分类号 TP311 密级 公开

UDC 编号



硕士研究生学位论文

**题 目 面向多方对话文本的抽取式机器阅读理解技术研究**

**Research on Extractive Machine Reading Comprehension Method of Multiparty Ddialogue**

**Title**

学院（所、中心） 信息学院

专业名称 计算机系统结构

研究方向 人工智能、自然语言处理

研究生姓名 朱星宇 学号 12020115053

导师姓名 王津 职称 副教授

2023年6月

**扉页1：**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **云 南 大 学 硕 士 研 究 生 学 位 申 请 简 况 表** | **论 文 预 审** | 论文预审结果： | | | |
| 专家姓名 | 职称 | 所在单位（校内：学院/校外：所在单位） | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
| **论 文 送 审** | 专家姓名 | 职称 | 所在单位（校内：学院/校外：所在单位） | 结果 |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
|  |  |  |  |
| **论 文 答 辩** | 答辩结果： | | | |
| 答辩专家 | 职称 | 所在单位（校内：学院/校外：所在单位） | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |
|  |  |  | |

扉页2：

论文独创性声明及使用授权

本论文是作者在导师指导下取得的研究成果。除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果，不存在剽窃或抄袭行为。与作者一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

现就论文的使用对云南大学授权如下：学校有权保留本论文（含电子版），也可以采用影印、缩印或其他复制手段保存论文；学校有权公布论文的全部或部分内容，可以将论文用于查阅或借阅服务；学校有权向有关机构送交学位论文用于学术规范审查、社会监督或评奖；学校有权将学位论文的全部或部分内容录入有关数据库用于检索服务。

（内部或保密的论文在解密后应遵循此规定）

研究生签名： 导师签名： 日 期：

# 摘要

相较于篇章段落式文本，多方对话文本更接近于人类的交流方式，针对其建模并完成抽取式阅读理解任务对探究机器对人类语言的理解能力有重要意义。由于口语化的多方对话文本包含多个说话者和话语片段，导致文本结构更为复杂，迫切需要针对上下文中的复杂信息交互进行有效建模。在之前的工作中，通过从说话者和话语片段两个层面将对话文本进行信息解耦能够有效帮助模型理解对话上下文。然而，除了保证模型能够充分理解给定的对话上下文外，针对从问题到答案的逻辑推理关系进行建模同样不可或缺，否则会导致模型预测的答案不够精确。因此，本文利用多任务学习的方式分别从说话者和话语片段两个角度解耦对话信息，再重点围绕增强对话信息的利用以及模型从问题到答案的逻辑推理能力进行研究。论文的主要工作内容如下：

(1) 针对对话文本信息利用不充分的问题，本文提出了基于关键话语片段的信息交互增强模型。将伪自监督的关键话语片段预测子任务给出的结果作为答案的候选序列并与问题拼接，模拟人类双向思考的过程来捕获对话上下文、问题、关键话语片段三者之间的信息交互，为最终答案的预测提供线索指导。并且，增加最小化互信息作为约束，减小子任务提取的说话者特征信息和话语片段特征信息之间的信息重叠，提高信息质量。通过在多方对话问答数据集上进行的充分的实验验证了所提出方法的有效性。在Molweni和FriendsQA数据集上的EM和*F*1指标均有所提升。

(2) 针对模型在面向问答任务时逻辑推理能力不足的问题，本文提出了逻辑推理记忆网络模型。引入篇章段落式问答数据集作为外部知识注入，通过逻辑推理记忆编码器将问答相关的推理知识整合为模型的潜在记忆并保存至下游的多方对话机器阅读理解任务中，有效提高模型的逻辑推理能力。实验中验证了所提出方法的有效性，在Molweni和FriendsQA数据集上的EM和*F*1指标显著提升，取得了最优结果

**关键词：**多方对话；机器阅读理解；问答逻辑推理；信息交互；外部知识注入

# Abstract

Compared with plain texts, multi-party dialogue texts are closer to human communication. Modeling the dialogue and completing extractive reading comprehension tasks is important for exploring the ability of machines to understand human language. Due to colloquial dialog text and complex text structure with multiple speakers and utterances, there is an urgent need to model information flow interactions. In previous work, decoupling information from the dialogue text at both speaker and utterance levels can effectively help the model understand the dialogue context. However, in addition to ensuring that the model fully understands the dialogue contexts, it is also essential to model the logical inference relations from questions to answers, otherwise the answers predicted by the model will be inaccurate. Therefore, this paper models dialogue information from speaker and utterance levels based on multi-task learning. And then, focuses on enhancing dialogue information utilization and the modeling of logical inference from questions to answers. The main work of the paper is as follows:

(1) To address the problem of inadequate utilization of dialogue text information, this paper proposed an enhanced key-utterance-based information interaction model. The results given by the pseudo-supervised key-utterance prediction subtask were used as the candidate sequences of answers and concatenated with the questions. And the bidirectional thinking process of humans was simulated to capture the information interactions among the dialogue context, questions, and key-utterance, providing clues to guide the prediction of the final answers. The effectiveness of the proposed method was verified by adequate experiments on multi-party dialogue MRC datasets. The EM and *F*1 metrics on the Molweni and FriendsQA datasets have improved significantly.

(2) To address the problem of the model's insufficient logical inference ability in MRC tasks, this paper proposed a memory network with logistic inference (LIMN). The LIMN introduced the plain article MRC dataset as external knowledge injection, and the logical inference memory encoder incorporated the MRC-related inference knowledge as latent memory and transferred it to the downstream multi-party dialogue MRC task, which effectively improves the model's logical inference ability from questions to answers. Extensive experiments verified the effectiveness of the proposed model. The EM and *F*1 metrics on the Molweni and FriendsQA datasets have improved by a large margin, achieving state-of-the-art results.

**Keywords**: Multi-party dialogue; Machine reading comprehension; Logistic inference; Information interaction; External knowledge injection

# 目录

[摘要 I](#_Toc127824249)

[Abstract II](#_Toc127824250)

[目录 IV](#_Toc127824251)

[插图 VII](#_Toc127824252)

[表格 IX](#_Toc127824253)

[第1章 绪论 1](#_Toc127824254)

[1.1 研究的背景和意义 1](#_Toc127824255)

[1.2 国内外研究现状分析 3](#_Toc127824256)

[1.2.1 面向篇章段落的机器阅读理解 3](#_Toc127824257)

[1.2.2 面向对话文本的机器阅读理解 4](#_Toc127824258)

[1.3 研究内容与论文组织结构 5](#_Toc127824259)

[1.3.1 研究内容 5](#_Toc127824260)

[1.3.2 组织架构 6](#_Toc127824261)

[第2章 面向多方对话的抽取式机器阅读理解技术 8](#_Toc127824262)

[2.1 面向多方对话的抽取式机器阅读理解问题定义 8](#_Toc127824263)

[2.1.1 任务描述 8](#_Toc127824264)

[2.1.2 抽取式多方对话问答数据集 8](#_Toc127824265)

[2.1.3 评价指标 9](#_Toc127824266)

[2.2 预训练语言模型 10](#_Toc127824267)

[2.2.1 Transformer 结构 10](#_Toc127824268)

[2.2.2 预训练语言模型 12](#_Toc127824269)

[2.3 面向对话文本的机器阅读理解技术研究 14](#_Toc127824270)

[2.3.1 基于关系图结构和语篇关系连接的对话建模方法 15](#_Toc127824271)

[2.3.2 基于多任务学习的对话建模方法 18](#_Toc127824272)

[2.3.3 基于外部知识注入的阅读理解增强方法 19](#_Toc127824273)

[2.4 本章小节 20](#_Toc127824274)

[第3章 基于关键话语片段交互增强的对话机器阅读理解方法 21](#_Toc127824275)

[3.1 引言 21](#_Toc127824276)

[3.2 基于关键话语片段交互增强的对话机器阅读理解模型 22](#_Toc127824277)

[3.2.1 对话上下文编码器 23](#_Toc127824278)

[3.2.2 双向多头共同注意力机制 24](#_Toc127824279)

[3.2.3 说话者预测模块 24](#_Toc127824280)

[3.2.4 关键话语片段预测模块 25](#_Toc127824281)

[3.2.5 基于关键话语片段的信息交互 26](#_Toc127824282)

[3.2.6 基于最小化互信息的信息解耦 26](#_Toc127824283)

[3.2.7 基于范围的答案抽取 27](#_Toc127824284)

[3.3 实验分析 28](#_Toc127824285)

[3.3.1 数据集与评价指标 28](#_Toc127824286)

[3.3.2 实验设置 28](#_Toc127824287)

[3.3.3 基线模型 29](#_Toc127824288)

[3.3.4 对比实验 30](#_Toc127824289)

[3.3.5 消融实验 31](#_Toc127824290)

[3.3.6 超参数分析 31](#_Toc127824291)

[3.3.7 案例分析 31](#_Toc127824292)

[3.4 本章小结 33](#_Toc127824293)

[第4章 基于外部知识注入增强的对话机器阅读理解方法 34](#_Toc127824294)

[4.1 引言 34](#_Toc127824295)

[4.2 面向多方对话的逻辑推理记忆网络模型 36](#_Toc127824296)

[4.2.1 对话上下文编码器 36](#_Toc127824297)

[4.2.2 推理记忆编码器 36](#_Toc127824298)

[4.2.3 说话者和话语片段信息解耦编码器 38](#_Toc127824299)

[4.2.3 信息融合与交互 38](#_Toc127824300)

[4.2.4 多任务学习策略 38](#_Toc127824301)

[4.3 实验分析 39](#_Toc127824302)

[4.3.1 数据集与评价指标 39](#_Toc127824303)

[4.3.2 实验设置 40](#_Toc127824304)

[4.3.3 基线模型 40](#_Toc127824305)

[4.3.4 对比实验 41](#_Toc127824306)

[4.3.5 消融实验 42](#_Toc127824307)

[4.3.6 超参数分析 43](#_Toc127824308)

[4.3.7 说话者和话语片段数量对模型性能的影响 44](#_Toc127824309)

[4.3.8 案例分析 46](#_Toc127824310)

[4.4 本章小结 47](#_Toc127824311)

[第5章 总结与展望 48](#_Toc127824312)

[5.1 论文的主要工作及贡献 48](#_Toc127824313)

[5.2 未来研究工作 49](#_Toc127824314)

[参考文献 50](#_Toc127824315)

[在学期间研究成果 56](#_Toc127824316)

# 插图

[图1.1 抽取式多方对话阅读理解任务 2](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793123)

[图2.1 Transformer Block结构（编码器） 11](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793124)

[图2.2 缩放点积注意力机制(左)、多头注意力机制(右) 12](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793125)

[图2.3 BERT的输入包括词嵌入、分段嵌入、位置嵌入三部分 13](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793126)

[图2.4 ELECTRA预训练架构与RTD任务 14](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793127)

[图2.5语篇关系图的构建 15](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793128)

[图2.6 说话者连接图 16](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793129)

[图2.7 语篇关系图 17](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793130)

[图2.8 说话者注意力掩码 18](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793131)

[图2.9 自监督的说话者预测任务和伪自监督的关键话语段预测任务 19](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793132)

[图3.1 错误预测示范 22](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793133)

[图3.2 模型的整体架构 23](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793134)

[图3.3 关键话语片段预测模块和说话者预测模块 25](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793135)

[图3.4 *N*sMHA对模型性能的影响 31](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793136)

[图4.1 通过语篇关系链接注释的多方对话文本 35](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793137)

[图4.2 LIMN模型的两阶段训练模式 36](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793138)

[图4.3 LIMN模型的整体架构 37](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793139)

[图4.4 LIMN模型在验证集上进行的超参数微调实验 43](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793140)

[图4.5 说话者和话语片段数量对模型性能的影响 44](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793141)

# 表格

[表2-1：抽取式多方对话机器阅读理解任务数据集 9](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793193)

[表2-2：单词精确率和召回率的计算 10](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793194)

[表2-3：话语片段之间的语篇关系 15](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793195)

[表3-1：实验中的超参数设置 28](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793196)

[表3-2：所提出模型在Molweni和FriendsQA数据集上的对比实验结果 29](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793197)

[表3-3：所提出模型在Molweni数据集上的消融实验结果 30](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793198)

[表3-4：案例分析 32](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793199)

[表4-1：LIMN模型在实验中的超参数设置 40](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793200)

[表4-2：LIMN模型继续学习阶段在SQuAD 2.0验证集上的测评结果 41](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793201)

[表4-3：LIMN模型在Molweni和FriendsQA数据集上的对比实验结果 42](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793202)

[表4-4：LIMN模型在Molweni和FriendsQA数据集上的消融实验 43](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793203)

[表4-5：LIMN模型在Molweni数据集上的案例分析 45](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793204)

[表4-6：LIMN模型在FriendsQA数据集上的案例分析 46](file:///C:\Users\Xingyu%20Zhu\Desktop\2-19-朱星宇论文.docx#_Toc127793205)

# 绪论

## 1.1 研究的背景和意义

人工智能是计算机科学的一个重要分支，其研究方向主要包括计算机视觉、自然语言处理、专家系统、启发式问题解决等。其中，自然语言处理技术(Natural Language Processing, NLP)旨在使计算机能够解读、处理和理解人类语言，并根据需要完成特定的任务，例如文本分类、文本摘要、特定信息抽取、文本生成等。NLP技术在日常生活中也同样被广泛应用，例如语音助手、翻译软件、搜索引擎、智能对话系统等。

机器阅读理解(Machine Reading Comprehension, MRC)[1]作为NLP中重要的一部分，其目的就在于利用计算机算法(模型)理解给定文本的语义信息并回答相关问题。机器阅读理解任务由给定文本、问题、答案三要素构成，根据任务形式的不同主要分为四类：选择式(给定候选答案并选择)、抽取式(从文章段落中抽取连续片段作为答案)、生成式(由机器生成问题答案)、填空式(填补文章中的空白词句)。机器阅读理解的研究对于人工智能的发展有着重要的意义，早在 20 世纪 50 年代，图灵就提出通过人机交互衡量机器的智能水平[2]；而在今天，例如智能机器人、自动客服系统、智能家居等许多人机交互的实际应用都离不开对机器阅读理解算法的研究。机器阅读理解作为一种测试机器对语言理解能力的重要方法有长达半个多世纪的研究历史，但受制于数据规模以及计算机硬件算力，并未取得重大的研究突破。近年来，随着硬件技术的飞速发展以及深度学习技术的兴起， 越来越多的研究者参与到机器阅读理解技术的研究中，每年在ACL和EMNLP等NLP领域顶级国际会议中有许多相关研究工作发表，在例如SQuAD[3]、SQuAD 2.0[4]、QuAC[5]、CoQA[6]、RACE[7]等众多机器阅读理解评测数据集上，机器的作答水平已经接近甚至超越了人类。

早期传统的机器阅读理解主要以篇章段落式的文章为主，内容多以独白的百科内容为主，其文本结构较为清晰，代表数据集有SQuAD。随着机器阅读理解技术的发展，对话这一特殊形式的文本开始被广泛关注。相较于篇章段落式文本，对话文本记录的是人与人的交流过程，与人机交互的形式更接近，所以机器能否准确地理解对话上下文并做正确推理将很大程度上影响智能对话系统的性能。而相较于两方对话，由三名及以上角色组成的“群聊”式多方对话文本则更具挑战性。研究面向多方对话文本的机器阅读理解技术不仅能够探索提升机器对于对话文本的理解能力，建立高效的智能问答系统，也可以运用于在例如博客、论坛、群聊天等多方参与的文本数据中进行数据挖掘、信息抽取等其他NLP任务中。



图1.1 抽取式多方对话阅读理解任务

多方对话的复杂性和建模难点主要包括两方面。数据结构上，由于引入了多个说话角色的缘故，各个角色发声的多个话语片段{说话者：话语片段内容}共同组成完整的对话上下文。同时，由于发言顺序的不确定性，某一信息的完整描述不一定存在于连续的话语片段中。内容上，对话文本相较于篇章式文本的表达更口语化，且每个说话者的语言风格和对话目的一般不同；在人数大于二的多方对话文本中，由于人数较多，上下文中通常存在大量的指代关系需要辨析。此外，各个角色及其话语片段之间也存在着复杂的信息流交互需要建模，增加了对文本的理解难度。例如，图 1.1 展示了从多方对话数据集Molweni[8]中抽取的一个例子，包含一段由5名说话者组成的对话以及3个相关的问题，对话的主要内容是*variant*帮助其他人解决在使用Ubuntu操作系统时遇到的一些问题。根据这段对话，算法模型需要从对话中找出三个问题的对应答案。

## 1.2 国内外研究现状分析

在机器阅读理解任务中，需要针对特定的文本和下游任务，设计适合的模型架构以实现高性能的阅读理解。在本节中，将分别介绍面向篇章段落的机器阅读理解技术以及面向对话文本的机器阅读理解技术。

### 1.2.1 面向篇章段落的机器阅读理解

2015年深度学习技术开始被用于机器阅读理解研究，文本形式以篇章段落为主。同年Hermann等人[1]发布大规模机器阅读理解数据集CNN/Daily Mail，并提出了基于双向循环神经网络LSTM的机器阅读理解模型Attentive Reader。2016年，基于神经网络的AS Reader[9]则在CNN/Daily Mail数据集上取得了大幅的性能提升。考虑到CNN/Daily Mail数据集为完形填空式任务，与人类的问答形式相差较大，2016年斯坦福发布了由人工标注的高质量抽取式机器阅读理解数据集SQuAD[3]，提高了任务难度，吸引了大量研究者。针对抽取式机器阅读理解任务，一般采用循环神经网络进行上下文编码，再利用注意力机制建模上下文和问题之间的联系。例如BiDAF[10]中先分别对文章和问题进行编码，再利用注意力机制将二者关联后找出问题答案，类似的工作还有 DCN[11]等。此外，一些模型还通过多次交互的方式来逐步推理问题答案，提高模型性能，例如GA Reader[12]、R-Net[13]、RM Reader[14]等。QANet[15]则是考虑到循环神经网络无法并行计算的问题，提出利用自注意力机制和卷积神经网络实现全局的信息交互，大幅提升了模型的速度。SQuAD数据集是机器阅读理解研究发展的一个里程碑，在其影响下，许多来自不同领域的高质量、高难度的机器阅读理解数据集不断涌现。例如选择式数据集SciQ[16]、RACE[7]，抽取式数据集SQuAD 2.0[4]、NewsQA[17]、SearchQA[18]，完形填空式数据集CBT[19]、BioRead[20]等。

2018年预训练语言模型(Pre-trained Language Model, PLM)的出现则开启了NLP领域新的研究范式，即通过大量语料库和预训练任务训练具有强大表征能力的预训练模型，再根据特定的下游任务来微调模型，将通识语言知识转移到具体的任务中，这种范式下模型所取得性能获得了极大提升。预训练语言模型大多以Transformer[21]结构为基础，最具代表性的是BERT[22]，其采用被遮挡单词预测和下一句子预测任务在大量语料库上训练，之后在诸多下游任务[23,24]上微调均取得了当时的最优性能，之后更多的模型研究针对预训练预料库和预训练任务的形式展开。例如，ALBERT[25]通过参数共享压缩模型大小并提升训练速度；XLNet[26]结合了自编码和自回归；RoBERTa[27]使用更大语料库并引入动态掩码；ELECTRA[28]引入生成器-判别器的预训练形式并使用替换词检测任务，提升了预训练模型的训练效率和性能。

### 1.2.2 面向对话文本的机器阅读理解

面向对话文本的机器阅读理解则在近几年被广泛研究，同样发布了大量高质量的测评数据集供研究者使用。例如响应选择式数据集MuTual[29]、DailyDialog[30]、DialogueNLI[31]、DREAM[32]等，基于背景知识的连续问答数据集QuAC[5]和CoQA[6]，抽取式问答数据集Molweni[8]和FriendsQA[33]等。

面向对话文本的机器阅读理解模型一般采用预训练语言模型作为编码器，但由于预训练语言模型所使用的语料库一般为篇章段落形式，与对话文本存在较大差距，直接微调一般无法取得理想的效果。因此，目前大部分的研究都使用预训练语言模型作为上下文编码器，之后根据对话的结构和具体阅读理解任务的形式设计更详细的对话建模策略，进一步完善特征的提取以提升模型性能。

一种建模方向为将整个对话上下文构建为关系图，图中的每个结点表示对话上下文中的一个话语片段，形式为{说话者：话语片段内容}，图中的边表示各个话语片段之间的关系，再利用图神经网络(Graph Convolutional Network, GCN)[34]来建模对话上下文，典型工作包括DialogueGCN[35]、DADgraph[36]等。另一研究方向为将复杂的对话结构根据不同元素进行信息解耦。MDFN[37]通过为多头注意力机制(Multi-head Attention, MHA)[21]添加不同掩码，分别针对说话者之间的关系和话语片段之间的关系进行建模。BiDeN[38]则使用类似的方式引入回溯阅读，通过多头注意力机制实现未来-现在、现在-现在、过去-现在三个角度的信息解耦。此外，文献[39]则结合关系图结构和说话者注意力掩码来建模对话上下文。SUP[40]则引入多任务学习，使用自监督的说话者预测任务来隐式建模说话者之间的信息交互，伪自监督的关键话语片段预测任务则用来建模话语片段与问题之间的联系。此外，也有一些针对对话文本的预训练任务，例如下一话语片段预测、话语片段排序、词和话语片段级别的掩码预测任务等[41,42]。

## 1.3 研究内容与论文组织结构

本节将从目前面向多方对话文本的机器阅读理解任务中存在的问题切入，引出并介绍本文的主要研究内容，最后给出本文的组织结构。

### 1.3.1 研究内容

根据上述对研究现状的简述，目前的相关研究与应用还存在以下几个问题：(1) **关键话语片段信息的利用及其与上下文之间的交互较弱。**目前的研究更注重对话上下文信息的解耦约建模，而针对抽取式机器阅读理解任务，关键话语片段是答案在话语片段层级的体现，包含了答案的位置信息以及其与上下文、问题之间的交互信息，对精确答案的预测具有重要的指导意义。然而，目前的研究针对问题、对话上下文、关键话语片段之间信息关联的建模较弱。(2) **针对问答任务中的逻辑推理关系建模不足。**在抽取式机器阅读理解任务中，除了对话上下文中复杂信息流的建模，从问题到答案的逻辑推理关系建模也同样重要。原因是在对话文本中存在复杂的指代关系或层次关系需要辨析，问答中通常存在需要进行多步推理才能得到的最终答案。

针对上述的两个问题，考虑到基于关系型图结构对话建模一般需要额外的话语关系标注数据来构建图，本文将从对话信息解耦的角度来建模多方对话并完成抽取式阅读理解任务，基于多任务学习，采用SUP模型中提出的说话者预测子任务和关键话语片段预测子任务来辅助模型建模对话上下文中的复杂信息。主要研究内容如下：

（1）基于关键话语片段的信息交互策略以及增强特征信息的解耦。本文将关键话语片段预测子任务给出的关键话语片段作为先验信息，将其与问题拼接后得到“问题+候选答案段”的形式，之后采用双向多头注意力机制来模拟人类完成阅读理解的过程，分别从“对话上下文=>问题及候选答案段”和“问题及候选答案段=>对话上下文”两个角度增强信息的交互，捕获上下文、问题、关键话语片段三者之间的关键信息。此外，我们还引入最小化互信息作为约束，减小两个子任务提取信息的冗余。在FriendsQA和Molweni数据集上的实验结果表明了所提出方法的有效性。

（2）整合外部问答知识作为潜在记忆以增强模型针问答任务的推理能力。本文提出了一种逻辑推理记忆网络模型(Memory Network with Logistic Inference, LIMN)。引入篇章段落型问答数据集SQuAD 2.0作为外部知识，使用LIMN将其保存后用来增强模型针对问答型对话阅读理解任务的逻辑推理能力。在FriendsQA和Molweni数据集上的实验结果证明，增强逻辑推理能力后的模型性能更优。

### 1.3.2 组织架构

全文共5个章节，各章节内容安排如下：

第1章为绪论，介绍了本文的研究背景及意义，之后从面向篇章段落的机器阅读理解模型和面向对话文本的机器阅读理解模型两方面简述了国内外相关工作和研究进展，以此为基础提出了目前面向多方对话文本的抽取式机器阅读理解任务的难点，最后总结本文的主要工作内容并给出文章组织架构。

第2章为面向对话文本的抽取式机器阅读理解相关技术研究，描述了问题的形式化定义、数据集和评测指标等，详细介绍了本文所使用的预训练语言模型，最后对目前较新的多方对话建模策略进行了分类梳理。

第3章针对模型中对话上下文、问题、关键话语片段三者之间信息交互不够充分的问题，采用双向多头注意力机制模拟人类阅读理解时双向思考的过程捕获三者之间的关键信息交互，并通过最小化互信息减少说话者和话语片段表征之间的信息重叠，增强模型性能。最后通过充分的实验验证了所提出方法的有效性。

第4章针对在多方对话问答模型中逻辑推理关系的弱建模问题，提出了一种逻辑推理记忆网络模型，通过引入外部的问答相关知识并将其整合为模型的潜在记忆以增加模型在下游任务中的从问题到答案的逻辑推理能力。最后通过充分的实验验证了所提出方法的有效性。

第5章对论文在面向多方对话文本的抽取式机器阅读理解任务中的相关工作进行总结，并对未来的研究方向做出展望。

# 面向多方对话的抽取式机器阅读理解技术

## 2.1 面向多方对话的抽取式机器阅读理解问题定义

### 2.1.1 任务描述

机器阅读理解任务属于有监督的机器学习问题，对话文本与篇章段落文本不同，由来自多个说话者(Speaker)发声的话语片段(Utterance)组成，每个话语片段形式为{说话者名称：说话内容}。在机器阅读理解任务中，给定一段包含*N*个话语片段的对话文本上下文，其中，表示第*n*个话语片段，*Sn*表示说话者的名称，表示说话的内容。*M*个与相关的问题，与其相对应的*M*个答案，其中，，表示第*m*个（问题，答案）对。抽取式机器阅读理解任务的目标是：对于给定的对话上下文和问题作为输入，建立机器学习模型*f*，输出问题对应的答案：



对于抽取式机器阅读理解任务，一定为中的一个连续片段，但当数据集中含有不可回答的问题时，可以为空序列，表示对应的问题无法回答。

### 2.1.2 抽取式多方对话问答数据集

在本文中，使用2个不同主题类型的多方对话文本问答数据集：Molweni[8]和FriendsQA[33]数据集。在多方对话的抽取式问答中，这些数据集被广泛的用于模型的性能验证，数据集的具体描述如下，表2.1展示了数据集的划分情况。

(1) Molweni数据集的内容主题是关于Ubuntu操作系统的交流对话。Li等人[8]从Ubuntu语料库[43]中抽取了话语片段数量为8~15，说话者数量为2-9的对话，采用人工注释的方式提出相应问题并标记答案。该数据集包含超过1万个对话上下文，25779个可回答的问题以及4287个不可回答的问题。

表2-1：抽取式多方对话机器阅读理解任务数据集

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Molweni | | | FriendsQA | | |
| 训练集 | 验证集 | 测试集 | 训练集 | 验证集 | 测试集 |
| 对话数量 | 8771 | 833 | 100 | 973 | 136 | 113 |
| 问题数量 | 24682 | 2513 | 2871 | 9874 | 1182 | 1201 |

(2) FriendsQA数据集的内容则主要是开放域的对话。该数据集是从著名的美国电视剧《老友记》的剧本中抽取得到，由人工针对对话上下文提出问题并标注答案。在本数据集中，共1222个对话上下文，10610个可回答的问题，没有不可回答的问题。此外，该数据集中的对话上下文比Molweni更长，并且含有更多的说话者。

### 2.1.3 评价指标

在抽取式机器阅读理解任务中，模型预测的答案一定来自于对话上下文，属于半客观式答案，评价模型性能的指标包括以下两项：

(1) EM (Exact Match) 表示完全匹配。将模型给出的答案记作*P=*[*Pstart*, *Pend*]，真实答案记作*y=*[*ystart*, *yend*]，当且仅当*Pstart=ystart*，且*Pend=yend*时，该条数据的EM值为1，否则EM值为0。EM在整个数据集上的计算方式如公式。



(2) *F*1 (*F*1 Score) 则被用来衡量答案的匹配程度。例如，正确答案为“深度学习”，模型给出的答案为“是深度学习”时，EM分数为0，但此时模型给出的答案已经十分接近正确答案，这种情况下*F*1能更好的反映出模型的性能。*F*1通过计算单词精确率(precision)和召回率(recall)的调和平均数得到，通过表2-2和公式可以计算出精确率和召回率：



*F*1计算过程则如公式所示。

表2-2：单词精确率和召回率的计算

|  |  |  |
| --- | --- | --- |
|  | 单词在正确答案中 | 单词不在正确答案中 |
| 单词在模型给出的答案中 | TP | FP |
| 单词不在模型给出的答案中 | TN | FN |



## 2.2 预训练语言模型

在预训练语言模型大规模应用之前，一般采用静态文本表示方法（例如Word2Vec、Glove）先获取词向量，之后利用循环神经网络（例如RNN、LSTM、GRU）对文本序列进行编码。

然而，静态文本表示方法是针对词向量进行训练，同一个词的向量表示不会因为上下文语境的变化而改变，但人类在阅读时通常会根据语境来理解词的意思，即相同的词在不同的语境中常表现出不同的意思。动态文本表示模型ELMo[44]的出现解决了这一问题，ELMo先用大量的语料库对模型进行预训练，再根据下游任务，通过权值的调整来调整词向量，解决了一词多义问题。

此外，循环神经网络因为其结构的原因，下一时刻的计算依赖于上一时刻的计算结果，因此难以实现并行计算，效率较低。Transformer[21]结构则基于注意力机制进行序列计算，可以实现高度并行。因此，以BERT[22]为代表的预训练语言模型基本都以Transformer结构为基础，结合大规模语料库进行预训练以获得强大的表征能力，被广泛应用于NLP下游任务。

### 2.2.1 Transformer 结构

最初提出Transformer时包括编码器和解码器两个部分，主要应用于机器翻译任务。以BERT为代表的预训练语言模型则采用了Transformer的编码器部分，由若干个Transformer Block组成，本节将重点介绍其编码器的结构。

传统的循环神经网络采用循环展开的机制进行计算以对序列进行建模，因此无法实现高度并行。Transformer则使用注意力机制实现对序列关系的计算。注意力机制使用全连接网络和矩阵乘法实现，可以实现高度并行；同时，Transformer在输入层引入位置嵌入(Positional Embedding, PE)来表示输入序列的位置信息，以保证序列编码的效果，计算方式如下：



图2.1 Transformer Block结构（编码器）



其中，*pos*表示词的位置，*dmodel*表示模型的向量维度，*i*表示向量维度中的具体位置。之后，PE将和原始输入序列词嵌入相加并送入后续网络。

Transformer Block是Transformer模型中的基本模块，其结构如图2.1所示。每个Transformer Block由多头注意力机制(Multi-Head Attention, MHA)和前馈神经网络(Feed Forward Network, FFN)组成，层间则使用残差连接(Residual Connection)和层归一化(LayerNorm)来增强层间关系。

如图2.2所示，多头注意力机制含有*h*个缩放点积注意力机制(Scaled Dot-Product Attention)，使用点积的方式进行计算，对于输入的上下文表征(*L*表示文本长度，*d*表示向量维度)，多头注意力的计算过程如下：



其中*WiQ*，*WiK*，*WiV*，*WO*表示可训练参数，对应图2.2中的线性层参数；除以为缩放操作，*dk*表示*K*的维度；*h*为“头数”，目的是将原始输入映射至不同的向量空间以得到多个视角上的信息，最后进行汇总以获得全面的表征输出。





图2.2 缩放点积注意力机制(左)、多头注意力机制(右)

前馈神经网络则由两个线性层组成，计算过程如公式所示。其中，*W1*，*W2*，*b1*，*b2*为可训练参数，采用ReLU作为激活函数。



最后，加入层间残差连接和层归一化操作，以上下文表征*H*为输入，Transformer Block完整的计算表示如下：



### 2.2.2 预训练语言模型

如今，基于Transformer结构的预训练语言模型具有其强大语言表征能力，被广泛应用于NLP领域。以下将简要介绍本文中所使用到的两种预训练语言模型BERT[22]和ELECTTRA[28]。

1. BERT (Bidirectional Encoder from Transformer)

BERT作为具有划时代意义的预训练语言模型，一经推出便大幅度刷新了NLP领域大多数任务的榜单，开启了NLP领域解决问题的新范式，即在大规模语料库上预训练之后针对下游任务微调，成为了预训练语言模型时代研究的标配基线模型。

BERT使用了Transformer模型的编码器部分，通过在嵌入层后堆叠多个Transformer Block作为编码器。相较于之前基于循环神经网络的模型，BERT不仅在训练和推理时更加高效，而且多层的Transformer Block堆叠保证了其强大的语言表征能力。与原始的Transformer不同，BERT在嵌入层中增加了分段嵌入部分(Segment Embedding)以表示前后两个句子或分组，针对文本输入序列*w =* [*w1,w2,…,wn*]，其嵌入层表示*E =* [*E1,E2,…,En*]由词嵌入、部分嵌入、位置嵌入三部分组成，如图2.3所示。



图2.3 BERT的输入包括词嵌入、分段嵌入、位置嵌入三部分

在预训练任务上，BERT采用掩码语言模型 (Masked Language Modeling, MLM)和下一句子预测 (Next Sentence Prediction, NSP)两个任务进行联合训练。MLM任务与完形填空类似，目的是训练深度的双向文本表征能力。原作者在输入序列中随机选择15%的单词，在被选择的15%单词中，80%被替换为特殊字符[MASK]，10%被随机替换为其他单词，10%则保持不变，最终模型需要恢复出被MASK掉的单词。具体方式为经过多层的编码后，从最终输出的上下文表征*H*中选择与[MASK]对应的向量，在词表上进行分类，而不是重建整个输入序列。NSP任务则是为了建模句间关系，即输入两个句子A和B，句间用特殊字符[SEP]分隔，模型需要判断B是否为A的下一句。

1. ELECTRA (Efficiently Learning an Encoder that Classifies Token Replacements Accurately)

MLM任务虽然能够有效的学习输入文本的上下文表征，但对于每个输入样本，只有15%的单词与损失函数的计算有关，文本的利用效率不够高，模型并未充分学习完整的文本特征，需要消耗大量的计算资源以保证模型被充分训练。

针对这些问题，ELECTRA提出了一种新的预训练任务：替换词检测(Replaced Token Detection, RTD)。基于RTD任务，ELECTRA在预训练时引入了类似对抗神经网络(Generative Adversarial Network, GAN)的架构如图2.4所示，包括一个生成器(Generator)和一个判别器(Discriminator)，两者均为基于Transformer的编码器。



图2.4 ELECTRA预训练架构与RTD任务

在ELECTRA的预训练过程中，生成器为一个参数较小的编码器，采用MLM任务训练。对于给定的原始序列*x =* {*x1,x2,…,xn*}，根据MLM任务将其替换为作为生成器输入*x*masked *=* REPLACE(*x, m,* [MASK])，输出为*x*corrupt *=* REPLACE(*x*, *m*, )，其中m表示被替换为[MASK]字符的位置，表示生成器将[MASK]还原后得到的字符。判别器则为参数较大的编码器，采用RTD任务训练，以*x*corrupt为输入，判断序列*x*corrupt中的每一个词是否为*x*中的原始序列。RTD任务本质上是一个在单词级别的二分类任务，与MLM任务相比，输入序列中的每一个词都参与到了损失函数的计算，大大提高了文本的利用效率，减少计算资源的使用；并且相较于MLM任务，RTD作为分类任务与众多下游任务的形式更接近。

值得注意的是，虽然模型架构类似于GAN，但生成器的训练目标是尽可能的让*x*corrupt与*x*一致而不是欺骗判别器；此外，判别器端RTD任务的梯度也不会反向传播至生成器中。预训练结束后，生成器将被抛弃，仅使用判别器来微调下游任务。

## 2.3 面向对话文本的机器阅读理解技术研究

预训练语言模型的出现极大地促进了自然语言处理领域的发展，在面向对话文本的机器阅读理解任务中，通常都使用预训练语言模型作为上下文编码器，借助其强大的文本表征能力来获取输入文本的表征矩阵，之后再针对对话文本的特点设计更细致的对话建模策略。再本小节中，主要将介绍基于关系图结构的对话建模方法、基于多任务解耦对话特征的建模方法、以及通过知识注入增强模型理解能力的方法。

表2-3：话语片段之间的语篇关系

|  |  |  |  |
| --- | --- | --- | --- |
| Comment | 评论 | Clarification question | 澄清问题 |
| Elaboration | 细化 | Acknowledgement | 致谢 |
| Continuation | 继续 | Explanation | 解释 |
| Conditional | 条件 | QAP | 问题-答案对 |
| Alternation | 交替 | Background | 背景 |
| Q-Elab | 问题-详细说明 | Result | 结果 |
| Narration | 旁白 | Correction | 更正 |
| Parallel | 平行 | Contrast | 对比 |

### 2.3.1 基于关系图结构和语篇关系连接的对话建模方法

1. 语篇关系感知型对话图网络 (Discourse-aware Dialogue Graph, DADgraph)

DialogueGCN[35]将每一个话语片段都作为图中一个结点，通过结点之间的边来描述各个话语片段之间的联系，对话上下文便构成了一张图。但考虑到多方对话文本中说话者众多，各个话语片段间的信息流交互极为复杂，为了更好的建模，DADgraph[36]引入了语篇关系依赖注释(discourse dependency links and discourse relations)。作者将话语片段之间的关系总结为16种如表格2-3所示，对原始文本拿数据进行注释后，将对话上下文构建为关系图，不同的语篇关系用不同类型的有向边描述，构建后的关系图如图2.5所示。



图2.5语篇关系图的构建 (a)为对话上下文；(b)为相应的问题和答案;(c)为根据语篇关系连接构建的关系图

模型的构建方面，DADgraph主要分为三步（a）利用预训练好的BERT获取完整的对话上下文表征矩阵。再分别将每一个话语片段单独作为输入，将输出的[CLS]向量作为该话语片段的句向量，得到各个话语片段的表征*u=*{*u1,u2,…,un*}。再通过Bi-GRU来建模多个话语片段之间的序列关系，得到输出*g=*{*g1,g2,…,gn*}。（b）利用*gi*初始化关系图的*n*个节点，再通过关系图卷积网络(Relational Graph Convolutional Network, R-GCN)[45]来建模话语片段之间的语篇关系，得到语篇关系隐含表征*h=*{*h1,h2,…hn*}。（c）最后将*H*和*h*融合，预测答案的范围。

1. 说话者感知增强模型

Ma等人[39]则提出了一种说话者感知增强模型。将预训练语言模型作为上下文编码器，得到上下文表征矩阵后，引入说话者图、语篇关系图、说话者掩码自注意力机制三个模块。

构建关系图时，从*H*中抽取相应的部分初始化各个结点，同时表示全局对话的结点与所有结点相连，各个结点均设置了一条自循环的边。在说话者连接图中，结点之间的边由各个话语片段的说话者是否相同决定，相同的说话者之间通过一条双向边连接，具体形式见图2.6；在语篇关系图中，结点之间的边则由语篇关系注释决定，通过话语片段结点单向连接至对应的语篇关系结点，再由语篇关系结点连接至另一话语片段结点，具体形式见图2.7。两部分都使用R-GCN进行特征提取，分别得到输出的上下文表征*HS*和*HG*。



图2.6 说话者连接图



图2.7 语篇关系图

此外，该模型将MDFN[37]中提出的掩码注意力机制扩展至多方对话，通过添加不同的掩码(公式)以分别捕获不同说话者之间(M1)、相同说话者之间(M2)的信息交互如图2.8所示，计算说话者之间的注意力表征*HC*，如公式和所示。







最后，将*H*、*HC*、*HS*和*HG*融合，进行最终答案的预测。

### 2.3.2 基于多任务学习的对话建模方法

虽然通过关系图网络建模对话比较直观，但这样的需要额外的数据注释，例如语篇依赖关系注释；并且关系图网络较为复杂还会导致计算速度较慢。对此，针对抽取型问答任务，SUP (Speaker and Utterance Prediction) [40]引入多任务学习，提出了自监督的说话者预测任务和伪自监督的关键话语片段预测任务。SUP模型同样使用预训练语言模型作为编码器获得上下文表征矩阵*H*，之后通过子任务的损失函数反向传播来优化对话上下文的说话者信息蕴含表征*HS*和关键话语片段信息蕴含表征*HK*，从说话者和话语片段两个层级隐式的建模对话文本中的复杂信息流。



图2.8 说话者注意力掩码

1. 自监督的说话者预测任务

在多方对话中含有多个说话者，考虑到发言顺序的不确定性，准确的建模说话者之间交互的信息流对理解整个对话文本具有重要意义。如图2.6，在自监督的说话者预测任务中，随机从对话上下文中选择一个话语片段*Ui*将其中的说话者名称*Si*掩盖住，模型则需要预测出该说话者的姓名。训练模型时通过该子任务损失函数的反向传播来优化说话者信息蕴含表征*HS*，隐式建模说话者之间的信息流。

1. 伪自监督的关键话语片段预测任务

对于抽取式机器阅读理解任务，问题的答案是对话上下文中的一个连续片段。由于对话文本的特殊性，答案一定来自于其中一个话语片段*Ui*。因此，如图2.9所示，对于一个可回答的问题*Qi*，在中一定存在一个对应的关键话语片段*UK*。相比于从整个对话上下文中找出正确答案的片段，从*N*个话语片段中找到关键话语片段*UK*相对要容易一些。所以，在训练集中根据真实答案来标记关键话语片段，训练模型时通过关键话语片段预测子任务损失函数的反向传播来优化关键话语片段信息蕴含表征*HK*，建模问题和上下文之间的联系。

在本文所提出的多方对话机器阅读理解模型中，将同样使用自监督的说话者预测任务以及伪自监督的关键话语片段预测任务来辅助模型建模对话。



图2.9 自监督的说话者预测任务和伪自监督的关键话语段预测任务。其中下划线标注的话语片段为问题对应的关键话语片段

### 2.3.3 基于外部知识注入的阅读理解增强方法

在许多之前关于机器阅读理解方法的研究中，引入外部知识的方法已经被证明能够有效提高模型对问题和上下文之间联系的理解能力，文献[46]也提到，相关知识的交互式检索有助于减少阅读上下文时产生的“幻觉”。迄今为止，多种类型的知识增强都显著提升了模型针对对话文本和问题的理解能力，例如常识性知识、知识图谱、句法结构注释、以及在2.3.1中提到的语篇关系注释等。

具体而言，文献[47]所提出的方法引入了“共同指代关系”[48]和“实体间的语义关系”[49]，在词的层级上将对话文本构建为关系图，再基于R-GCN进行知识整合。Knowledge Reader[50]则将常识性知识[51,52]和上下文表征结合，提升模型再完形填空型阅读理解任务中的性能。文献[53]则从知识图谱ConceptNet[52]中抽取与对话数据集相关的知识项三元组{关系，原因，结果}，例如{导致，病毒，疾病}。之后针对不同的对话上下文和问题，匹配相关的多个知识项将其融合编码，让表征更加细化，改进对答案的预测。

## 2.4 本章小节

本章主要内容是任务定义与相关技术描述，包括面向对话文本的抽取式阅读理解任务定义和测评，预训练语言模型，多方对话文本建模策略。该部分是论文的基础，同时为后文研究方法的介绍进行铺垫。

# 基于关键话语片段交互增强的对话机器阅读理解方法

## 3.1 引言

在文本形式上，篇章段落型的文本大多以百科、新闻等语料为主，叙事方式则大多呈线性。但多方对话文本不同，在一段多方对话文本中包含两个以上说话者，每个说话者的发生目的以及说话方式都有差异，导致多方对话文本内容更加混乱；此外由于发言顺序的不确定性，各个说话者间信息流的交互也更加复杂。此外，针对抽取型的问答任务，要从充满噪声的对话上下文中准确的找到问题的答案也更具挑战。在2.3.2中，SUP[40]模型提出了自监督的说话者预测任务和伪自监督的关键话语片段预测任务证明，通过子任务损失函数的反向传播来帮助模型更好的建模对话上下文中各个说话者间的信息流交互以及上下文和问题之间的联系，在本文中将沿用这两个子任务，采用多任务学习的方式建立模型。

关键话语片段是问题的答案在话语片段层级的体现，其蕴含了答案的位置信息，以及对话上下文和问题之间丰富的交互信息，仅仅通过子任务损失函数的反向传播优化编码器输出的上下文表征，关键话语片段的信息并没有被充分利用。例如在图3.1中，尽管模型预测的答案中包含了正确答案*rhythmbox*，但仍然受到下一个话语片段中*call*的影响，最终给出的答案范围不够精确。此外，考虑到说话者预测任务和关键话语片段预测任务是分别从说话者和话语片段两个层级来解耦对话信息，分别优化说话者信息蕴含表征*HS*和关键话语片段信息蕴含标准*HK*，在信息融合前，应该尽可能的减少两个表征之间的信息重叠以增强信息提取的效率。

针对上述两个问题，本文提出了基于关键话语片段交互增强的机器阅读理解模型。相较于直接从上下文中抽取答案，在选择式的问答任务中，问题和若干个选项绑定后为阅读者提供线索，降低了难度。关键话语片段与选项的作用类似，也能够起到为阅读者提供线索的作用。因此，本文将子任务预测得到的关键话语片段作为先验信息，将其与问题拼接后得到“问题+关键话语片段”的形式。之后，为了模拟人类阅读理解时换位思考的过程[10]，本文将双向多头共同共同注意力机制(Dual Multi-head Co-Attention, DUMA)[54]扩展至多方对话任务中，带着线索从“对话上下文=>问题”和“问题=>对话上下文”两个角度，捕获对话上下文、问题、关键话语片段三者之间的关键信息交互，这一部分得到的交互感知表征*Hqa*将用于协助最终答案的预测。此外，我们引入最小化互信息(Mutual Information, MI)的方法，将关键话语片段信息蕴含表征*HS*和说话者信息蕴含表征*HK*之间的互信息最小化，借由互信息的约束来减小两部分之间的信息重叠和相互影响，以增强两个辅助任务学习到的特征并提高两部分对最终答案推理的贡献。



图3.1 错误预测示范

本章工作主要内容如下：

(1) 提出了关键话语片段信息交互增强模型，利用多任务学习建模多方对话，再将关键话语片段预测子任务的结果作为先验信息，采用双向多头共同注意力机制从“上下文=>问题”、“问题=>上下文”两个角度增强对话文本、问题、关键话语片段之间的关键信息交互，指导精确答案的预测。

(2) 利用最小化户信息的方法，减小两个子任务之间的信息重叠，增强其对最终答案推理的贡献。

(3) 通过在Molweni和FriendsQA数据集上的对比实验和分析实验，验证了所提出的方法针对多方对话文本抽取式阅读理解任务的有效性。

## 3.2 基于关键话语片段交互增强的对话机器阅读理解模型

本章所提出模型的整体架构如图3.2所示，模型分为四部分，分别是对话上下文编码器、说话者预测模块、关键话语片段预测模块、答案抽取模块。

### 3.2.1 对话上下文编码器



图3.2 模型的整体架构

在下游任务上微调预训练语言模型 (PLM) 以利用其强大的表征能力已成为目前NLP任务中的常见范式，因此本文中将分别使用BERT和ELECTRA两种PLM作为对话上下文编码器。

对于给定的对话上下文，其中。以及问题中的一个问题*Qi*，将编码器的输入处理为文本长度为*L*的序列：



其中[CLS]和[SEPn] ()，为用作分隔各个话语片段和问题的特殊字符。在下文中，[SEPn]也将被用作表示各个话语片段的句向量。在PLM中，为了将离散标记（或单词）转换为密集向量以进行深度学习模型优化，嵌入技术(Embedding) 用于生成单词嵌入**，向量维度为*d*。最后，由*N*PLM层Transformer Block堆叠的PLM将*E*编码后得到输入序列*x*的上下文表征*H*：



### 3.2.2 双向多头共同注意力机制

人类进行阅读理解时通常会从“上下文=>问题”和“问题=>上下文”两个角度寻找答案，在本模型中，将双向多头共同注意力机制(DUMA)扩展至多方对话以模拟这一过程。DUMA基于多头注意力机制(MHA)，分别计算两个方向的注意力后再将其融合。DUMA的输入为上下文表征和问题表征，计算过程如下：



其中*M1*和*M2*分别表示“上下文=>问题”和“问题=>上下文”的注意力，mean为按*L*的方向取均值，Fusion为一种启发式匹配机制[55]：



### 3.2.3 说话者预测模块

为了从说话者(Speaker)的角度解耦对话信息，说话者预测模块通过自监督的说话者预测任务损失函数的反向传播来隐式建模多个说话者之间的信息流，获得蕴含说话者信息的上下文表征。该任务通过二分类的方式实现，对于给定的对话上下文，随机选择其中一个话语片段将其中的说话者名字遮罩，记作*Umask*， 。这意味着在训练过程中，其对应的说话者姓名*Smask*对模型来说将不可见。之后，将*Umask*与其他*n-*1个未处理过的话语片段*Uunmask* ()一一组合为*n*-1对(*Umask*, *Uunmask*)，模型需要判断话语片段*Umask*和*Uunmask*中的说话者是否为同一人，即是否*Smask*= *Uunmask*。

实现细节如图3.3所示，将预训练语言模型编码器输出的上下文表征*H*从计算图中分离以防止说话者标签泄露，得到*Hdetach*后添加新的注意力掩码，通过*N*sMHA个堆叠的多头注意力层(记作sMHA)来捕获说话者间的信息流交互，获得蕴含说话者信息的上下文表征*HS*：



图3.3 关键话语片段预测模块和说话者预测模块



之后从*HS*中分离出和*N*-1个以表示每一个话语片段。针对每一对(*Umask*, *Uunmask*)，对应的和按维度拼接后被送入由全连接层和sigmoid激活函数构成的分类器。最后计算*N-*1对话语片段组合的二元交叉熵损失函数：



损失函数的梯度将反向传播，优化蕴含说话者信息的上下文表征*HS*。

### 3.2.4 关键话语片段预测模块

为了从话语片段(Utterance)角度解耦对话信息并建模问题和上下文之间的关系，关键话语片段预测模块采用伪自监督的说话者预测任务损失函数的反向传播来优化蕴含关键话语片段信息的上下文表征*HK*。与*HS*需要从*H*分离后再经过*N*sMHA个堆叠的多头注意力层建模不同，*HK*即为PLM编码器的输出*H*，在训练过程中直接由关键话语片段预测子任务的损失函数反向传播优化。

实现细节如图3.3所示，将上下文表征分解为对话上下文表征和问题表征，。通过DUMA来计算双向交互特征向量：



之后，与说话者预测模块类似，从中收集*N*个，表示*N*个话语片段。最后，将分别与各个按维度拼接，再将*N*个按列组合，送入由全连接层和softmax激活函数组成的分类器并计算交叉熵损失函数：



损失函数的梯度将反向传播，优化蕴含关键话语片段信息的上下文表征*HK。*

### 3.2.5 基于关键话语片段的信息交互

为了充分利用3.3.4中子任务得到的关键话语片段*UK*，本文将其作为先验信息，将其与问题拼接成“问题-候选答案”的形式，通过DUMA来捕获对话上下文、问题、关键话语片段三者之间的信息交互，为精确答案的预测提供线索指导。

实施细节上，从*H*中分离出关键话语片段对应部分的词表征片段，将其与问题表征按*L*方向拼接，记作。最后，计算后将其重复*L*次，得到信息交互表征*HQA*：



### 3.2.6 基于最小化互信息的信息解耦

此外，为了更好地从话语片段和说话者两个层级将对话的特征信息解耦，模型训练时在*HS*和*HK*之间引入了最小化互信息(MI)的约束。MI 衡量一个变量对另一个变量的信息量。 不幸的是，MI在连续和高维的情况下很难计算。根据参考文献[56,57]，MI 的计算可以近似为具有异方差均值和方差的变分分布 log q(*HK*|*HS*)，具体计算过程为：



其中，表示变量*HS*的均值，通过线性变换参数化。则表示其方差，通过softplus激活函数保证其为正值:



其中 δHs ∈ R 是训练过程中要优化的标量参数，ε > 0 是为数值稳定性引入的最小方差。最后，为了最小化互信息，将该部分的损失函数定义为：



### 3.2.7 基于范围的答案抽取

考虑到并不是每一个答案都能从上下文中找到合理的答案，因此对于含有不可回答问题的数据集，还需要判别问题是否可被回答。对此，我们将中得到的“上下文-问题”双向交互特征向量*HA*作为输入，利用一个全连接层和sigmoid激活函数进行问题可回答判定，计算二元交叉熵损失函数如公式，表示真实值，*pa*表示预测值。



对于可回答的问题则在整个序列上分别预测答案的开始位置*pstart*和结束位置*pend*。首先，将三种蕴含不同信息的表征*HS*、*HK*、*HQA*融合得到最终用于计算的增强上下文表征*Henhance*：



之后，通过两个全连接层和softmax激活函数分别计算*start*和*end*位置在对话中的概率分布如公式，其中和表示可训练参数。



最后，计算交叉熵损失函数如公式，其中*t*表示真实值，*p*表示预测值。



最终，模型的整体损失函数如下：



## 3.3 实验分析

### 3.3.1 数据集与评价指标

为了验证所提出模型的有效性，本节采用了2.1.2中介绍的多方对话问答数据集Molweni和FriendsQA进行实验，详细描述如表2-1。在模型性能的评估上，则采用2.1.3中设定的两种评价指标 EM和*F*1对模型的性能进行评价。

表3-1：实验中的超参数设置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 超参数 | Molweni | | FriendsQA | |
| BERT | ELECTRA | BERT | ELECTRA |
| Learning rate | 1.2e-5 | 6e-6 | 8e-6 | 8e-6 |
| Max sequence length | 390 | 390 | 512 | 512 |
| Batch size | 4 | 2 | 4 | 1 |
| *N*sMHA | 3 | 3 | 3 | 3 |

### 3.3.2 实验设置

在本节中，将介绍本章所提出模型的实验设置细节。在数据预处理阶段，根据训练集中问题对应的答案标记出关键话语片段，作为模型训练中关键话语片段预测子任务的标签数据。同样的，对于说话者预测子任务，同样在数据预处理阶段在训练集中针对每一条数据，随机选择一个话语片段的说话者，并标记其与其他几个话语片段说话者之间的关系，用作说话者预测子任务训练时的标签数据。

实验中主要使用的库版本为Pytorch 1.7.1[58]和Transformers 3.5.0[59]。为了与之前的研究保持一致，本实验采用BERT和ELECTRA作为对话上下文编码器。其中，BERT使用默认设置的bert-base-uncased作为预训练模型检查点，由12层Transformer Block组成，隐藏向量维度*d*为768。ELECTRA则默认设置选择electra-large作为预训练模型检查点，由24层Transformer Block组成，隐藏向量维度*d*为1024。模型训练时采用AdamW[60]优化器，具体的参数设置如表3-1所示。

### 3.3.3 基线模型

在对比实验中，我们选取了一些最近几年针对多方对话的机器阅读理解模型如下：

**PLM：** 我们采用 BERT-base[22]和 ELECTRA-large[28]作为baseline模型。通过PLM编码得到对话上下文表征*H*后，直接通过全连接层提取问题的答案。

**DADgraph**[36]**：**DADgraph使用BERT获取对话上下文的表征，但并不对其进行微调。之后，DADgraph使用语篇依赖链接和语篇关系将对话上下文构建为关系图，通过R-GCN进一步建模。

表3-2：所提出模型在Molweni和FriendsQA数据集上的对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Molweni | | FriendsQA | |
| EM | *F*1 | EM | *F*1 |
| BERTbaseline | 45.8 | 60.2 | 43.3 | 59.3 |
| BERTDADgraph | 46.5 | 61.5 | - | - |
| BERTUML+UOP | - | - | 46.3 | 63.1 |
| BERTWord Graph | - | - | 46.4 | **64.3** |
| BERTSUP | 49.2 | 64.0 | **46.9** | 63.9 |
| BERTours | **51.1** | **64.7** | 45.5 | 62.5 |
| ELECTRAbaseline | 56.8 | 70.6 | 52.8 | 70.1 |
| ELECTRASUP | 58.0 | 72.9 | 55.8 | 72.3 |
| ELECTRAours | **58.7** | **73.1** | **57.1** | **73.0** |

**SUP**[40]**:** SUP采用多任务学习策略，在说话者和话语片段两个层次引入自监督的说话者预测任务和伪自监督的关键话语片段预测子任务，通过子任务损失函数的反向传播来优化特定信息表征，隐式地对说话者信息进行建模并捕获问题答案的线索。

**UML+UOP**[41]**:** 该模型先使用话语片段级别的掩码任务(Utterance level LM, ULM) 和话语片段顺序预测(Utterance order prediction，UOP)任务对BERT进行预训练，在针对对话MRC任务进行微调。

表3-3：所提出模型在Molweni数据集上的消融实验结果

|  |  |  |
| --- | --- | --- |
| Model | EM | F1 |
| BERTours | 51.1 | 64.7 |
| w/o key-utterance-based Interaction | 50.6 | 64.1 |
| w/o Minimize MI | 50.7 | 64.5 |

**Word Graph**[47]：该模型基于图进行知识集成，引入了共指知识和实体关系知识注释，将对话上下文从单词级别构建为关系图，随后由 R-GCN 编码。

### 3.3.4 对比实验

为了验证所提出模型的有效性和泛化性，我们在Molweni和FriendsQA两个面向多方对话的机器阅读理解数据集上进行对比实验中，并且分别采用基础的预训练语言模型BERT-base以及性能更强大的ELECTRA-large作为对话上下文编码器，进行了4组对比实验结果如表格3-2所示。

在以BERT-base作为骨架网络时，由于没有为多方对话机器阅读理解任务设计特殊的交互策略，baseline模型的结果不够理想。通过不同方法将对话上下文构建为关系图，再借助R-GCN网络建模的方法DADgraph和Word Graph则在性能上有一定的提升，但这需要额外的数据注释，且网络较为复杂。最后，基于多任务学习的SUP模型则通过自监督的说话者预测任务、伪子监督的关键话语片段预测任务损失函数的反向传播来优化蕴含特定信息的上下文表征，隐式的建模说话者之间的关系并从话语段片的角度寻找答案的线索，这一方法与之前的模型相比性能显著提升。

本节中所提出的模型同样使用了SUP中的两项子任务，增加了对话上下文、问题、关键话语片段三者之间的信息交互，并通过最小化互信息的方式减少说话者和话语片段两个层级之间的信息重叠，有效提升了模型的性能，在Molweni数据集上均优于基线模型。然而，在FriendsQA数据集上进行的实验中，ELECTRA-large作为骨架模型时性能有所提升，但BERT-base作为骨架模型时性能有所下降。这是由于BERT-base模型属于基础模型，性能有限，但FriendsQA数据中的对话较长，含有较多的话语片段，关键话语段预测子任务的准确率不够，导致模型的最终性能下降。

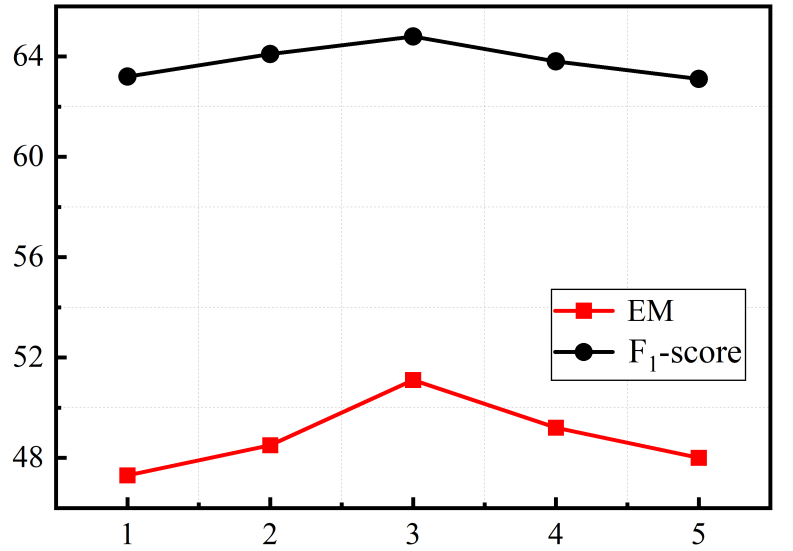


图3.4 *N*sMHA对模型性能的影响

### 3.3.5 消融实验

本节在Molweni数据集上进行消融实验，以验证本章所提出的基于关键话语片段的信息交互策略和基于最小化互信息的信息解耦策略的效果。具体而言，针对基于关键话语片段的信息交互策略，消融实验中将其移除并替换为SUP中的设置，即将上下文和问题的表征结合以预测关键话语段并通过其损失函数的反向传播来优化上下文表征。针对基于最小化互信息解耦策略，消融实验中则将该限制移除。消融实验结果如表格3-3所示，两个模块均对最终性能的提升有所贡献。

### 3.3.6 超参数分析

由于在说话者预测模块中多头注意力机制的层数*N*sMHA对模型的性能有一定影响，因此我们研究了使用BERT-base作为编码器时不同层数*N*sMHA对模型性能的影响，如图3.4所示。如图所示，当*N*sMHA为3时，模型的性能最佳。

### 3.3.7 案例分析

本小节选择了三个在Molweni数据集上进行预测的例子如图3.5所示，其中基线模型的预测结果与参考答案不完全匹配，所提出模型的预测与参考答案完全匹配，以直观地呈现除所提出模型的改进。

表3-4：案例分析

|  |
| --- |
| *Example 1*  …  ***Shamus:*** *i’m in FILEPATH now, but what do i do now ?*  ***Twohats:*** *the soundblaster is no longer listed pc-speaker is in its place*  ***DasEi:*** *look up proper dir name, then cd psp-whatever*  …  Question: What do I do when I’m in FILEPATH now?  Answer: look up proper dir name, then cd psp-whatever |
| BERTbaseline: the soundblaster is no longer listed  BERTSUP: look up proper dir name  BERTours: look up proper dir name, then cd psp-whatever |
| *Example 2*  …  ***mrwes:*** *and when you did a mkdir you used sharedfiles or server ?*  ***Doonz:*** *the extra must be something pastebin*  ***mrwes:*** *what fs is the share ? ntfs or ext3 ?*  ***Doonz:*** *i can connect to the share through places - connect to server*  …  Question: How can doonz connect to share ?  Answer: through places - connect to server |
| BERTbaseline: through places  BERTSUP: through places  BERTours: through places - connect to server |
| *Example 3*  …  ***Bacon5o:*** *i dont want ubuntu , it does n't support my internet , thus i can not use it*  ***morfic:*** *my ati has no aiglx support so i ca n't speak for how FILEPATH is*  ***morfic:*** *your internet is different from mine ? damn bush and his internets !*  …  Question: Who is annoyed about different internets ?  Answer: morfic |
| BERTbaseline: -  BERTSUP: morfic: your internet is different from mine ? damn bush  BERTours: morfic |

在第一个例子中，根据问题和对话上下文，可以发现问题中的*I*即指*shamus*。 baseline只是简单地将*shamus*之后的第一句话作为答案。SUP和所提出的模型则准确定位到了答案所在的关键话语片段，但SUP给出的答案不够完整，只包含了第一个步骤。由所提出的模型则通过对话上下文、问题、关键话语片段三者之间的深度信息交互，给出了完全匹配的答案。

在第二个例子中，因为问题中包含了直接的人物名称*Doonz*以及对应的目的*connect to share*，所以更容易定位到答案所在的关键话语片段。然后，baseline和SUP都只得到了部分正确的答案，而所提出的模型凭借丰富的信息交互得到了准确的答案。

在第三个例子中，由于问题中需对指代关系进行判定，没有针对说话者信息流进行建模的baseline无法给出答案。SUP模型虽然定位到了*annoy*所对应*damn*及其说话者 *morfic*，但是预测出来的答案过于粗糙。然而，所提出的方法凭借于信息交互，以及说话者和话语段特征信息之间的解耦，给出了准确的答案*morfic*。

## 3.4 本章小结

通过自监督的说话者预测任务和伪自监督的关键话语片段预测任务能够有效的帮助模型从说话者和话语片段两个层面对多方对话上下文进行建模，但仅通过子任务损失函数的反向传播优化特定信息表征无法对问题、关键话语片段、对话上下文之间的信息交互进行有效建模。针对此，本章中将子任务预测的关键话语片段作为先验信息，采用双向多头注意力机制捕获对话上下文、问题、关键话语片段三者之间的关键信息交互，为最终答案的预测提供线索指导。并且，还在两个子任务分别优化的上下文表征中增加了最小化互信息的约束，以减小两部分之间的信息重叠，保证其对最终答案预测的贡献。通过充足的实验，验证了本章中所提出方法的有效性，并取得了优于基线模型的性能。此外，消融实验以及案例分析也直观的展示了所提出方法的有效性。

# 基于外部知识注入增强的对话机器阅读理解方法

## 4.1 引言

考虑到多方对话文本中复杂的信息流交互，本文在上一章中采用了多任务学习的方式，引入自监督的说话者预测子任务和伪自监督的关键话语片段预测任务来辅助模型从说话者和话语片段两个角度来解耦对话信息。并且，将子任务预测的关键话语片段作为先验信息与问题拼接成“问题-候选答案”的形式，利用双向多头注意力机制来捕获其与对话上下文之间的信息交互为最终答案的推理提供指导，在本章中，将沿用上述方法。

此外，在面向多方对话文本的抽取式问答任务中，除了针对对话文本信息的建模让模型能够充分理解对话上下文之外，提高模型基于上下文内容，实现问题到答案的逻辑推理能力同样重要。目前绝大多数的模型均采用预训练语言模型对上下文和问题进行编码，将其在大量文本中学习的通识文本表示转移至下游任务。然而，在预训练时使用的语料基本以百科类的篇章段落式文本为主，与对话文本有较大差异；并且在2.2.2中所提到的常用预训练任务MLM、NSP以及RTD都与抽取式问答任务有着显著的不同。因此，数据分布以及下游任务的差距导致直接在多方对话问答数据集上微调预训练语言模型模时，针对抽取式问答任务中逻辑推理能力的建模较弱，模型性能不理想。

针对上述两个问题，在本章中我们提出了一种面向多方对话的逻辑推理记忆网络(Memory Network with Logistic Inference, LIMN) 模型，着重于增强模型针对抽取式问答任务的逻辑推理能力。通过对多方对话文本的分析，我们发现虽然多方对话文本中各个说话者的发言具有一定的随机性，但其整体的叙述逻辑仍然是顺序的。例如，如图4.1所示的多方对话文本中，我们根据语篇依赖关系对其进行了注释，虽然话语片段之间的链接存在一定跳跃性，但整体的对话叙述顺序仍为顺序，与篇章段落型的文本叙述顺序一致，这也是文本型数据的一项重要特征。因此，考虑到 SQuAD 2.0[4]是面向篇章段落文本的抽取式问答数据集，数据量丰富且下游任务形式与本文所研究相同，在LIMN中将其作为外部问答知识注入，优化预训练语言模型编码器，并通过记忆编码器从中学习并记忆与问答相关的逻辑推理知识，提高模型针对多方对话问答任务的逻辑推理能力。



图4.1 通过语篇关系链接注释的多方对话文本。其中蓝色、红色、绿色的连接反别表示Question-answer pair、Clarification-question、Comment三种语篇关系。红色的字体则表示问题的答案。

LIMN分为“继续学习”和“微调”两个阶段，如图4.2所示。在“继续学习”阶段，将SQuAD 2.0作为外部知识注入，一方面利用该数据集丰富的数据继续训练预训练语言模型编码器让其更加适应抽取式问答任务；另一方面则通过基于Transformer的记忆编码器来学习并记忆基于问答任务的逻辑推理知识，用于后续多方对话问答中的推理。在“微调”阶段则采用多方对话数据集训练模型，记忆编码器的参数被冻结，将从纯文本问答数据集中学习的逻辑推理知识迁移至多方对话问答任务，增强模型的逻辑推理能力。同时LIMN也保存了上一章所使用的说话者和关键话语片段信息解耦模块以及基于关键话语片段的信息交互。

本章工作主要内容如下：

(1) 提出了一种面向多方对话问答任务的逻辑推理记忆网络(Memory Network with Logistic Inference, LIMN) 模型，通过将外部知识注入整合为模型的潜在记忆，增强模型处理抽取式问答任务时的逻辑推理能力。

(2) 通过在Molweni和FriendsQA数据集上的对比实验和分析实验，验证了所提出的方法针对多方对话文本抽取式阅读理解任务的有效性。并且，两个数据集上的评测结果显著提高，达到了最优性能(state-of-the-art, SOTA)。

## 4.2 面向多方对话的逻辑推理记忆网络模型

所提出的LIMN模型整体架构如图4.3所示，分为预训练语言模型编码器、推理记忆编码器、说话者和话语片段信息解耦编码器和多任务学习四个部分。

### 4.2.1 对话上下文编码器



图4.2 LIMN模型的两阶段训练模式

与上一章相同，分别采用预训练语言模型BERT和ELECTRA作为对话上下文编码器，将输入构建为：



通过嵌入技术得到单词嵌入**，再由PLM编码得到上下文表征*H*：



### 4.2.2 推理记忆编码器

研究自然语言文本下的概念和语义关系可以促进对复杂对话信息的理解和分析。尽管篇对话文本的信息流交互更复杂，但整体的叙述逻辑以及面向问答任务的逻辑推理过程仍然与篇章段落型的文本基本相同，故LIMN中引入大量的篇章段落型问答数据集作为外部知识注入以提升模型的针对问答任务的逻辑推理能力。因此，为了建模对话文本中问题到答案之间的推理关系，本节提出了推理记忆编码器(Inference Memory Encoder, IME)，由*N*IME层Transformer Block组成，进一步从对话上下文表征*H*提取蕴含逻辑推理信息的上下文表征*HI*：



在针对问答逻辑推理知识的“继续学习”阶段，为了确保这种上下文表征*HI*能在篇章段落文本和多方对话文本中均能捕获基于问答的逻辑推理信息，推理记忆编码器在纯文本问答数据集上使用抽取式问答任务进行训练。对于下游的抽取式多方对话问答任务，推理记忆编码器的参数被冻结，用于从对话上下文表示生成蕴含逻辑推理信息的上下文表征*HI*，并且基于 PLM 的对话上下文编码器和基于范围的答案抽取模块仍在改进和调整。通过这种方式，可以将LIMN从SQuAD 2.0问答数据集中学习的问答逻辑推理能力保留到面向多方对话的问答任务中。



图4.3 LIMN模型的整体架构

### 4.2.3 说话者和话语片段信息解耦编码器

之前的大量工作证明，将对话上下文从不同角度进行信息解耦将有效提高模型对其的理解能力[40,37,61,39]。因此本章中将保留从说话者和话语片段两个角度通过多任务学习的方式解耦对话信息。

与上一章相同，说话者和话语片段信息解耦编码器(Speaker & Utterance Decouple Encoder)基于说话者预测子任务以及关键话语片段预测子任务，分别获得蕴含关键话语片段信息的上下文表征*HK*，蕴含说话者信息的上下文表征*HS*，整体过程与3.2.3和3.2.4相同，如图3.3所示。

### 4.2.3 信息融合与交互

对话上下文表征*H*经过推理记忆编码器得到蕴含逻辑推理信息的表征*HI*，经过说话者和话语片段信息解耦编码器后得到了蕴含说话者信息的表征*HS*，蕴含关键话语片段信息的表征*HK*。首先将三种表征融合得到*HF*：



与第3章不同，考虑到基于关键话语片段的信息交互策略(3.2.5)能够实现从对话上下文、问题、关键话语片段三者之间的深度信息交互，且LIMN种引入了新的逻辑推理信息表中HI，因此先进行表征融合，再进行信息交互，借此促进三种不同信息表征的融合。从*HF*中分离出对话上下文、问题、关键话语片段，计算双向注意力：



再将*HF*与融合得到最终表征*Henhance*：



### 4.2.4 多任务学习策略

多任务学习策略与第3章类似，包含问题可回答判定、关键话语片段预测、说话者预测、基于范围的答案抽取四部分。

（1）关键话语片段预测

该子任务仅在“微调”阶段进行。如图3.3所示，从*H*中收集*N*个话语片段的句表征向量，分别与*HA*拼接，通过全连接层和softmax激活函数找到关键话语片段，计算交叉熵损失函数。

（2）说话者预测

该子任务仅在“微调”阶段进行。如3.3所示，从*HS*中收集1个说话者被遮罩的话语片段表征，*N-*1个未经处理的话语片段表征，将与*N*-1个分别拼接后通过全连接层和sigmoid激活函数判断说话者是否相同，计算二元交叉熵损失函数。

（3）问题可回答判断定

该子任务在“继续学习”和“微调”阶段均进行。从上下文编码器的直接输出*H*中分离上下文*HC*和问题*HQ*并计算，再通过全连接层和sigmoid激活函数进行判定，计算二元交叉熵损失函数。

（4）基于范围的答案抽取

该任务为主任务，在“继续学习”和“微调”阶段均进行。在“继续学习”阶段，直接使用蕴含逻辑推理信息的上下文表征*HI*作为输入，通过两个线性层和softmax激活函数分别预测答案的开始和结束位置，计算交叉熵损失函数。

在“微调”阶段，使用由*HI*、*HS*、*HK*、*HQA*融合得到的对话上下文表征*Henhance*作为输入，使用同样的两个线性层和softmax激活函数分别预测答案的开始和结束位置，计算交叉熵损失函数。

“继续学习”阶段的整体损失函数为：



“微调”阶段的整体损失函数为：



## 4.3 实验分析

### 4.3.1 数据集与评价指标

“继续学习”阶段，将篇章段落型问答数据集SQuAD 2.0作为外部问答知识注入，该数据集来源于维基百科，包含人物、电子产品、城市、宗教等不同主题，问题中含有不可回答的问题。其中，训练集根据442篇文章人工提出了约13万个问题，其中约4.3万个不可回答的问题；验证集根据35篇文章人工提出了约1.1万个问题，其中约5000个不可回答的问题。

表4-1：LIMN模型在实验中的超参数设置

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 超参数 | Molweni | | FriendsQA | |
| BERT | ELECTRA | BERT | ELECTRA |
| Learning rate | 1.4e-6 | 6e-6 | 1.2e-5 | 8e-7 |
| Max sequence length | 390 | 390 | 512 | 512 |
| Batch size | 20 | 8 | 16 | 4 |
| *N*sMHA | 3 | 1 | 3 | 4 |
| *N*IME | 3 | 3 | 3 | 3 |

“微调阶段”则继续采用多方对话问答数据集Molweni和FriendsQA，详细描述如表2-1。在模型性能的评估上，同样采用两种评价指标EM和*F*1对模型的性能进行评价。

### 4.3.2 实验设置

实验中主要使用的库版本为Pytorch 1.7.1和Transformers 3.5.0。数据预处理中，关键话语片段的标签以及说话者的随即遮罩与第三章相同。同样采用BERT-base和ELECTRA-large分别作为上下文编码器以验证所提出LIMN在不同强度预训练语言模型下的有效性。实验中的超参数设置如表格4-1所示。

### 4.3.3 基线模型

除3.3.3中所提到的基线模型外，本章额外增加了MaskAttn+Graph以及SUP+BiDeN两种基线模型。

**MaskAttn+Graph**[39]:该模型主要从说话者角度增强针对多轮对话内容的建模，除关系图网络分别构造基于对话上下文的语篇关系图和说话者关系图外，还通过掩码自注意力机制(公式2.10)捕获相同说话者和不同说话者之间的信息交互。通过三个通道获取具有不同特征息的上下文表征，融合后进行最终答案的预测。

**SUP+BiDeN**[38]：该模型引入了回溯阅读[62]的思想，在预训练语言模型编码器后添加基于掩码自注意力机制的双向对话编码器，分别从过去<=>现在、现在<=>现在、将来<=>现在三个角度来建模对话上下文。之后，引入混合专家模块(Mixture of Experts Module, MoE)[63]形成最终的对话历史表征。此外，该模型同样使用了自监督的说话者预测任务以及伪自监督的关键话语片段预测任务辅助对话上下文的建模。

表4-2：LIMN模型继续学习阶段在SQuAD 2.0验证集上的测评结果

|  |  |  |
| --- | --- | --- |
| Model | EM | *F*1 |
| BERT-base | 71.5 | 75.6 |
| ELECTRA-large | 86.0 | 89.5 |

**SUP+MI**[38,61]：本文于第3章中所提出的模型，通过双向多头注意力机制捕获对话上下文、问题、关键华语片段之间的信息交互，并结合最小化互信息减小说话者和话语片段层级的信息重叠。

### 4.3.4 对比实验

对比实验分为两组，分别采用BERT-base和ELECTRA-large作为对话上下文编码器。表 4-2 展示了在继续学习阶段，所提出的模型在 SQuAD 2.0 数据集上的性能。 表 4 -3则分别展示了微调阶段在Molweni和FriendsQA数据集上的比较结果。在SQuAD 2.0和Molweni、FriendsQA数据集上的比较结果也表明，面向多方对话文本的机器阅读理解任务比纯文本更具挑战性。

总的来说，以前的研究通过SUP在说话者和话语级别对对话进行建模有显著的性能提升。另一方面，MaskAttn+Graph模型中分别基于语篇关系链接、说话者关系链接构建关系图网络、利用掩码自注意力机制以捕获说话者层面的信息交互三种通道共同来建模对话文本进一步提高了模型的性能。然而，话语关系图需要额外的数据注释。虽然上述方法与baseline模型相比取得了显着改进，但这些方法在对话问答任务中对逻辑推理关系的建模较弱。

表4-3的对比实验结果表明，所提出的LIMN模型在两个基准数据集上都取得了最先进的结果。此外，无论是使用 BERT-base 模型还是更强大的 ELECTRA-large 模型作为模型主干，所提出的模型都显着提高了性能，这证明了我们方法的通用性。高性能结果的主要原因可能在于以下几点：1）所提出的LIMN模型通过SUP子任务将对话中的说话人和关键话语信息分离。 2）LIMN引入篇章段落型问答数据集作为外部知识注入，帮助模型学习和记忆问答相关知识，提高对话问答任务的逻辑推理能力。 3）基于关键话语片段的信息交互策略被用来有效地结合由多个辅助子任务解耦的表示，它侧重于对话上下文、问题和关键话语之间的深度信息交互。

表4-3：LIMN模型在Molweni和FriendsQA数据集上的对比实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Molweni | | FriendsQA | |
| EM | *F*1 | EM | *F*1 |
| BERTbaseline | 45.8 | 60.2 | 43.3 | 59.3 |
| BERTDADgraph | 46.5 | 61.5 | - | - |
| BERTUML+UOP | - | - | 46.3 | 63.1 |
| BERTGraph | - | - | 46.4 | 64.3 |
| BERTSUP | 49.2 | 64.0 | 46.9 | 63.9 |
| BERTSUP+MI | 51.1 | 64.7 | 45.5 | 62.5 |
| BERTMaskAttn+Graph | 49.7 | 64.4 | 47.0 | 63.0 |
| BERTSUP+BiDeN | 50.3 | 65.9 | - | - |
| BERTLIMN | **53.0** | **67.5** | **49.4** | **65.4** |
| ELECTRAbaseline | 56.8 | 70.6 | 52.8 | 70.1 |
| ELECTRASUP | 58.0 | 72.9 | 55.8 | 72.3 |
| ELECTRASUP+MI | 58.7 | 73.1 | 57.1 | 73.0 |
| ELECTRAMaskAttn+Graph | 58.6 | 72.2 | 58.7 | 75.4 |
| ELECTRALIMN | **59.5** | **74.5** | **60.5** | **75.9** |

### 4.3.5 消融实验

为了研究LIMN种每个模块的效果，本节采用ELECTRA-large作为上下文编码器，在Molweni和FriendsQA数据集上进行了消融实验。具体来说，为了验证引入外部问答知识的作用，我们删除了推理记忆编码器和外部知识注入。为了验证推理记忆编码器的影响，我们将其移除，但仍使用外部篇章段落型问答数据集对PLM和最终答案抽取层进行继续学习。此外，我们还验证了基于关键话语的交互策略和推理记忆编码器的冻结操作的影响。

如表4-4所示，实验结果表明每个部分均对所提出模型的性能有显着影响。引入篇章段落型的问答知识对增强模型的逻辑推理能力有显著提升，使用推理记忆编码器和冻结操作则可以让模型更好地学习和记忆这种推理能力。最后，基于关键话语的交互策略也提高了模型的性能。



**(a) FriendsQA**



**(b) FriendsQA**

图4.4 LIMN模型在验证集上进行的超参数微调实验

表4-4：LIMN模型在Molweni和FriendsQA数据集上的消融实验

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Model | Molweni | | FriendsQA | |
| EM | *F*1 | EM | *F*1 |
| ELECTRALIMN | **59.5** | **74.5** | **60.5** | **75.9** |
| w/o External knowledge injection | 58.0 | 72.9 | 55.8 | 72.3 |
| w/o Inference Memory Encoder | 58.7 | 73.1 | 57.1 | 73.0 |
| w/o key-utterance-based Interaction | 58.6 | 72.2 | 58.7 | 75.4 |
| w/o Freeze Operation | **59.5** | **74.5** | **60.5** | **75.9** |

### 4.3.6 超参数分析

在实验中，对结果产生较显着影响的超参数主要包括学习率(Learning rate)、说话者预测模块中堆叠MHA层的个数*N*sMHA和推理记忆编码器中Transformer Block的数量*N*IME。

在这一部分中，使用ELECTRA-large作为对话上下文编码器。考虑到*N*IME的数量会影响模型在SQuAD 2.0数据集上的继续学习，其初始值被设置为3。学习率和*N*MHA则在微调阶段通过网格搜索确定。之后，我们固定学习率和*N*MHA来探索*N*IME对模型性能的影响。在Molweni和FriendsQA的验证集上进行超参数微调实验的结果如图4.4所示。实验结果表明，学习率对结果的影响最为明显，*N*MHA则对敏感数更为敏感据集。因此，我们根据超参数分析实验结果选择合适的学习率和*N*MHA。由于*N*IME的个数对结果的影响比较稳定，所以其被统一设置为3。



1. **Molweni**



1. **FriendsQA**

图4.5 说话者和话语片段数量对模型性能的影响

### 4.3.7 说话者和话语片段数量对模型性能的影响

为了研究说话者和话语片段的数量如何影响对话的复杂性，我们在Molweni 和FriendsQA数据集上进行了分析实验。图4.5显示了LIMN在处理具有不同数量的说话者和话语片段的对话上下文时的表现。

如图4.5 (a) 所示，Molweni数据集中对话文本的大小变化较小，只有8-15个话语和2-9个说话者。说话者和话语片段的数量对模型性能的影响较小。与基线模型相比，所提出的LIMN模型在每个划分区间的性能都有显着提高。

FriendsQA数据集中对话文本的大小跨度较大。如图4.5 (b) 所示，随着说话者和话语片段数量的增加，所比较的三个模型都受到不同程度的影响。与baseline模型相比，LIMN模型和SUP模型受到的影响更小，因为它们通过子任务将对话与说话人和话语水平分离。此外，由于引入了外部知识来帮助LIMN模型学习并记忆针对问答任务的逻辑推理关系，因此LIMN在大多数范围上的表现更好。同时，由于逻辑推理能力的提升并增加了基于关键话片语的信息交互，当对话文本更复杂时，LIMN模型表现更好。

表4-5：LIMN模型在Molweni数据集上的案例分析

|  |
| --- |
| *Example 1*  …  ***BlueEagle:*** *i already pasted the two non-commented lines in that file .*  ***Techsupport:*** *permission denied lol cant save the file.*  ***BlueEagle:*** *that 's because it 's owned by root .*  ***Techsupport:*** *with the user that i 'm looged in as*  ***BlueEagle:*** *but you already know that , do n't you ?*  …  Question: Why can not it save the file ?  Answer: because it 's owned by root |
| ELECTRAbaseline: permission denied  ELECTRASUP: permission denied lol  ELECTRALIMN: because it 's owned by root |
| *Example 2*  …  ***slerder:*** *thanks . is dpkg for installing things ?*  ***ziroday:*** *yep , apt is a frontend to dpkg*  ***ziroday:*** *can you please do sudo apt-get install install pastebinit and then do cat FILEPATH pastebinit*  *ziroday: okay , you installed the nvidia driver .*  …  Question: What is the first step of ziroday's advice ?  Answer: do sudo apt-get install install pastebinit |
| ELECTRAbaseline: installed the nvidia driver  ELECTRASUP: do sudo apt-get install install pastebinit and then do cat FILEPATH pastebinit  ELECTRALIMN: do sudo apt-get install install pastebinit |

### 4.3.8 案例分析

为了直观地展示所提出模型的改进，我们选择了四个不同问题类型预测示例。 如图4.6所示，例1&2和例3&4分别从Molweni和FriendsQA数据集中提取。

例1是*Why*类型的问题。对于文题中文件保存失败的原因，baseline和SUP模型预测答案为*permission denied*，但这只是表面原因。所提出的LIMN模型准确地识别了*permission denied*背后的原因，即该文件归*root*用户所有，这体现了 LIMN的逻辑推理能力更强。

表4-6：LIMN模型在FriendsQA数据集上的案例分析

|  |
| --- |
| *Example 3*  …  ***Monica Geller:*** *And well , we probably should n't see each other anymore . I 'm sorry .*  ***Peter Becker:*** *Okay , yeah . I mean ... If that 's , if that 's really what you want , okay .*  ***Monica Geller:*** *Okay , bye . #NOTE#: ( She kisses him on the cheek , and he kisses her back on the mouth . )*  ***Peter Becker:*** *I 'm sorry things did n't work out ...*  …  Question: How does Monica kiss Peter ?  Answer: on the cheek |
| ELECTRAbaseline: on the cheek , and he kisses her back on the mouth .  ELECTRASUP: on the mouth  ELECTRALIMN: on the cheek |
| *Example 4*  …  ***Monica Geller:*** *You were the next caller five hours ago . You must be going crazy .*  ***Phoebe Buffay:*** *Nah . I kept myself busy . #NOTE#: ( Both Rachel and Monica walk into their bedrooms , stop , and come back into the living room with confused looks on their faces . )*  ***Phoebe Buffay:*** *Oh , okay , yeah . I put your stuff in her room , and her stuff in your room .*  …  Question: Where did Phoebe not put Rachel and Monica 's things ?  Answer: their bedrooms |
| ELECTRAbaseline: i put your stuff in her room , and her stuff in your room .  ELECTRASUP: in her room , and her stuff in your room .  ELECTRALIMN: their bedrooms |

例2是What类型的问题。baseline模型预测的结果偏差较大。SUP模型虽然准确地定位了关键话语片段，但它没有关注到问题中的*first step*，给出的答案包含两个步骤，不够精确，反观LIMN模型则预测了完全匹配的答案。

例3是How类型的问题。所有三个模型都找到答案所在的关键话语片段。但是，baseline模型无法继续区分问题中Monica和Peter之间的指代关系。SUP模型则将指代关系识别错误。LIMN模型正确识别*Monica*对应于*she*，*Peter*对应于*he*，并找到了正确答案。

例4是Where类型的问题。baseline和SUP模型未能准确理解问题种的*not put*。对于问题语义的不清晰识别是因为针对问答任务的训练数据较小。baseline和SUP直接根据*put*寻找答案，得到与参考答案完全相反的结果。相反，所提出的LIMN模型则正确地理解了问题并找到了准确的答案。

综合上述示例可以发现，所提出的LIMN模型能够更好地理解对话上下文中的复杂信息，例如引用关系、深层次逻辑推理等。它能够更好的揭示问题和上下文之间的关系，并推理出更准确的答案。

## 4.4 本章小结

在面向多方对话文本的抽取式机器阅读理解任务中，一方面需要针对对话上下文和问题进行充分建模以确保模型对上下文的理解，另一方面则需要注重以对话上下文为基础，从问题到答案的逻辑推理过程进行建模。针对此，本章提出了一种面向多方对话的逻辑推理记忆网络(LIMN)模型。LIMN模型通过引入与问答相关的外部知识注入，在继续学习阶段通过逻辑推理记忆编码器将其整合为模型的潜在记忆，增强模型针对抽取式问答任务的逻辑推理能力并将其保留至下游任务中。在对话上下文和问题的建模方面，LIMN模型则延用了第3章中采用多任务学习进行建模，进行对话上下文、问题、关键话语片段三者之间的信息交互为寻找最终答案提供线索。通过充足的实验，验证了LIMN模型的有效性，并且在对比实验中取得了最优的性能。此外，还提供了消融实验、分析实验以及案例分析，进一步验证了所提出方法的有效性。

# 总结与展望

## 5.1 论文的主要工作及贡献

机器阅读理解是自然语言处理中的一个重要领域，对于人工智能技术的发展有重大意义。相较于篇章段落式的文本，对话文本则更接近人类的日常交流，并且与人机对话更加契合。然而，对话文本相较于篇章段落式的文本而言结构更加复杂，更具挑战性。因此，本文的工作重点是面向多方对话文本的机器阅读理解模型，主要针对以下两个问题进行研究：

(1) 通过多任务学习的方式从说话者和话语片段两个层级将对话上下文进行信息解耦后，如何充分利用各部分解耦的信息，并针对对话上下文、问题、关键话语片段之间的信息交互进行建模，为最终答案的预测提供线索。

(2) 除了针对对话上下文进行建模以外，在机器阅读理解中针对问题=>答案的逻辑推理关系进行建模同样至关重要。针对问答任务中逻辑推理关系的建模较弱会导致答案的预测不够精确。

针对上述问题，本文的主要工作内容总结如下：

1. 提出了一种基于关键话语片段的信息交互增强模型。通过子任务找出答案所在的关键话语片段，将其作为先验信息。再利用双向多头注意力机制模拟人类双向思考的过程，捕获对话上下文、问题、关键话语片段之间的信息交互，为最终答案的预测提供线索指导。同时，为了保证从对话中高效的解耦说话者和话语段层级的特征信息，模型中通过最小化互信息的方法降低两部分之间的信息重叠。在Molweni和FriendsQA数据集上进行的实验证明了所提出方法的有效性。

2. 提出了一种整合外部知识以增强逻辑推理能力的逻辑推理记忆网络模型(LIMN)。通过引入外部问答知识并由逻辑推理记忆编码器学习并保存其中的相关知识，再将其转移至多方对话问答任务中。该方法将篇章段落型文本和对话文本的阅读理解任务进行有机结合，有效提高了模型的逻辑推理能力。在Molweni和FriendsQA数据集上进行的实验验证了所提出方法的有效性，并取得了当前最优的性能。

## 5.2 未来研究工作

本文围绕面向多方对话问的抽取式机器阅读理解任务展开了深入的研究，取得了一定的成果，但在实验过程中仍然发现了一些值得深入研究的地方：

(1) 针对对话文本的知识整合及预训练。在本文所提出的方法中，仅使用10余万条数据进行逻辑推理知识的整合，就让下游的对话机器阅读理解任务的准确率得到了显著提升。因此，未来工作可以通过整合针对机器阅读理解任务相关的语料库，构建更加庞大且高质量的预训练数据集，设计相应的预训练任务建立专门面向阅读理解任务的预训练语言模型。

(2) 中文对话抽取式机器阅读理解数据集。现有的抽取式机器阅读理解数据集多局限于英文数据，面向中文的抽取式对话机器阅读理解研究相对较少。因此，再未来工作中基于中文媒体平台，例如微博、知乎等构建抽取式对话机器阅读理解数据集显得极为重要。

(3) 机器阅读理解模型的压缩。在智能对话模型的实际应用中，通常需要模型在做到实时响应的同时保持较高的准确率。大部分研究都着眼于提升模型预测答案的准确率，通常采用参数非常庞大的预训练语言模型进行编码，再设计复杂的交互网络，训练过程中需要耗费极大的计算资源。因此，如何对庞大的机器阅读理解模型进行压缩，同时将其准确率保持在一个较高的水准也是一个重要的研究方向。

# 参考文献

[1] HERMANN K M, KOCISKÝ T, GREFENSTETTE E, et al. Teaching Machines to Read and Comprehend[C]//Advances in Neural Information Processing Systems 28: Annual Conference on Neural Information Processing Systems 2015 (NeurIPS-2015). Montreal, Quebec, Canada, 2015: 1693-1701.

[2] TURING A M. Computing Machinery and Intelligence[M]//The Philosophy of Artificial Intelligence. 1990: 40-66.

[3] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100, 000+ Questions for Machine Comprehension of Text[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP-2016). Austin, Texas, USA, 2016: 2383-2392.

[4] RAJPURKAR P, JIA R, LIANG P. Know what you don’t know: Unanswerable questions for SQuAD[C/OL]//proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018): Vol. 2. 2018.

[5] CHOI E, HE H, IYYER M, et al. QuAC: Question Answering in Context[C]//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (EMNLP-2018). Brussels, Belgium, 2018: 2174-2184.

[6] REDDY S, CHEN D, MANNING C D. CoQA: A Conversational Question Answering Challenge[J]. Trans. Assoc. Comput. Linguistics, 2019, 7: 249-266.

[7] LAI G, XIE Q, LIU H, et al. RACE: Large-scale ReAding Comprehension Dataset From Examinations[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP-2017). Copenhagen, Denmark, 2017: 785-794.

[8] LI J, LIU M, KAN M Y, et al. Molweni: A Challenge Multiparty Dialogues-based Machine Reading Comprehension Dataset with Discourse Structure[C]//Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020). Barcelona, Spain (Online), 2020: 2642-2652.

[9] CHEN D, BOLTON J, MANNING C D. A thorough examination of the CNN/daily mail reading comprehension task[C/OL]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016): Vol. 4. 2016: 2358-2367.

[10] SEO M J, KEMBHAVI A, FARHADI A, et al. Bidirectional Attention Flow for Machine Comprehension[C]//Proceedings of the 5th International Conference on Learning Representations (ICLR-2017). Toulon, France.

[11] XIONG C, ZHONG V, SOCHER R. Dynamic Coattention Networks For Question Answering[C]//Proceedings of the 5th International Conference on Learning Representations (ICLR-2017). 2017.

[12] DHINGRA B, LIU H, YANG Z, et al. Gated-Attention Readers for Text Comprehension[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL-2017). 2017: 1832-1846.

[13] WANG W, YANG N, WEI F, et al. Gated Self-Matching Networks for Reading Comprehension and Question Answering[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, (ACL-2017). 2017: 189-198.

[14] HU M, PENG Y, HUANG Z, et al. Reinforced Mnemonic Reader for Machine Reading Comprehension[C]//Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-2018). 2018: 4099-4106.

[15] YU A W, DOHAN D, LUONG M T, et al. QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension[C]//6th International Conference on Learning Representations, (ICLR-2018). 2018.

[16] WELBL J, LIU N F, GARDNER M. Crowdsourcing Multiple Choice Science Questions[C]//Proceedings of the 3rd Workshop on Noisy User-generated Text, (NUT@EMNLP-2017). 2017: 94-106.

[17] TRISCHLER A, WANG T, YUAN X, et al. NewsQA: A Machine Comprehension Dataset[C]//Proceedings of the 2nd Workshop on Representation Learning for NLP, Rep4NLP@ACL 2017. 2017: 191-200.

[18] DUNN M, SAGUN L, HIGGINS M, et al. SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine[J]. CoRR, 2017, abs/1704.0.

[19] HILL F, BORDES A, CHOPRA S, et al. The Goldilocks Principle: Reading Children’s Books with Explicit Memory Representations[C]//4th International Conference on Learning Representations, (ICLR-2016). 2016.

[20] PAPPAS D, ANDROUTSOPOULOS I, PAPAGEORGIOU H. BioRead: A New Dataset for Biomedical Reading Comprehension[C]//Proceedings of the Eleventh International Conference on Language Resources and Evaluation, (LREC-2018). 2018.

[21] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is All you Need[C]//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017 (NeurIPS-2017). Long Beach, CA, {USA}, 2017: 5998-6008.

[22] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2019). Minneapolis, MN, USA, 2019: 4171-4186.

[23] WANG A, SINGH A, MICHAEL J, et al. Glue: A multi-task benchmark and analysis platform for natural language understanding[C/OL]//Proceedings of the 7th International Conference on Learning Representations (ICLR-2019). 2019.

[24] WANG A, PRUKSACHATKUN Y, NANGIA N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS-2019): Vol. 32. Vancouver, BC, Canada, 2019: 3261-3275.

[25] LAN Z, CHEN M, GOODMAN S, et al. ALBERT: A Lite BERT for Self-supervised Learning of Language Representations[C]//Proceedings of the 8th International Conference on Learning Representations (ICLR-2020). Addis Ababa, Ethiopia, 2020.

[26] YANG Z, DAI Z, YANG Y, et al. XLNet: Generalized Autoregressive Pretraining for Language Understanding[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019 (NeurIPS-2019). Vancouver, BC, Canada, 2019: 5754-5764.

[27] LIU Y, OTT M, GOYAL N, et al. RoBERTa: A Robustly Optimized BERT Pretraining Approach[J]. CoRR, 2019, abs/1907.1.

[28] CLARK K, LUONG M T, LE Q V, et al. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators[C]//Proceedings of the 8th International Conference on Learning Representations (ICLR-2020). Addis Ababa, Ethiopia, 2020.

[29] CUI L, WU Y, LIU S, et al. MuTual: A Dataset for Multi-Turn Dialogue Reasoning[J/OL]. 2020: 1406-1416.

[30] LI Y, SU H, SHEN X, et al. DailyDialog: A Manually Labelled Multi-turn Dialogue Dataset[C]//Proceedings of the Eighth International Joint Conference on Natural Language Processing, (IJCNLP-2017). 2017: 986-995.

[31] WELLECK S, WESTON J, SZLAM A, et al. Dialogue Natural Language Inference[C]//Proceedings of the 57th Conference of the Association for Computational Linguistics, (ACL-2019). 2019: 3731-3741.

[32] SUN K, YU D, CHEN J, et al. DREAM: A Challenge Dataset and Models for Dialogue-Based Reading Comprehension[J]. Trans. Assoc. Comput. Linguistics, 2019, 7: 217-231.

[33] YANG Z, CHOI J D. FriendsQA: Open-Domain Question Answering on TV Show Transcripts[C]//Proceedings of the 20th Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL-2019). Stockholm, Sweden, 2019: 188-197.

[34] KIPF T N, WELLING M. Semi-Supervised Classification with Graph Convolutional Networks[C]//Proceedings of 5th International Conference on Learning Representations, (ICLR-2017). 2017.

[35] GHOSAL D, MAJUMDER N, PORIA S, et al. DialogueGCN: A Graph Convolutional Neural Network for Emotion Recognition in Conversation[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP-2019). Hong Kong, China, 2019: 154-164.

[36] LI J, LIU M, ZHENG Z, et al. DADgraph: A Discourse-aware Dialogue Graph Neural Network for Multiparty Dialogue Machine Reading Comprehension[C]//International Joint Conference on Neural Networks (IJCNN-2021). Shenzhen, China, 2021: 1-8.

[37] LIU L, ZHANG Z, ZHAO H, et al. Filling the Gap of Utterance-aware and Speaker-aware Representation for Multi-turn Dialogue[C]//Thirty-Fifth AAAI Conference on Artificial Intelligence (AAAI-2021), Thirty-Third Conference on Innovative Applications of Artificial Intelligence (IAAI-2021), The Eleventh Symposium on Educational Advances in Artificial Intelligence (EAAI-2021). Virtual Event, 2021: 13406-13414.

[38] LI Y, ZHAO H, ZHANG Z. Back to the Future: Bidirectional Information Decoupling Network for Multi-turn Dialogue Modeling[J]. CoRR, 2022, abs/2204.0.

[39] MA X, ZHANG Z, ZHAO H. Enhanced Speaker-aware Multi-party Multi-turn Dialogue Comprehension[J/OL]. CoRR, 2021, abs/2109.0.

[40] LI Y, ZHAO H. Self- and Pseudo-self-supervised Prediction of Speaker and Key-utterance for Multi-party Dialogue Reading Comprehension[C]//Findings of the Association for Computational Linguistics (EMNLP-2021). Virtual Event / Punta Cana, Dominican Republic, 2021: 2053-2063.

[41] LI C, CHOI J D. Transformers to Learn Hierarchical Contexts in Multiparty Dialogue for Span-based Question Answering[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL-2020). 2020: 5709-5714.

[42] LIU Y, FENG S, WANG D, et al. A Graph Reasoning Network for Multi-turn Response Selection via Customized Pre-training[C]//Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, (AAAI-2021). 2021: 13433-13442.

[43] LOWE R, POW N, SERBAN I, et al. The Ubuntu Dialogue Corpus: A Large Dataset for Research in Unstructured Multi-Turn Dialogue Systems[C]//Proceedings of the SIGDIAL 2015 Conference, The 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL-2015). Prague, Czech Republic, 2015: 285-294.

[44] PETERS M E, NEUMANN M, IYYER M, et al. Deep Contextualized Word Representations[C]//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2018). New Orleans, Louisiana, USA, 2018: 2227-2237.

[45] SCHLICHTKRULL M, KIPF T N, BLOEM P, et al. Modeling Relational Data with Graph Convolutional Networks[C/OL]//Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics): 10843 LNCS. 2018: 593-607.

[46] SHUSTER K, POFF S, CHEN M, et al. Retrieval Augmentation Reduces Hallucination in Conversation[C/OL]//Findings of the Association for Computational Linguistics, Findings of ACL (EMNLP-2021). 2021: 3784-3803.

[47] LIU J, SUI D, LIU K, et al. Graph-Based Knowledge Integration for Question Answering over Dialogue[C]//Proceedings of the 28th International Conference on Computational Linguistics (COLING-2020). Barcelona, Spain (Online), 2020: 2425-2435.

[48] CHEN H Y, ZHOU E, CHOI J D. Robust Coreference Resolution and Entity Linking on Dialogues: Character Identification on TV Show Transcripts[C]//LEVY R, SPECIA L. Proceedings of the 21st Conference on Computational Natural Language Learning, (CoNLL-2017). 2017: 216-225.

[49] YU D, SUN K, CARDIE C, et al. Dialogue-Based Relation Extraction[C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, (ACL-2020). 2020: 4927-4940.

[50] MIHAYLOV T, FRANK A. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL-2018). 2018: 821-832.

[51] MILLER G A, RICHARDBECKWITH, CHRISTIANEFELLBAUM, et al. Introduction to WordNet: An On-line Lexical Database\*[C].

[52] SPEER R, CHIN J, HAVASI C. ConceptNet 5.5: An Open Multilingual Graph of General Knowledge[C/OL]//SINGH S, MARKOVITCH S. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, {USA}. 2017: 4444-4451.

[53] ZHANG Z, LI J, ZHAO H. Multi-Turn Dialogue Reading Comprehension With Pivot Turns and Knowledge[J]. IEEE ACM Trans. Audio Speech Lang. Process., 2021, 29: 1161-1173.

[54] ZHU P, ZHANG Z, ZHAO H, et al. DUMA: Reading Comprehension With Transposition Thinking[J]. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2022, 30: 269-279.

[55] MOU L, MEN R, LI G, et al. Natural Language Inference by Tree-Based Convolution and Heuristic Matching[C]//Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL-2016). Berlin, Germany, 2016.

[56] AHN S, HU S X, DAMIANOU A C, et al. Variational Information Distillation for Knowledge Transfer[C]//IEEE Conference on Computer Vision and Pattern Recognition (CVPR-2019). Long Beach, CA, USA, 2019: 9163-9171.

[57] LI J, HE R, YE H, et al. Unsupervised Domain Adaptation of a Pretrained Cross-Lingual Language Model[C]//Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, (IJCAI-2020). 2020: 3672-3678.

[58] PASZKE A, GROSS S, MASSA F, et al. PyTorch: An Imperative Style, High-Performance Deep Learning Library[C]//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, (NeurIPS-2019). 2019: 8024-8035.

[59] WOLF T, DEBUT L, SANH V, et al. Transformers: State-of-the-Art Natural Language Processing[C]//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations (EMNLP-2020). Online, 2020: 38-45.

[60] LOSHCHILOV I, HUTTER F. Decoupled Weight Decay Regularization[C]//Proceedings of the 7th International Conference on Learning Representations (ICLR-2019). New Orleans, LA, USA, 2019.

[61] ZHU X, WANG J, ZHANG X. An Enhanced Key-utterance Interactive Model with Decoupled Auxiliary Tasks for Multi-party Dialogue Reading Comprehension[C]//2022 International Joint Conference on Neural Networks (IJCNN-2022). 2022: 1-8.

[62] SUN K, YU D, YU D, et al. Improving Machine Reading Comprehension with General Reading Strategies[C]//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2019). 2019: 2633-2643.

[63] JACOBS R A, JORDAN M I, NOWLAN S J, et al. Adaptive Mixtures of Local Experts[J]. Neural Comput., 1991, 3(1): 79-87.

# 在学期间研究成果

发表的学术论文

1. **Xingyu Zhu**, Jin Wang, and Xuejie Zhang, "An Enhanced Key-utterance Interactive Model with Decoupled Auxiliary Tasks for Multi-party Dialogue Reading Comprehension", in Proceedings of the 2022 International Joint Conference on Neural Networks (IJCNN-2022), pp. 1-8, 2022. (CCF-C类推荐会议，EI)
2. **Xingyu Zhu**, Jin Wang and Xuejie Zhang, "YNU-HPCC at SemEval-2021 Task 6: Combining ALBERT and Text-CNN for Persuasion Detection in Texts and Images", in Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021), pp. 1045–1050, 2021. (EI)
3. **Xingyu Zhu**, You Zhang, Jin Wang, Xiaobing Zhou and Xuejie Zhang, "Incorporating External Knowledge as Latent Memory for Multiparty Dialog Machine Reading Comprehension", Knowledge-Based Systems. 已投稿. (CCF-C类推荐期刊，SCI)

主持或参与的实践项目

1. 参与国家自然科学基金NSFC地区项目“基于多维度连续情感空间的文本生成技术研究”，批准号：61966038，起止时间 2020.1-2023.12
2. 参与国家自然科学基金NSFC地区项目“面向小语种的高性能文本情感分析关键技术研究”，批准号：61762091，起止时间：2018.1-2021.