词标记为 1，非关键词标记为 0。通过这种标记，模型在处理文本时能够明确区分哪些词是关键词，从而在生成问题时给予更多关注。例如，在分析一个句子如“葡萄干有助于缓解咳嗽”，如果“葡萄干”和“咳嗽”被标注为关键词，那么这两个词在二值化过程中会被分别赋值为 1。模型在生成问题时，会着重考虑这些标记为 1 的词，因为它们携带着回答问题所需的关键信息。这种二值化方法的应用不仅简化了模型对文本的处理过程，而且提升了模型在辨识和利用关键信息上的能力。在中医药领域，这一点尤为重要，因为准确识别和理解文献中的专业术语和关键症状对于生成相关和准确的问题至关重要。总体来说，手动标注和二值化方法的结合提供了一种简洁有效的手段，通过显著地减少非关键词对模型训练的干扰，确保了问题生成模型能够集中资源处理那些对理解问题最为关键的信息。

3.4. 句法结构

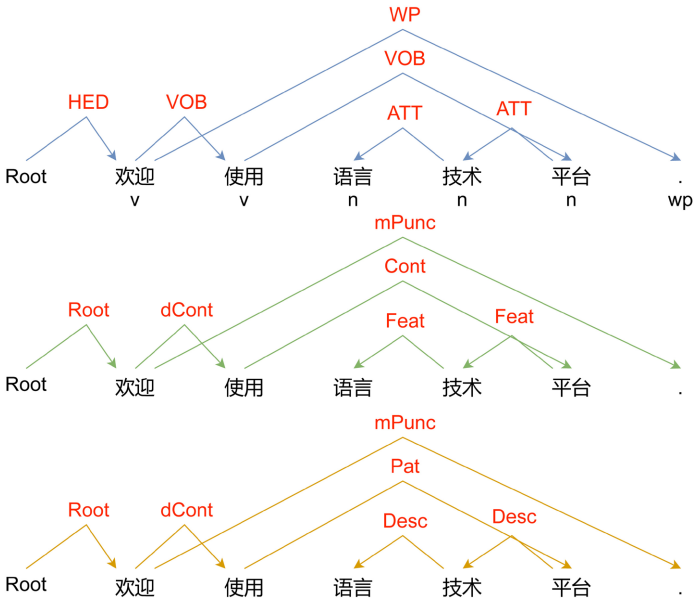

Figure 2. Syntax structure diagram
图 2. 句法结构图

哈工大 LTP (Language Technology Platform)是一个集成了多种语言处理功能的工具，包括但不限于词性标注、命名实体识别、句法依存分析等。在您的研究中，这些功能被用于增强模型对文本的语法和语义分析能力。

词性标注是识别文本中每个词的词性(如名词、动词等)。这一过程对于理解文本的语法结构至关重要。例如，在生成针对中医药领域的问题时，正确区分名词(如草药名称)和动词(如治疗方法)有助于提高问题的准确性和相关性。本模型通过 LTP 工具标记出输入文本中每个词的词性信息，建立了一个词性表。通过标注的词性信息，对输入文段进行编号。

句法依存分析揭示文本中词与词之间的依存关系。本模型通过简历依存关系表来对，词与词的依存关系。如图 2 所示。

3.5. 特征融合方式

在构建针对中医药领域的问题生成模型时，采用特征向量相加的融合方法，能够有效地整合多源信息并增强模型的综合性能。例如，中医药文献中充满了复杂的概念和术语，如“气虚”、“血瘀”等，这些术语背后承载着深厚的理论和实践内涵。在处理这些术语时，Word2vec 向量可以捕获词汇的语义空间位置，反映出“气虚”和“血瘀”在中医理论中的语义联系和差异。而哈工大 LTP 的句法分析则能揭示这些术语在句子中的作用，比如“气虚”可能是症状描述的主体，而“血瘀”可能是导致某些病症的因素。同时，人工标注的关键词信息能够指出问题生成中的重点，确保模型在提出相关问题时，能够准确地聚焦于核心概念。

将这些特征向量相加融合，模型不仅能够以更低的复杂度处理这些信息，还能在不同的特征之间建立联系，实现更为丰富的语义表示。这种融合不仅提升了模型对中医药专有术语的理解能力，还保证了在生成问题时，语言的自然性和专业性得以兼顾。例如，在生成关于“如何治疗气虚导致的头晕？”这样的问题时，模型能够综合利用“气虚”在语义上的含义、其在句法结构中的作用，以及关键词的标注信息，生成一个既符合中医药理论，又表达自然的问题。

此外，这种特征融合方式具有增强模型泛化能力的优点，使其在面对多样化的中医药文本输入时，能够稳定输出高质量的问题。在实际应用中，这意味着无论面对古籍引文、现代研究论文，还是临床案例记录，模型都能够准确捕捉并反映出中医药的专业知识和语言特点，进而生成符合实际应用需求的问题。这对于提升中医药智能问答系统的实用性和用户体验是极其有益的，因为它直接关系到系统能否准确理解用户的咨询内容，并提供专业可靠的回答。通过这样的特征融合，模型能够更好地服务于中医药领域的专业人士和普通用户，帮助他们更深入地探索和理解中医药的深奥知识。

3.6. UniLM

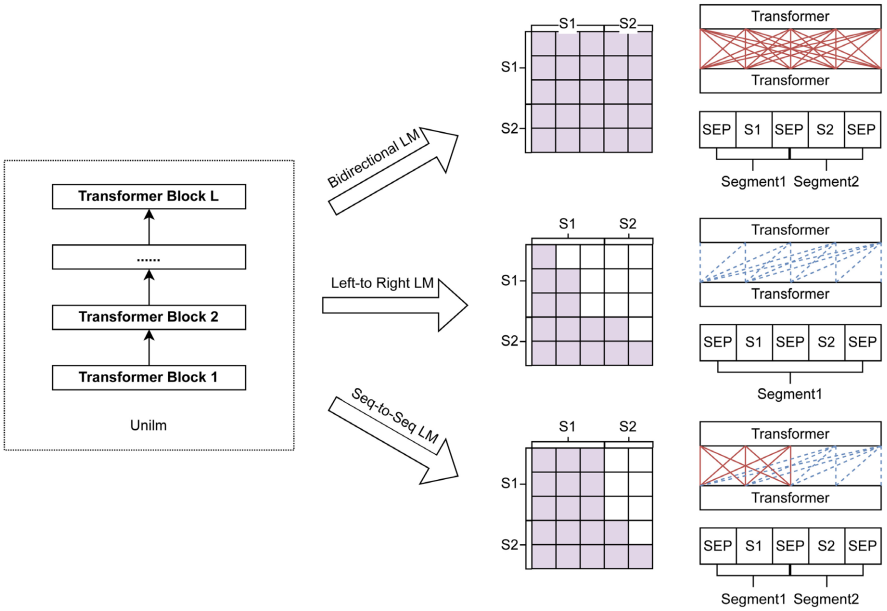

Figure 3. UniLm architecture diagram
图 3. UniLm 架构图

自然语言处理领域中的语言模型可大致分为两类：自编码语言模型和自回归语言模型。其中，自回归语言模型则是基于前文预测后续单词的概率分布，常用于自然语言生成任务中；而自编码语言模型的主要任务是将输入语句编码为一定维度的向量，并在此基础上重构输入语句。自回归语言模型根据前面(或后面)出现的词来预测当前时刻的词，代表模型有 ELMO、GPT 等，其训练目标符合生成式任务的生成过程而对生成任务(NLG)友好；缺点是只能利用单向语义而不能同时利用上下文信息以至于损失了很多文本信息。自编码语言模型通过上下文信息来预测当前被 MASK 的词，代表有 BERT、Word2Vec 等，其优点是能够很好的编码上下文语义信息，在自然语言理解(NLU)相关的下游任务上表现突出；但相对应的生成式问题的支持能力较差，而且由于训练中采用了 MASK 标记，导致预训练与微调阶段不一致。为了在同一模型中结合两种语言模型的优点，UniLM 使用三种不同的任务对模型进行训练，从而使得模型可以用于 NLG，同时在 NLU 任务获得和 BERT 一样的效果，如图 3 所示。

如图 3 所示，UniLM 通过联合训练三种语言模型(包括：双向语言模型，单向语言模型和序列到序列语言模型)，能够在单个模型中实现多种不同类型的语言生成任务，如生成式阅读理解、问答、文本摘要、机器翻译等。此外，该模型还采用了单词层面和句子层面的交替训练方式，有助于提高模型对输入序列的理解能力和生成能力。在许多自然语言处理任务上，UniLM 模型取得了领先的性能。

本章利用 UniLM 语言模型完成对槽位上下文的向量化表示。首先将被匹配句子表示为 $D=\left\{d_1,d_2,\cdots,d_j\right\}$，其中 j 为被匹配句子的长度，槽位的上一句表示为 $S^1=\left\{s_1^1,s_2^1,\cdots,s_m^1\right\}$，其中 m 为句子长度，槽位的下一句表示为 $S^2=\left\{s_1^2,s_2^2,\cdots,s_n^2\right\}$，其中 n 为句子长度。

将被匹配句子和上下句信息拼为 $\text{Tokens}=\left\{[\text{CLS}]s_1^1,s_2^1,\cdots,s_m^1,[\text{SEP}],d_1,d_2,\cdots,d_m,[\text{SEP}],s_1^2,s_2^2,\cdots,s_n^2,[\text{SEP}]\right\}$ 的字符序列形式作为 BERT 的输入，其长度为 $m+n+j+4$，将其输入至 UniLM 模型，得到标题和段落的向量化表示。经过 UniLM 输出的向量表示融合了 Token Embedding、Segment Embedding 和 Position Embedding 三部分内容，因此包含了尽可能多的文本上下文信息。其中 Position Embedding 可以对句子的前后关系做出约束，结合语义信息经过模型训练可以学习到被匹配文本与槽位上下句的适配度，提高槽位填充的可读性。

4. 实验结果和分析

4.1. 实验数据集

由于问题生成领域的主要研究方向为英文数据集合，且对于中医药领域的数据集相对比较少，本次通过从中国中医网和一些中医药典籍网站中通过爬虫技术进行爬取数据。在关注中医药领域内的常见及复杂病症，例如“肝病”和“风湿”等方面，本文通过开发一个精确的网络爬虫程序，自动化地从目标网站中抓取相关信息。爬虫的操作过程包括发送请求至目标网站，接收网站的响应，以及解析得到的网页内容。

数据清洗

原始爬取的数据中包含大量杂乱无章和不规范的信息，例如不当的标点使用、中英文字符混杂等问题。这些问题需要通过精细的数据清洗和预处理工作来解决，尤其是在标点符号和字符格式的统一方面。此外，爬取过程中不可避免地会引入无关的 HTML 标签、换行符、多余的空格及其他特殊符号，这些都需要借助正则表达式等技术手段进行有效清除。最终，通过上述处理步骤，构建出一个以(原文，答案，问题)三元组为核心的数据集，如图 4 所示。每个三元组包含三个部分：原文是从中医药相关网站上爬取的原始文本，答案是从原文中提取并保留的关键信息，而问题则是根据原文和答案生成的相关问题。这种三元组数据集不仅提升了数据的质量，还为中医药相关研究提供了规范化的数据基础，有助于后续的

自动化信息抽取、问答系统以及知识图谱的构建。将爬取的三元组数据集按比例划分为训练集、测试集和验证集，其中训练集占 80%，测试集和验证集各占 10%。

```
{
  "src_text": " 胆石症的治疗应区别不同情况分别处理，无症状胆囊结石可不作治疗。",
  "answer_text": "无症状胆囊结",
  "tgt_text": "什么类型的胆囊结石可不作治疗?"
}
```

Figure 4. Data triplet format
图 4. 数据三元组格式

4.2. 实验设置和评价指标

实验设置

本实验构建于 Cent OS 7.9 操作系统之上，采用 Python 作为主要的编程语言。表 1 详细介绍了系统开发过程中所涉及的关键软件与硬件环境配置：

Table 1. Software and hardware parameters
表 1. 软硬件参数

名称	型号
中央处理器	12th Gen Intel (R) Core (TM) i5-12400
内存	金士顿 32 G
显卡	RTX 3060Ti
磁盘	金士顿 500 G
系统	Cent OS 7.9
框架	Pytorch2.1
编程语言	Python3.8.1

本文采用了 MFFQG 模型，利用 RoBERTa 结合主副关键词、五笔信息和句法依存度信息进行融合。RoBERTa 作为词嵌入层，由 12 层网络构成，每个词汇被表示为 768 维的向量。这些由多种特征生成的 768 维向量进行了归一化处理。考虑到输入句子长度的统一性，本实验将其设定为 128，并将批处理大小设置为 32，进行了 30 个训练周期。为了降低过拟合的风险，设定了 0.25 的 Dropout 比率。此外，自注意力机制配置了 12 个注意力头。相关配置详情参见表 2 所示。

Table 2. Model parameters
表 2. 模型参数

参数名	参数值
序列最大长度	128
批次尺寸	32
训练轮数	30
隐层数	12
学习率	4 e−7
Dropout	0.25
注意力头数	12

4.3. 实验结果

4.3.1. 基准模型

1) Seq2seq 模型：是一种经典的用于序列到序列任务的深度学习架构，通过编码器－解码器结构实现输入序列到目标序列的转换。

2) NQG 模型：是一种基于神经网络的自动问题生成系统。它利用编码器－解码器架构，

通过学习从文本中生成自然语言问题。NQG 模型添加了注意力机制和词嵌入技术，以提高生成问题的质量和相关性。

3) BERT + UniLm 模型：该模型基于序列到序列架构，将编码器替换为双向 Transformer 编码模块，并使用 UniLM 作为解码器。编码器的参数通过 BERT 模型进行初始化，而解码器则从头开始训练。

4.3.2. 实验结果

从表 3 可以看出，与传统的 Seq2Seq 模型相比，NQG 模型通过添加注意力机制和词嵌入技术，使模型在生成问题时能够更好地关注文本的重点。在 Rouge-1、Rouge-2 和 Rouge-L 指标上，NQG 模型分别提升了 8.28%、3.94%和 9.28%。此外，BERT + UniLM 模型引入了预训练模型的优势，进一步提升了效果。本文采用的 MFFQG 模型，通过融合中医药领域的多种特征，展示了卓越的性能。在 Rouge-1、Rouge-2 和 Rouge-L 这三项指标中，MFFQG 模型均取得了最好的结果。具体而言，相较于 BERT + UniLM 模型，MFFQG 模型在 Rouge-1、Rouge-2 和 Rouge-L 上分别提升了 6.36%、5.12%和 7.14%。这些提升不仅表明 MFFQG 模型在中医药问题生成任务中的有效性，还突显了多特征融合策略在提高生成问题质量方面的重要作用。尤其是在中医药知识的处理上，MFFQG 模型通过结合主副关键词、五笔信息和句法依存度信息，显著提升了生成问题的准确性和相关性。这一改进使得模型能够更好地理解和利用中医药领域的专业知识，从而生成更加贴近实际应用的问题。此外，模型的训练和验证均在 GPU 上进行，确保了高效的计算性能和可靠性。

综上所述，MFFQG 模型的实验结果表明，通过融合多种特征，可以显著提升中医药领域问题生成的质量。这一发现不仅为中医药自动问答系统的发展提供了新的思路，也为其他领域的自然语言处理任务提供了参考和借鉴。通过进一步优化和改进，MFFQG 模型有望在更多实际应用中发挥更大的作用。

Table 3. Comparison of experimental results
表 3. 对比实验结果

Model	Rouge-1	Rouge-2	Rouge-L
Seq2seq	40.95%	23.72%	39.05%
NQG	49.23%	27.66%	48.34%
BERT + UniLm	57.30%	29.45%	55.91%
MFFQG	64.93%	34.57%	63.05%

5. 结语

本文面向中医药领域，提出一种基于多特征融合的中医药问题生成模型 MFFQG，分别使用主副关键词标注和句法依存度分析方法以改善现有方法中领域关键词信息缺失和生成表达不规范问题，有效提升了中医药领域的问题生成质量并通过对比实验验证了本文方法的有效性。

然而，本文在特征融合阶段只是采用了简单的向量归一化的方式，此处仍有提高空间，未来会研究如何在融合过程中引入深度学习模型，使模型可以自主学习为不同特征赋予不同权重以提升模型效果；同时本文在中医药数据上取得了较好的结果，未来可以研究如何推广到其他领域。

参考文献

[1] Dhole, K. and Manning, C.D. (2020) Syn-QG: Syntactic and Shallow Semantic Rules for Question Generation. *Proceedings of the* 58*th Annual Meeting of the Association for Computational Linguistics*, Online, 5-10 July 2020, 752-765. https://doi.org/10.18653/v1/2020.acl-main.69

[2] Labutov, I., Basu, S. and Vanderwende, L. (2015) Deep Questions without Deep Understanding. *Proceedings of the* 53*rd Annual Meeting of the Association for Computational Linguistics and the* 7*th International Joint Conference on Natural Language Processing* (*Volume* 1: *Long Papers*), Beijing, 26-31 July 2015, 889-898. https://doi.org/10.3115/v1/p15-1086

[3] Mazidi, K. and Nielsen, R.D. (2014) Linguistic Considerations in Automatic Question Generation. *Proceedings of the* 52*nd Annual Meeting of the Association for Computational Linguistics* (*Volume* 2: *Short Papers*), Baltimore, 23-25 June 2014, 321-326. https://doi.org/10.3115/v1/p14-2053

[4] Heilman, M. and Smith, N.A. (2010) Good Question! Statistical Ranking for Question Generation. *Proceedings of Human Language Technologies*: *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, Los Angeles, 2-4 June 2010, 609-617.

[5] Chali, Y. and Hasana, A. (2012) Towards Automatic Topical Question Generation[C]//*Proceedings of the* 24*th International Conference on Computational Linguistics*, Mumbai, 2012: 475-492.

[6] Duan, J.-Y., Xu, L.-S., Liu, J., *et al*. (2022) Question Generation Based on Sememe Knowledge and Bidirectional Attention Flow. *Data Analysis and Knowledge Discovery*, **6**, 44-53.

[7] Zeng, H., Zhi, Z., Liu, J. and Wei, B. (2021) Improving Paragraph-Level Question Generation with Extended Answer Network and Uncertainty-Aware Beam Search. *Information Sciences*, **571**, 50-64. https://doi.org/10.1016/j.ins.2021.04.026

[8] Hu, Y. and Zhou, G.-Y. (2022) Question Generation from Knowledge Base with Graph Transformer. *Journal of Chinese Information Processing*, **36**, 111-120.

[9] Fei, Z., Zhang, Q. and Zhou, Y. (2021) Iterative GNN-Based Decoder for Question Generation. *Proceedings of the* 2021 *Conference on Empirical Methods in Natural Language Processing*, Punta Cana, 7-11 November 2021, 2573-2582. https://doi.org/10.18653/v1/2021.emnlp-main.201

[10] Serban, I.V., Garcia-Durán, A., Gulcehre, C., *et al*. (2018) Generating Factoid Questions with Recurrent Neural Networks: The 30M Factoid Question-Answer Corpus. *Proceedings of Annual Meeting of the Association for Computational Linguistics*, Berlin, 7-12 August 2016, 588-598.

[11] Ma, H., Wang, J., Lin, H. and Xu, B. (2022) Graph Augmented Sequence-to-Sequence Model for Neural Question Generation. *Applied Intelligence*, **53**, 14628-14644. https://doi.org/10.1007/s10489-022-04260-2

[12] Li, Y.-F., Ye, D.-Y. and Chen, Y.-Z. (2023) Knowledge Enhanced Biograph Interaction Neural Network for Question Generation. *Journal of Chinese Computer Systems*.