# 摘要

近年来，机器问答已经成为自然语言处理领域的研究热点之一。该任务的研究成果可以直接反应出机器对于语言文本的理解能力，也是机器智能水平的直接体现。全领域的机器问答系统更加具有挑战性，它要求机器对于来自各个领域的问题都具有理解和回答能力。语义推理和语义表示是机器问答系统的两大重要构成，本文主要对基于语义推理方法和语义表示能力的研究，来进一步提升全领域机器问答系统的性能。

首先，对于文本的语义推理，给定文档、问题以及候选答案，传统的方式是直接将整个文档与问题候选答案组合进行相似度匹配，考虑的是句子与整个文档的相关性，缺少更加细粒度的隐层关系推理。事实上，文档中与问题相关的句子只有个别几句，其它非相关的句子都可能对答案推理产生干扰。隐变量模型可以较好的捕捉到文本的潜在信息，首先从文档中找到与问题的相关句，再在相关句上进行答案推理。而一个好的隐变量组织结构可以更好的发掘文本之间的潜在关系，本文提出了两种新的隐变量组织结构（依赖解析树和语义框架）来提高文本的深层次的语义推理能力，并在MCTest阅读理解数据中进行验证，取得当前领先水平。

其次，分布式词向量是文本语义的一种重要的数字化表示，它也是深度神经网络模型的主要输入特征。一般的做法是通过对文档和问题句子进行编码，得到相应的隐层表示，再通过对文档和问题的隐藏表示的相似度计算进行答案推理。这种方法缺少了文档与问题之间的交互表示。本文研究多种语义层次级别的交互式表示，并在SQuAD机器问答数据集进行实验验证，说明多层级的语义交互表示对于提高机器理解能力的有效性。

**关键词：全领域；机器问答；语义推理；语义表示；依赖解析树；语义框架；隐变量模型；神经网络**

# 绪论

## 研究背景和意义

随着互联网的不断发展，人们对机器机器智能化的需求日益增加，而网络中大量的文本信息成为了提高机器智能化水平的重要数据。自然语言处理技术的发展使得这些信息数据可以更好的转化到实际的应用当中，为人工智能的发展提供了无限的可能。近年来，全领域的机器问答系统已经成为自然语言处理研究的主要任务之一。该问答系统不局限于某一特定专业领域，它对于来自任何方面的问题都能给予回答。该任务的研究成果可以直接反应出机器智能化的发展水平。

目前，在互联网上机器问答的应用也随处可见，如搜索引擎中的知识搜索，问答机器人等。但大部分的机器问答应用还是基于知识检索和规则匹配方法实现的常识问答，它们将外部知识的表示和利用置于研究的中心位置。但为了更好的检测机器对语言的理解能力，对于具有受限的外部知识（不利用或少利用外部知识）的机器阅读理解任务的研究才更具有深远有意义，它不需要人工进行过多的规则干涉而能够从语义表示层面上进行问题推断。

## 机器阅读理解的国内外研究现状

机器问答已经逐渐成为自然语言处理领域的研究热点。机器问答系统的研究有两大主流方法，一种是基于知识检索，一种是基于语义理解。一个完整的全领域机器问答系统一般是由这两部分共同组成。而基于知识检索的研究缺少灵活性，已经逐渐呈现饱和趋势。目前，大部分的机器问答任务都是把研究注意力放在了机器阅读理解任务上。

目前，用于阅读理解任务研究的数据集种类很多。比如，CBT和CNN/Daily Mail是做完形填空任务的。Algebra是做代数运算任务的。bAbI是做问题逻辑推理任务的。Science exam是做科学问答单选任务的。SQuAD是一个大规模的基于WikiPedia的全领域阅读理解数据集，该数据集中的每个问题是没有候选答案的，答案是文章中的一个短语或片段。TriviaQA是一个与SQuAD数据结构类似，但更加具有挑战性的全领域机器阅读理解数据集，它具有更加棘手的问题，并且需要跨多个句子进行问题推理，答案不限于文档中的连续片段。MCTest是做全领域的单项选择的任务的，它的文章是由人工虚构而成的高质量故事。除此之外，对于中文的机器阅读理解研究目前也有很多新兴的任务，如百度的DuReader阅读理解任务。现在基于该数据集的工作主要可以分为两类：1.基于传统特征工程的方法2.基于深度学习的方法。

MCTest数据集是由Richardson提出的，并且他在文中提出了两种基准方法。第一种是基于词袋滑窗的方法，他将问题句与候选答案句进行字符串拼接形成假设句，然后计算假设句与文章中句子的词重叠程度，选择具有最高词重叠的答案作为预测输出。第二种是基于词距的方法，计算问题词与答案词在文章中出现的最近距离。Smith之后对基于词袋滑窗的基准方法进行了改进，他设置不同大小的窗口长度来捕捉文章与问题更加全面的相关信息，从而进一步提高了问答准确率。简单的滑窗模型对于回答单一支持句的问题已经可以达到不错的效果，但对于回答多支持句的问题还是显得乏力。Narasimhan和Barzilay使用了隐变量模型来指示问题句的候选支持句，并在模型中加入了话语关系特征。Wang在他的隐变量模型中加入了句法、语义和框架结构特征，同时又引入了已经训练好的词向量信息。Sachan提出了一种隐变量模型来学习潜在的答案蕴含结构，他在模型中引入了名词实体的指代关系链以及句子之间的释义关系链来进行隐变量的指示，同时他又在模型中加入了语义相似特征以及字符串编辑特征来更加细化的进行文本识别与匹配。

随着近些年深度学习框架的不断开发，CNN、RNN、LSTM以及注意力机制和记忆力机制也被广泛地应用到了机器阅读理解的任务中来，并在大规模语料的数据集中取得了不错的性能效果。Weston提出了Memory Network机制，将文档内容作为外部知识数据进行存储，将输入的问题咋在外部记忆框架中进行推理找到与问题相关的支持句，再在支持句中找到问题的答案。该模型在许多逻辑推理任务中都取得了不错的性能效果（例如bAbI任务）。之后，Sukhbaatar又提出了一种End-To-End Memory Network模型，解决了Memory Network中需要额外标注支持句标签的问题。Herman首次提出了基于机器阅读任务的注意力机制，加入了文档与问题的交互式表示，增强了文本的语义表示能力，并在CNN/Daily Mail任务中取得了不错的效果。Hill提出了基于窗口的记忆力网络机制，并在CBT数据集上取得了不错的性能效果。Seo提出了一种双向的注意力机制，同时学习到文档对于问题的表示和问题对于文档的表示。Xiong提出了一种Co-Attention机制来对学习文档的叠加式注意力表示，并且加入了强化学习的损失函数，使得学习目标与真实回答目标更加一致，并在SQuAD数据集中取得了领先水平。而对于小规模机器阅读数据集（如MCTest），深度学习的方法并不能得以很好的发挥。Yin提出一种基于注意力机制的卷积神经网络模型来解决这一问题，他通过卷积和池化操作分别得到句子层面和篇章层面的表示。最终通过高速通道网络获得最终的整体表示。但该模型的最终实验效果并不理想。在MCTest数据集上深度学习具有较好发挥效果的有Trischler提出的基于多层面相似度计算的神经网络模型，分别从词层面和句子层面进行编码。但该模型存在一个局限问题是，模型的初始化参数都是通过经验设置，而且参数的初始化设置对于最终模型的效果是十分重要的。如果是仅仅随机初始化并不能达到理想的实验效果。Wang使用RNN模型对句子进行建模表示，并另外结合使用了外部的答案选择模型和蕴涵关系推理模型。他的整个模型是将神经网络模型与传统特征相结合，并且使用了外部的知识数据（Wikipedia）进行模型训练。可以看出，如果仅仅使用神经网络模型，没有先验知识与外部资源的支持，是很难在小规模语料数据集上得到理想结果的。

## 与其他相关任务的联系

自然语言处理研究任务可以按由低到高的层级关系划分为如图1所示的体系结构。其中，词性标注任务是确定句子中各个单词的词性，如名词、动词和形容词等。它是自然语言处理领域中的底层任务。句法分析的任务是要确定句子的

**词性标注**

**句法分析**

**语义分析**

**上层应用**

图 1 自然语言处理任务层级关系简图

句法结构，如主语、谓语、宾语和定语等等。语法分析的任务是要抽取出句子的语法组织结构，如一个行为的发出者是谁，发生了什么行为，行为的施受者是谁。上层应用可以看作是低层任务的综合体现，如情感分析、机器翻译、机器问答和聊天系统等。

机器问答在这个体系结构中处于高层次级别，它依赖于低层任务的实现，如单词的POS Tag、命名实体标注和句子的语义框架抽取都可以作为机器问答系统的重要特征，对系统的性能提升有着不可忽略的影响。在上层应用中，机器问答和聊天系统有着较高的相似性，但相比与聊天系统，机器问答系统是专注于对于输入问题的回答，而聊天系统的任务是要对输入话语进行自然的对话，输入的话语不限于一个问题。

## 论文主要研究内容和成果

本文主要研究内容如下：

1. 传统特征工程在小语料阅读理解任务上的实用性
2. 隐变量模型对于问题推理的帮助价值
3. 不同隐变量的组织结构对于问题推理的优势
4. 隐变量组织结构中高效的推理算法
5. 神经网络模型在大语料阅读理解任务上的优势
6. 多层级的注意力机制对于机器阅读系统性能的提升
7. 增强的语义表示对于机器阅读系统性能的提升

本文主要成果如下：

1. 提出了两种隐变量模型组织结构增强机器阅读的推理能力
2. 采用了一种高效的推理算法用于本文提出的隐变量模型的解码
3. 比较了不同隐变量模型的对于机器阅读的优劣势
4. 提出的隐变量模型在MCTest数据集上进行实验，并取得了世界领先水平
5. 提出了一套多层级注意力交互式用于机器阅读的神经网络系统
6. 验证了神经网络中多层交互式注意力机制对于系统性能提升的帮助
7. 验证了语义表示的增强对于神经网络提升机器阅读系统的帮助

## 论文组织结构