# 面向中文多方对话的机器阅读理解研究

**蒋玉茹[1], 李宇[1*], 那婷婷[1], 张仰森[1]**
1.北京信息科技大学计算机学院，北京
{jiangyuru,2021020666}@bistu.edu.cn

## 摘要

在机器阅读理解领域，处理和分析多方对话一直是一项具有挑战性的研究任务。鉴于中文语境下相关数据资源的缺乏，本研究构建了DialogueMRC数据集，旨在促进该领域的研究进展。DialogueMRC数据集作为首个面向中文多方对话的机器阅读理解数据集，包含705个多方对话实例，涵盖24451个话语单元以及8305个问答对。区别于以往的MRC数据集，DialogueMRC数据集强调深入理解动态的对话过程，对模型应对多方对话中的复杂性及篇章解析能力提出了更高的要求。为应对中文多方对话机器阅读理解的挑战，本研究提出了融合篇章结构感知能力的中文多方对话问答模型（Discourse Structure-aware QA Model for Chinese Multi-party Dialogue，DSQA-CMD），该模型融合了问答和篇章解析任务，以提升对话上下文的理解能力。实验结果表明，相较于典型的基于微调的预训练语言模型，DSQA-CMD模型表现出明显优势，对比基于Longformer的方法，DSQA-CMD模型在MRC任务的F1和EM评价指标上分别提升了5.4%和10.0%；与当前主流的大型语言模型相比，本模型也展现了更佳的性能，表明了本文所提出方法的有效性。

**关键词：** 中文多方对话；机器阅读理解；篇章解析

# Research on Machine Reading Comprehension for Chinese Multi-party Dialogues

**Yuru Jiang[1], Yu Li[1*], Tingting Na[1], Yangsen Zhang[1]**
1. School of Computer Science,
Beijing Information Science and Technology University
{jiangyuru,2021020666}@bistu.edu.cn

## Abstract

In the field of machine reading comprehension (MRC), processing and analyzing multi-party dialogues remains a challenging research task. Given the lack of related data resources in the Chinese context, this study has developed the DialogueMRC dataset to facilitate progress in this area. As the first MRC dataset tailored to Chinese multi-party dialogues, the DialogueMRC dataset comprises 705 multi-party dialogue instances, encompassing 24,451 utterance units and 8,305 question-answer pairs. Unlike previous MRC datasets, the DialogueMRC dataset emphasizes a deep understanding of dynamic conversational processes, imposing higher demands on models in handling the

* 通讯作者

**基金项目**：北京市自然科学基金项目（4242019），题目：智慧养老服务中对话机器人关键技术研究

complexities of multi-party dialogues and discourse parsing capabilities. To address the challenges of Chinese multi-party dialogue MRC, this study introduces a discourse structure-aware question answering model for Chinese multi-party dialogue (DSQA-CMD). This model integrates question-answering and discourse parsing tasks to enhance the comprehension of dialogue context. Experimental results show that the DSQA-CMD model exhibits significant advantages over typical fine-tuned pretrained language models. Compared to methods based on Longformer, the DSQA-CMD model shows improvements of 5.4% and 10.0% in F1 and EM metrics, respectively. Moreover, the model outperforms current mainstream large language models, demonstrating the effectiveness of the proposed approach.

**Keywords:** Chinese Multi-party Dialogue , Machine Reading Comprehension , Discourse Parsing

## 1 引言

随着人工智能和自然语言处理技术的飞速发展，机器阅读理解（Machine Reading Comprehension，MRC）在智能信息获取、问答系统等领域中扮演着关键角色。尤其是将对话文本作为一种日益重要的信息载体，其MRC任务远比格式化良好的正式文档更加复杂且具有挑战性。对话文本MRC面临的主要挑战体现在模型需要深入理解动态的对话过程、处理信息的多维度特性以及准确识别对话意图上。在中文多方对话领域，由于其固有的交叉依赖特性，这些挑战变得尤为明显。目前，相关研究大多集中在英语资源上，而面向中文资源的研究则明显不足。鉴于此，本研究提出了DialogueMRC数据集，这是一个专为中文多方对话设计的片段抽取式MRC数据集。该数据集旨在促进MRC模型对中文多方对话文本结构的理解，以及提升问题答案提取的准确度，图1展示了DialogueMRC数据集经过标注的一个问答样例，其中不仅包含可回答的问题，还包含无法回答的问题。对于可回答的问题，每个问题的答案都对应上下文中的一段连续文本片段。此外，图中还展示了之前研究者所完成的篇章解析（Discourse Parsing，DP）标注成果(Jiang et al., 2023)。基于此，本研究提出了创新的多任务学习模型，旨在更有效地解析和处理复杂的对话结构以提升MRC模型的性能。本文研究的主要贡献如下：

1)提出了多阶段MRC数据标注流程，并借此构建了DialogueMRC数据集[0]，这是首个面向中文多方对话的大规模MRC数据集，包含8305个问答对，旨在提供一个全面且真实的中文对话理解环境。

2)提出了融合篇章结构感知能力的中文多方对话问答模型（DSQA-CMD模型），通过联合对话问答任务和篇章解析任务来增强模型对于复杂对话结构的理解能力。

3)实验结果表明，相较于典型的基于微调的预训练语言模型，DSQA-CMD模型表现出明显优势；与当前主流的大型语言模型相比，本模型也展现了更佳的性能，证明了本文所提出方法的有效性。

## 2 相关工作

### 2.1 机器阅读理解

机器阅读理解任务旨在训练机器阅读并理解文章内容，以便对给定问题从相关文本中找到答案(张开颜et al., 2021)。Hermann et al. (2015)于2015年构建了大规模数据集CNN/DailyMail，并提出了基于注意力的LSTM模型——THE ATTENTIVE READER，这标志着机器阅读理解技术进入神经网络时代。虽然CNN/DailyMail数据集的规模足以满足深度学习模型的训练需求，但由于其完形填空的问题形式并不符合人类自然语言的使用方式。为了突破这一限制，Rajpurkar et al. (2016)在2016年推出了SQuAD数据集，这是首个大规模机器阅读理解数据集，每个问题的答案都直接来自相关段落文本。而在中文机器阅读理解领域，由于数据资源的匮乏发展相对滞后。最早的重要贡献包括Cui et al. (2016)手工构建的首个中文阅读理解数据集，涵盖14个领域的121篇文章。此外，哈工大讯飞联合实验室（HPL-RC）

[0]https://github.com/LIyu810/DialogueMRC

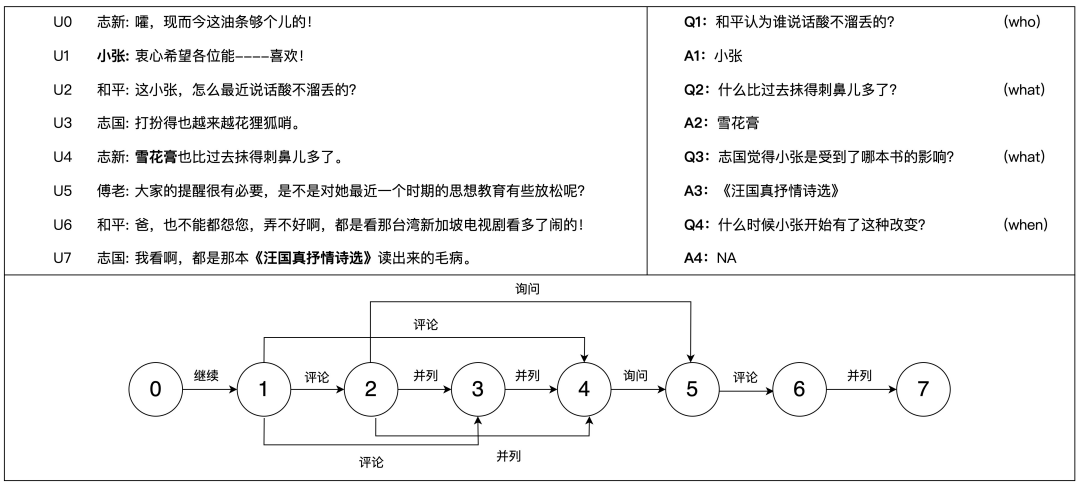



图 1. DialogueMRC数据集样例

于2019年使用人民日报和儿童读物作为语料发布了大规模中文机器阅读理解数据集，Duan et al. (2019)还基于该数据集开发了效果良好的基于注意力机制的深度阅读理解模型。

### 2.2 篇章解析

篇章解析是自然语言处理领域的基础任务，旨在分析文本中的语篇结构和语义关系，这一任务涵盖了诸如评论、致谢及问答等多种关系类型。目前，篇章解析技术主要基于宾州树库（PDTB）和修辞结构理论树库（RST-DT）等数据集进行研究。PDTB(Webber et al., 2019)和RST-DT(Hou et al., 2020)的发布显著推动了语篇关系及篇章语义分析技术的发展，这两个语料库的原始数据均来源于华尔街日报。然而，随着研究的深入，学界对于偏口语化的多方对话篇章解析表现出增加的兴趣。在这一领域，STAC数据集(Afantenos et al., 2015)成为主流语料库之一。该数据集由Afantenos等人在2016年发布，基于游戏《The Settlers of Catan》的在线版本进行构建。此游戏涉及多玩家的合作与竞争交流，为篇章分析提供了丰富的口语化文本资源。

### 2.3 融合篇章解析的机器阅读理解研究

近年来，将篇章解析融入机器阅读理解的研究受到了广泛关注。Molweni数据集(Li et al., 2020)作为该研究领域的重要里程碑，包含10000个多方对话，88303个篇章单元和30066个问答对，此外，研究者还标注了78245个篇章关系。Molweni要求模型不仅需要理解每个发言的表面文字信息，还要分析不同篇章单元之间的篇章关系，从而实现对上下文更深入的理解。Molweni数据集来源于Ubuntu操作系统的在线技术论坛，其相较于日常生活场景的发言内容，在交流形式上存在显著差异。通常，日常对话中的发言内容更加普遍，一次只能由一人发言，而且对话是即时进行的，不具有滞后性。这些特征赋予了基于日常场景的多方对话数据集其独特的价值，为了更好的挖掘这些日常生活中对话的特点，Jiang et al. (2023)提出了DialogueDSA的数据集。该数据集精心挑选并标注了包括因果、转折、并列等在内的16种篇章关系类型。相关研究表明，篇章结构的理解对于机器准确回答问题至关重要，特别是在需要长距离推理和把握复杂对话流的情境下。此外，针对Molweni的实验结果也揭示了DP具有提高MRC任务性能的潜力(He et al., 2021)。在本文中，将进一步探讨融合篇章解析的机器阅读理解技术，以及该技术在处理多方对话文本中的应用效果和潜在的改进方向。

## 3 DialogueMRC数据集的构建与分析

本文所构建的数据集（以下简称为DialogueMRC）语料来自1994年上映的120集情景喜剧《我爱我家》的剧本文本。该剧情景包含117位出场角色，其中8位为主要角色。下述列出了选择《我爱我家》剧本文本作为标注基础语料库的三个原因。

1)该剧内容与日常生活紧密相连，涵盖了广泛的话题，如家庭关系、职场矛盾、邻里交往等。这种丰富多彩的生活场景为机器阅读理解研究提供了一个接地气的、现实生活中的应用背

景，非常适合探索和解决日常生活场景下的机器阅读理解问题。

2)该剧本涵盖了117位角色，每个角色都有其独特的性格、语言风格和情感表达。这种角色多样性对数据集的标注提出了挑战，同时也提供了丰富的标注维度，这对于构建一个能够理解复杂人物关系和情感交互的机器阅读理解模型至关重要。

3)目前已有研究基于《我爱我家》的剧本文本标注了关系抽取(Jiang et al., 2022)、篇章解析(Jiang et al., 2023)等数据集，并在此基础上完成了相关实验，提出了CSE-DPM(Jiang et al., 2023)等模型，这些工作为本文的研究提供了坚实的基础。

在问答设计中，DialogueMRC特别考虑到多方对话的复杂性，其中对话参与者可能多于两人。这一设定增加了问题设计的难度，因为需要准确识别和理解每个参与者的发言及其在对话中的角色和立场。此外，多方对话的动态性要求数据标注过程中必须考虑对话流的连贯性和上下文变化，这对于提高模型在真实场景中的应用能力至关重要。通过这种针对性的设计，DialogueMRC旨在提升模型对复杂对话结构的理解力，从而更好地服务于多参与者交互的实际应用场景。

DialogueMRC数据集的场景选择继承自CRECIL和DialogueDSA所采用的同一批705个对话场景，这种继承性不仅保证了数据的一致性，也为后续的多任务联合训练研究提供了便利。其中每个场景都被视作一个独立的对话单元，以便于进行更精确的数据处理与分析。DialogueMRC可以被看作是片段抽取任务，即在给定的对话上下文中提出问题，并要求模型在对话内容中找到答案片段。为了确保DialogueMRC数据集在精确性和可靠性方面达到最高标准，本研究提出了多阶段机器阅读理解数据标注流程，并建立了详尽的数据质量检查体系。通过这些措施，力求使DialogueMRC数据集成为机器阅读理解领域的一个可靠和有价值的资源。

### 3.1 多阶段机器阅读理解数据标注流程

在数据标注领域，人工标注因其提供高度准确且可靠的数据而长期被视为标准。但是，人工标注的缺点是成本高且耗时。随着人工智能技术的发展，使用文本生成技术进行数据标注成为了一种新兴方法，利用类似GPT-3.5的大语言模型可以高效地生成大量标注数据，提高了标注效率。但这种方法的挑战在于生成的标签可能包含噪声，特别是在特定领域或复杂任务中更为明显。鉴于以上挑战，本文提出了一种多阶段机器阅读理解数据标注流程，如图2所示，旨在结合人工标注的准确性与使用大语言模型自动标注的高效性，以达到更优的标注结果。

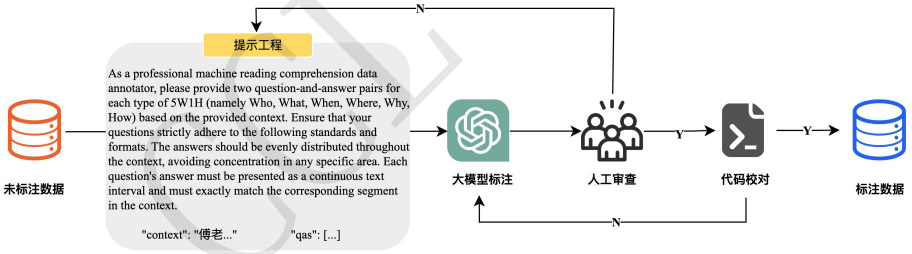



图 2. 多阶段机器阅读理解数据标注流程图

1)提示工程：流程的首要步骤是设计初始的提示。此步骤涉及输入特定的提示和对话内容，引导大型语言模型生成相应的问答对。然后，通过人工审查这些问答对，识别并解决诸如问题类型不匹配、问答对在对话中分布不均、生成的答案与上下文连续片段不匹配等问题。通过这种持续的反馈和优化过程，逐步完善提示语的设计，确立数据标注的准确提示语。

2)大语言模型标注：本研究选择目前业界前沿的大语言模型ChatGPT-4.0来完成标注工作，此举旨在提升标注效率(Gilardi et al., 2023)。借助大语言模型强大的语言理解能力，本研究高效地标注了大量的高质量问答对。该模型通过分析对话内容生成不同类型的问题，然后根据上下文内容生成准确的答案。这一步骤大幅减少了传统手工标注所需的时间和人力成本。

3)人工审查：鉴于大型语言模型可能出现的幻觉等问题，为确保标注数据的高质量，本研究采用人工审查方式。审查内容包括模型生成问答对的类型匹配、5W1H（何时、何地、何人、何物、为何、如何）问题的数量分布、有答案与无答案问题的准确性和鲁棒性等。在发现问答对存在问题时，标注人员将对提示语进行微调，并重新生成问答对，直至通过人工审查。

4)代码校对：通过编写专门的脚本，对生成的答案进行审核，确保答案是对话上下文中的连续区间，并检查答案的起始位置是否准确。若发现错误，则对错误样例重新进行大语言模型标注，直至通过代码校对。

总结而言，本文提出的多阶段机器阅读理解数据标注流程，有效地融合了手工标注的精准性与自动化标注的高效性。

### 3.2 DialogueMRC语料库数据分析

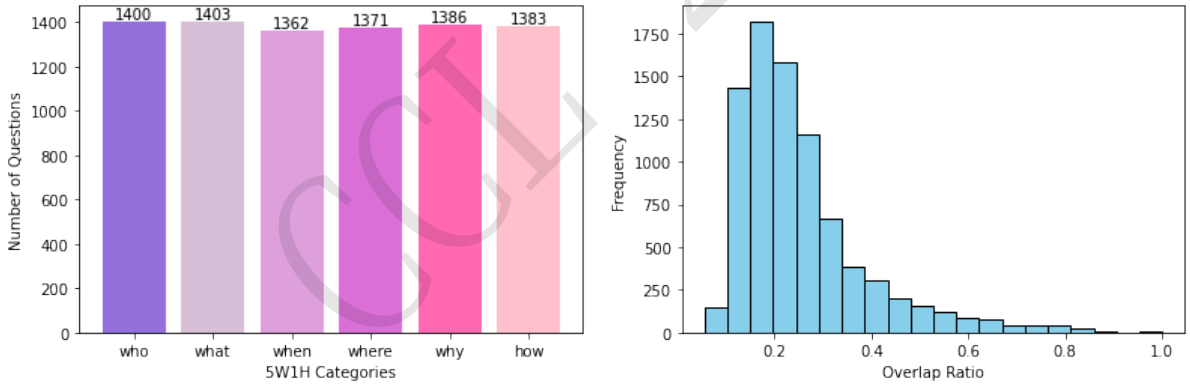
标注工作完成后，得到了一个详尽且均衡的机器阅读理解数据集，如表1所示，DialogueMRC数据集包含了总计705个对话实例，其中训练集占565个，开发集和测试集各占70个。

| | 训练集 | 验证集 | 测试集 |
|---|---|---|---|
| 对话数量 | 565 | 70 | 70 |
| 话语单元数量 | 19556 | 2576 | 2319 |
| 问题数量 | 6654 | 830 | 818 |
| 可回答问题数量 | 5526 | 662 | 689 |
| 不可回答问题数量 | 1128 | 168 | 129 |

表 1. DiologueMRC数据集概览

#### 3.2.1 5W1H问题类型分布

DialogueMRC数据集中问题类型在5W1H范式下的分布均衡，如图3(a)所示。这种均衡对于确保机器阅读理解系统的全面性和鲁棒性至关重要，有助于系统对各类问题进行平等处理，减少偏差。此外，数据集的精心设计与平衡标注有利于发展广泛适用的模型，并公平评估各类问题上的模型性能。




(a) 5W1H问题类型分布图　　(b) 问题与上下文的词汇重叠率分布图

图 3. 问题类型及词汇重叠分析示意图

#### 3.2.2 问题与上下文的词汇重叠率分布

图3(b)展示了DialogueMRC数据集中问题与其上下文之间词汇重叠率的分布情况。词汇重叠率是衡量问题词汇与其上下文共享程度的指标，这一指标反映了问题相对于对话内容的复杂性和独特性。在该数据集中，一个明显的高频区间发生在10%到30%的词汇重叠率范围内，表明相当一部分问题与其上下文共享了适度的词汇量。在更高的重叠率范围内，频率显著下降，只有2个问题的重叠率在90%至100%之间。这一急剧的下降表明与上下文几乎完全相同的问题极为罕见，这意味着大多数问题需要对对话内容有更深层次的理解才能确定正确答案。该问题分布证明了DialogueMRC语料库经过了精心设计，提供了有效训练机器阅读理解模型的数据。

第二十三届中国计算语言学大会论文集，第650页-第661页，太原，中国，2024年7月25日至28日。
卷1：主会论文

## 4 融合篇章结构感知的多方对话问答模型设计

本研究提出了融合篇章结构感知能力的中文多方对话问答模型（Discourse Structure-aware QA Model for Chinese Multi-party Dialogue，DSQA-CMD），如图4所示，问答任务通过预训练模型的输出来预测答案在文本中的位置，而篇章解析任务则通过构建依赖树捕获话语间的结构关系，这两个任务的联合训练旨在加深对多方对话的理解，从而提升MRC模型的性能。

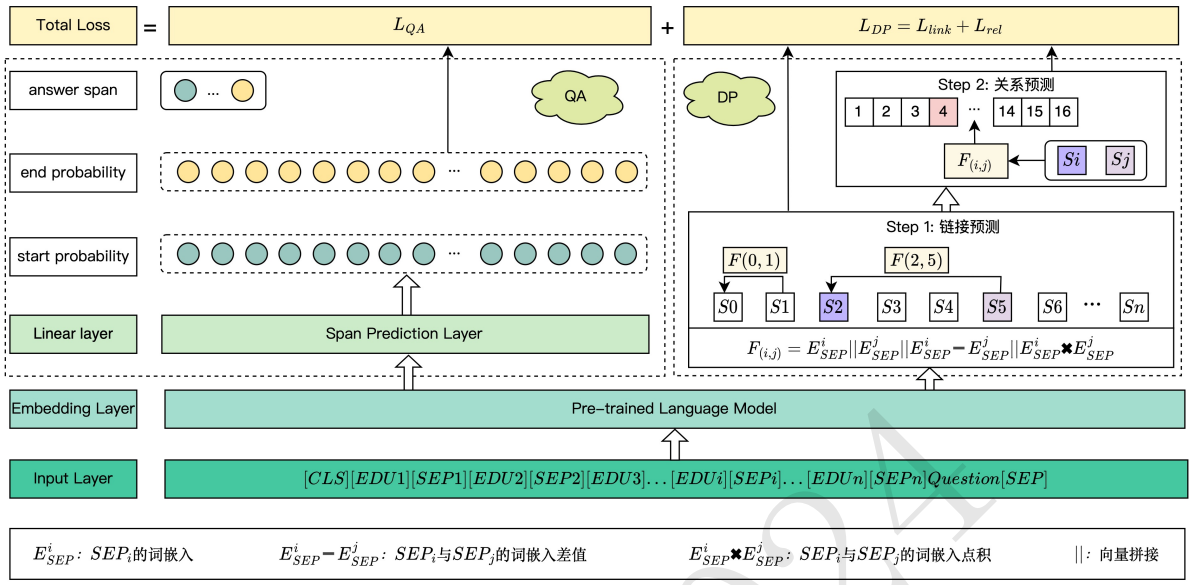



图 4. 融合篇章结构感知能力的中文多方对话问答模型图

### 4.1 输入特征表示

对话上下文与问题在预处理阶段通过预训练语言模型进行编码。为明确区分问题与对话内容以及对话中的不同篇章单元，特殊分隔符号（$[SEP]$）被插入相应位置。这样不仅有助于辨识问题与上下文的边界，还为篇章解析任务提供了话语链接预测的关键线索。同时，对话中的每个分隔符位置都被准确记录，以确保篇章解析任务中能够单独提取出每个话语对应的特征表示。最后，问题Q与对话中所有的篇章单元（$EDU$）连同特殊符号一同作为输入，得到融合了问题与上下文信息的序列特征表示S，具体表示为$S = encoder\left([CLS], [EDU1], [SEP1], \ldots, [EDUn], [SEPn], Q, [SEP]\right)$

### 4.2 问答任务

问答任务直接使用从预训练语言模型中得到的输出特征。模型通过全连接层为给定对话中的每个词元（token）预测作为答案开始和结束的概率，并据此从对话文本中提取答案片段。起始位置的预测如公式1所示，其中$P_s$表示预测的起始位置，下标$i$表示词元序列中的索引，$W_s$为用于预测起始位置的权重矩阵，$S$为从预训练语言模型中获得的文本特征。在处理无法回答的问题时，首先计算最可能的答案范围分数，并将其与无答案的预设阈值进行比较，以此来决定问题是否有确定答案。

$$P_s = argmax_i \left(softmax\left(W_s S\right)_i\right) \tag{1}$$

问答任务旨在准确预测答案在文本中的起始和结束位置。该任务被构建为一个分类问题，假定输入嵌入中共有$M$个词元，每个词元对应一个可能的起始或结束概率。利用$softmax$函数对模型输出的logits进行归一化，得到各个位置的概率分布。然后通过交叉熵损失函数计算起始和结束位置的预测损失。如公式2所示，$M$表示问答任务的标签数量，起始位置标签$y_s^m$为1表示第$m$个词元为开始位置，否则为0。同样，结束位置标签$y_e^m$为1表示第$m$个词元为结束位置，否则为0。模型预测的起始位置概率为$P_s^m$，结束位置概率为$P_e^m$。QA任务的总损失$L_{QA}$通过起始

和结束位置预测损失的均值来表示。

$$L_{QA} = -\frac{1}{2}\sum_{m=0}^{M-1}(y_s^m logP_s^m + y_e^m logP_e^m) \tag{2}$$

### 4.3 篇章解析任务

篇章解析任务需要进一步处理经过预训练语言模型得到的特征，以表征话语之间的依赖关系。特别地，篇章单元所对应的分隔符向量被选定为话语特征的代表，进一步计算出用于篇章解析的关系特征向量。这一过程结合了欧几里得距离和余弦相似度的计量，旨在更全面地捕捉话语间的细微关系。为了解析篇章单元$[EDUi]$和$[EDUj]$所对应的篇章关系，本模型构建了特征向量的联合表示$F_{i,j}$（$F_{i,j} = \left(E_{SEP}^i, E_{SEP}^j, E_{SEP}^i - E_{SEP}^j, E_{SEP}^i \times E_{SEP}^j\right)$）。其中，$E_{SEP}^i$和$E_{SEP}^j$分别对应篇章单元$[EDUi]$和$[EDUj]$的输出特征，$F_{i,j}$代表对不同表示的拼接，模型综合了分隔符对应的向量及其交互信息，从而更好地揭示篇章单元之间的深层结构关系。

篇章解析任务通过构建依赖树来表征话语之间的依赖关系。篇章解析任务分为两个部分：链接预测和关系预测。对于链接预测，模型旨在确定话语间是否存在依赖关系。对于每个话语，通过计算特征向量来预测其依赖对话中的哪个话语。当话语不依赖于其他话语时，将其指定为依赖于一个虚构的根节点。如公式3所示，$L_i$表示预测的依赖话语索引，$a$表示遍历所有可能的依赖目标话语的索引，$W_l$表示用于链接预测的权重矩阵。而在关系预测中，模型需确定话语间的确切关系类型，如公式4所示，$R_{i,j}$表示预测的第$i$个话语和第$j$个话语之间的关系类型，$k$表示所有可能的关系类型，$W_r$表示用于关系预测的权重矩阵。

$$L_i = argmax_a\left(softmax\left(W_l\left[F_{i,1}, F_{i,2}, \ldots, F_{i,t}\right]\right)_a\right) \tag{3}$$

$$R_{i,j} = argmax_k\left(softmax\left(W_r F_{i,j}\right)_k\right) \tag{4}$$

DialogueMRC数据集单一对话中最大基本篇章单元数量为$T\,(T=90)$，篇章关系类型数量为16。链接预测和关系预测被视为多类分类问题，分别具有$T+1$和16个类别，其中$T+1$中的额外一个标签代表根节点。交叉熵损失函数用于分别计算链接预测$L_{link}$和关系预测$L_{rel}$的损失，对其求和得到DP任务的总损失$L_{DP}$。

链接预测的损失函数定义如公式5所示，如果第$t1$个话语依赖于第$t2$个话语，那么$y_l^{t1,t2}$为1，否则为0。$p_l^{t1,t2}$表示模型预测第$t1$个话语依赖于第$t2$个话语的概率值。

$$L_{link} = -\frac{1}{T}\left(\sum_{t1=0}^{T-1}\sum_{t2=0,t2\neq t1}^{T-1} y_l^{t1,t2} logP_l^{t1,t2}\right) \tag{5}$$

关系预测的损失函数$L_{rel}$如公式6所示，其中，如果第$n$个样本中第$t$个话语与任何其他话语存在第$i$种类型的关系，则$y_r^{t,i}$等于1，否则等于0。同时，$p_r^{t,i}$表示模型预测第$t$个话语与任何其他话语存在第$i$种关系的概率。

$$L_{rel} = -\frac{1}{T}\left(\sum_{t=0}^{T-1}\sum_{i=0}^{15} y_r^{t,i} logP_r^{t,i}\right) \tag{6}$$

如公式7所示，DP任务的总损失$L_{DP}$是链接预测和关系预测损失的和。

$$L_{DP} = L_{link} + L_{rel} \tag{7}$$

将DP任务的总损失$L_{DP}$与QA任务的损失$L_{QA}$相加，得到DSQA-CMD模型的总损失$L$。

$$L = L_{QA} + L_{DP} \tag{8}$$

# 5 实验与分析

## 5.1 评价指标

实验采用F1分数和精确匹配（EM）来作为问答（QA）任务的评价指标。F1分数是基于精确率（P）和召回率（R）的调和平均值，公式定义如9所示。其中，精确率度量了预测为正例的样本中真实为正的比例，召回率度量了在所有真实正例中被正确预测为正的比例。

$$F1 = \frac{2 \times P \times R}{P + R} \tag{9}$$

EM（精确匹配）是在问答系统中广泛应用的评价指标，用于衡量预测答案与任何一个真实答案完全一致的比例，公式定义如公式10所示。对于篇章解析任务，本研究使用F1分数来分别评估链接预测和关系预测的性能。在关系预测中，只有当链接和关系预测都正确时，才会被视为结果正确。

$$EM = \frac{\text{完全匹配的预测数量}}{\text{总预测数量}} \tag{10}$$

## 5.2 实验设置

本研究选取两种预训练语言模型，分别为RoBERTa(Liu et al., 2019)和MacBERT(Cui et al., 2020)，作为对比实验的基础。这两种模型分别被设定为实验的基线（benchmark）和当前最优（state-of-the-art）性能指标，这两种模型均具有768维的隐藏层。在对话上下文处理中，由于数据集存在上下文长度超过512个词元的实例，为适应模型输入限制并保持上下文完整性，本研究设置最大序列长度为512，并采用文档滑动窗口为128的策略。此外，问题和答案的最大长度分别设定为64和60，以涵盖大多数潜在答案。实验中设置每个对话的最大话语数为40，模型训练采用2e-5的学习率，并设定丢弃率为0.2，以抑制过拟合现象。权重衰减设置为0.01，以进一步正则化模型训练。经过初步实验验证，所有模型均在10轮内完成训练，批次大小为18。

## 5.3 对比模型

### 5.3.1 基于微调的预训练对比模型

为了全面评价本研究提出的多任务学习模型在多方对话机器阅读理解任务上的表现，本研究设计了一系列对比实验，涉及多种基于微调的深度学习模型。本研究基于MacBERT构建了多任务学习模型（以下简称DSQA-CMD（Mac））和一个采用RoBERTa作为预训练模型的多任务模型（以下简称DSQA-CMD（Ro））。此外，为了确保评估的全面性，本研究还同其他专门针对问答和篇章解析任务的模型作为对比，包括针对长文本的Longformer机器阅读理解模型，以及专注于篇章解析的Deep sequential模型和CSE-DPM模型。这些对比实验不仅有助于揭示DSQA-CMD (Mac)在多方对话MRC任务中的相对优势，也为深入理解不同模型架构和训练策略在复杂对话环境中的适应性和效果提供了丰富的数据支持。

### 5.3.2 基于提示驱动的大语言对比模型

在自然语言处理领域，大型预训练语言模型如GPT-3.5-turbo已广泛应用于多种任务，并展现出卓越的性能。本研究将进一步探索这些模型在MRC任务中的应用潜力，特别是在处理DialogueMRC数据集时的表现。本研究采用基于提示的评测方法，该方法相较于传统微调，允许模型直接在多样化对话场景中进行理解和推理，无需进行特定的微调。为此，本研究设计了一系列提示语，引导模型准确理解并回答源自复杂多方对话的问题。

为了全面评估方法的有效性，本研究选取了几种最新的大型模型进行对比，包括ChatGLM3-6b(Du et al., 2022)、Vicuna-13b(Zheng et al., 2024)、Baichuan2-7b(Yang et al., 2023)、GPT-3.5-turbo(Brown et al., 2020)以及Qwen1.5-72b-chat(Bai et al., 2023)。这些模型均通过不同的架构和预训练策略，在NLP任务中取得了显著成绩。通过与上述模型的综合比较，本研究旨在探索基于提示的评测方法在激发模型深层次理解和推理能力方面的有效性。

## 5.4 实验结果与分析

### 5.4.1 与基于微调的预训练模型的对比结果与分析

如表2所示，在QA任务中，DSQA-CMD(Mac)模型分别在F1和EM指标上达到了45.9%和36.3%，显著优于先前的模型。特别是相较于使用传统方法，如不对文本进行切分处

理，同采用基于Longformer(Beltagy et al., 2020)架构的方法相比，DSQA-CMD(Mac)在F1指标上提高了5.4%，EM指标上提升了10.0%。这一结果突显了在问答任务中，多任务训练模型在性能均衡和优势方面的显著优越性。

| 模型 | QA | | DP | |
|---|---|---|---|---|
| | F1(%) | EM(%) | Link(%) | Link&Rel(%) |
| Longformer | 40.5 | 26.3 | - | - |
| Deep sequential | - | - | 73.1 | 42.2 |
| CSE-DPM | - | - | **76.2** | **46.7** |
| DSQA-CMD(Ro) | 41.1 | 33.0 | 61.4 | 36.1 |
| DSQA-CMD(Mac) | **45.9** | **36.3** | 61.6 | 37.2 |

表 2. 基于微调的预训练对比模型实验结果

### 5.4.2 与基于提示驱动的大语言模型对比实验结果与分析

表3展示了多任务学习模型与当前主流大型语言模型在DialogueMRC上的性能对比。结果表明，DSQA-CMD(Mac)在F1和EM指标上均优于参考的大型模型，尤其是在EM指标上相比于最接近的Qwen1.5-72b-chat模型，实现了12.8%的显著提高，同时在F1分数上也取得了2.3%的领先。分析大型语言模型在DialogueMRC数据集上的表现时，可以发现它们在生成答案时倾向于生成完整的句子或短语。这种生成式的回答格式常常无法严格对应到原文的确切位置，导致即使答案内容正确，也因格式不符合抽取式任务的要求而被评估为错误。这一现象强调了大模型在处理具体、特定领域任务时可能的局限性。随着模型参数数量的增加，其性能也随之提升，这强调了大语言模型参数量对于增强模型理解能力和处理复杂任务方面的关键作用。此外，实验表明尽管大模型在规模上的优势有助于提升通用任务的处理能力，但当面对具体的、特定领域的任务时，针对具体下游任务设计模型和训练显得更为关键。尤其是在中文多方对话理解任务中，大语言模型可能无法充分捕捉到任务的特定需求和细微差异。而多任务学习模型通过对问答和篇章解析任务的联合优化，不仅提升了模型对对话上下文的深层理解，还增强了针对具体任务需求的适应能力，从而实现了在关键性能指标上的显著提升。

| 模型 | QA | |
|---|---|---|
| | F1(%) | EM(%) |
| ChatGLM3-6b | 16.7 | 2.9 |
| Vicuna-13b | 21.1 | 5.9 |
| Baichuan2-7b | 25.1 | 8.7 |
| GPT-3.5-turbo | 34.5 | 15.2 |
| Qwen1.5-72b-chat | 43.6 | 23.3 |
| DSQA-CMD(Mac) | **45.9** | **36.3** |

表 3. 基于提示驱动的大语言对比模型实验结果

### 5.4.3 消融实验结果及分析

为了深入了解多任务学习模型在中文多方对话机器阅读理解任务中的表现，本文设计了一系列消融实验。实验结果如表4所示，其展示了DSQA-CMD模型相较于单任务模型在F1和EM指标上的提升。实验结果有力地证实了多任务学习策略的有效性。值得注意的是，多任务模型的出色表现揭示了任务间的正向互动效应。具体来说，篇章解析任务的融合促进了模型在共享特征空间中的表示学习，从而增强了问答性能。这种正向效应可归因于两个任务在语义理解和信息提取方面的内在联系。通过联合训练，模型能够在篇章解析中学习深层次语义关系和对话结构知识，进而在问答任务中更有效地定位和推理答案。

| 模型 | QA | | DP | |
|---|---|---|---|---|
| | F1(%) | EM(%) | Link(%) | Link&Rel(%) |
| QA-only | 43.2 | 33.7 | - | - |
| DP-only | - | - | 60.6 | 35.3 |
| DSQA-CMD(Mac) | **45.9** | **36.3** | **61.6** | **37.2** |

表 4. 消融实验结果

### 5.4.4 基于5W1H问题类型的模型性能评估

为深入理解本文提出的多任务学习模型在处理不同类型问题时的性能，本文分析了模型在5W1H问题类型上的表现。表5展示了各问题类型的评估结果，以及它们相较于全量数据集训练的平均性能的变化。本研究发现模型在处理“Who”和“Where”问题时表现尤为出色，这主要归因于其强大的实体识别能力，特别是在定位对话中的人物和地点方面，这些问题类型通常不需要复杂推理即可有效解答。相反，“Why”和“How”问题的表现较差，主要因为这些问题需要深度理解和推理，包括因果关系和方法论的把握，此外，这类问题的答案往往较长，包含多个信息点，要求模型不仅准确理解问题意图，还需从广泛的上下文中提取准确的信息片段，这大大增加了答案精准匹配的难度。模型在这些问题类型上的局限性揭示了其在处理复杂语义、逻辑推理以及长篇信息方面的不足。

| 问题类型 | 样本数 | F1(%) | EM(%) |
|---|---|---|---|
| What | 139 | 47.4(↑1.5) | 29.5(↓6.8) |
| Who | 138 | 69.7(↑23.8) | 68.1(↑31.8) |
| Why | 138 | 27.8(↓18.1) | 12.3(↓24.0) |
| How | 134 | 24.8(↓21.1) | 9.7(↓26.6) |
| When | 134 | 51.7(↑5.8) | 47.8(↑11.5) |
| Where | 135 | 53.6(↑7.7) | 50.4(↑14.1) |

表 5. 基于5W1H问题类型的模型性能评估结果

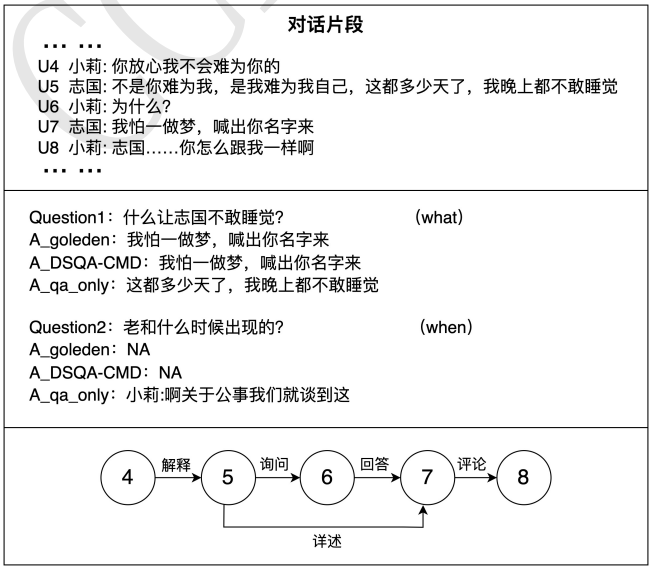



图 5. 案例分析

### 5.4.5 案例分析

图5所示对话片段涉及两个角色，分别是小莉和志国，他们之间的对话展示了角色之间的情感动机和个人困扰。本研究通过两个具体问题来评估模型的性能：一是询问导致志国失眠的原因，二是探究一个特定事件的出现时间。在回答“什么让志国不敢睡觉？”时，DSQA-CMD模型通过捕捉对话中的关键信息，志国担心梦中无意间喊出小莉的名字，成功地给出了问题的答案，这体现了其在理解对话中隐含的情绪和深层含义方面的出色能力。在这一对话片段中，每一轮交流都建立在前一轮的基础上，形成了逻辑上的连续性。例如，小莉的询问（U6）引出了志国的回答（U7），而志国的回答进一步详述了他的担忧（U7对U5的详述）。篇章解析的融入使模型能够识别这些细微的逻辑连接，进而在回答问题时提供更加准确和深入的理解。面对第二个问题“老和什么时候出现的？”时，由于问题存在歧义且信息不足。DSQA-CMD模型预测答案不存在，这展现了其在处理不确定性问题时的鲁棒性，同时体现了篇章解析在进一步整合和理解对话结构方面的重要作用。

## 6 结束语

本研究构建了面向中文多方对话的机器阅读理解数据集DialogueMRC，为中文MRC的研究提供了高质量的数据资源。此外，通过将篇章解析与问答任务相结合，本研究提出了融合篇章结构感知能力的中文多方对话问答模型，并在DialogueMRC数据集上证明了篇章解析在提升多方对话机器阅读理解能力方面的关键作用。本文的研究成果不仅推动了中文多方对话MRC领域的技术创新，还为智能客服、智能家居等相关应用提供了强有力的技术支持。未来，将试图融合图网络技术建模，以提出更加创新的多方对话MRC模型，为人机交互和自然语言处理的发展提供新的思路和方向。

## 参考文献

Stergos Afantenos, Eric Kow, Nicholas Asher, and Jérémy Perret. 2015. Discourse parsing for multi-party chat dialogues. pages pp–928.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Yiming Cui, Ting Liu, Zhipeng Chen, Shijin Wang, and Guoping Hu. 2016. Consensus attention-based neural networks for chinese reading comprehension. *COLING 2016 - 26th International Conference on Computational Linguistics, Proceedings of COLING 2016: Technical Papers*, pages 1777 – 1786.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. 2020. Revisiting pre-trained models for chinese natural language processing. *Findings of the Association for Computational Linguistics Findings of ACL: EMNLP 2020*, pages 657 – 668.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 1:320 – 335.

Xingyi Duan, Baoxin Wang, Ziyue Wang, Wentao Ma, Yiming Cui, Dayong Wu, Shijin Wang, Ting Liu, Tianxiang Huo, Zhen Hu, et al. 2019. Cjrc: A reliable human-annotated benchmark dataset for chinese judicial reading comprehension. pages 439–451.

Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. Chatgpt outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.

Yuchen He, Zhuosheng Zhang, and Hai Zhao. 2021. Multi-tasking dialogue comprehension with discourse parsing. In Kaibao Hu, Jong-Bok Kim, Chengqing Zong, and Emmanuele Chersoni, editors, *Proceedings of the 35th Pacific Asia Conference on Language, Information and Computation*, pages 551–561, Shanghai, China, 11. Association for Computational Linguistics.

Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. 2015. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28.

Shengluan Hou, Shuhan Zhang, and Chaoqun Fei. 2020. Rhetorical structure theory: A comprehensive review of theory, parsing methods and applications. *Expert Systems with Applications*, 157:113421.

Yuru Jiang, Yang Xu, Yuhang Zhan, Weikai He, Yilin Wang, Zixuan Xi, Meiyun Wang, Xinyu Li, Yu Li, and Yanchao Yu. 2022. The crecil corpus: a new dataset for extraction of relations between characters in chinese multi-party dialogues.

Yuru Jiang, Yu Li, Weikai He, Jie Chen, Yanchao Yu, and Yangsen Zhang. 2023. A new dataset and parsing model for chinese multiparty dialogue discourse structure. pages 221–227.

Jiaqi Li, Ming Liu, Min-Yen Kan, Zihao Zheng, Zekun Wang, Wenqiang Lei, Ting Liu, and Bing Qin. 2020. Molweni: A challenge multiparty dialogues-based machine reading comprehension dataset with discourse structure. *arXiv preprint arXiv:2004.05080*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100,000+ questions for machine comprehension of text. *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 2383 – 2392.

Bonnie Webber, Rashmi Prasad, Alan Lee, and Aravind Joshi. 2019. The penn discourse treebank 3.0 annotation manual. *Philadelphia, University of Pennsylvania*, 35:108.

Aiyuan Yang, Bin Xiao, Bingning Wang, Borong Zhang, Ce Bian, Chao Yin, Chenxu Lv, Da Pan, Dian Wang, Dong Yan, et al. 2023. Baichuan 2: Open large-scale language models. *arXiv preprint arXiv:2309.10305*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

张开颜, 张伟男, and 刘挺. 2021. 基于深度学习的多方对话研究综述. 中国科学: 信息科学, 51(8):1217–1232.