1.1研究意义及应用背景

近年来，随着人工智能学科的长足发展，使之成为了一门广泛的交叉和前言学科，越来越多的进入人们的视野，2017年12月，人工智能入选“2017年度中国媒体十大流行语”，使其再一次备受关注，作为计算机科学的前沿学科，人工智能研究的一个主要目标是使机器能够胜任一些通常需要人类智能才能完成的复杂工作，从视觉、听觉、触觉、感觉及思维方式等方面模拟人类，这其中涉及到计算机视觉、语音识别、自然语言处理等各个领域的诸多问题；

问答系统，作为实现机器和人类无障碍交流的一个重要切入点，成为近年来国内外学者研究的热点问题，纵观自动问答系统的技术发展历史，从1950年代因图灵测试而诞生至今，已经有几十年的历史。但真正在产业界得到大家的广泛关注，则得益于2011年Siri和Watson的成功所带来的示范效应。自此，自动问答系统较以往任何时候都显得离实际应用更近。这一方面归功于机器学习与自然语言处理技术的长足进步，另一方面得益于维基百科等大规模知识库以及海量网络信息的出现。

随着人工智能、自然语言处理等相关技术的发展，针对不同的数据形态的变化也衍生出不同种类的问答系统。早期由于智能技术和领域数据规模的局限性，问答系统主要是面向限定领域的AI系统或专家系统，例如STUDENT[1]、LUNAR[2]系统。该时期的问答系统处理的数据类型主要是结构化数据，系统一般是将输入问题转化为数据库查询语句，然后进行数据库检索反馈答案。随着互联网的飞速发展以与自然语言处理技术的兴起，问答系统进入了面向开放领域、基于自由文本数据的发展时期，例如英文问答式检索系统Ask Jeeves (http://www.ask.com)、START (http://start.csail.mit.edu)。。这种问答系统的处理流程主要包括：问题分析、文档及段落检索、候选答案抽取、答案验证。特别自1999年文本检索会议（Text Retrieval Conference，简称TREC）引入问答系统评测专项（Question Answering Track，简称QA Track）以来，极大推动了基于自然语言处理技术在问答领域中的研究发展。随着苹果公司Siri系统的问世，问答系统进入了智能交互式问答的发展阶段，这种形式的问答系统能够让用户体验更为自然的人机交互过程，并且也使信息服务的相关应用更为方便可行。

随后网络上出现的社区问答（community question answering, CQA）提供了大规模的用户交互衍生的问题答案对（question-answer pair, QA pair）数据，这为基于问答对的问答系统提供了稳定可靠的问答数据来源。社区问答系统是以用户为中心，用户参与的问答系统，所以包含了大量用户生成内容，有各种开放域的社区问答如Yahoo! Answers 、Quora 网站、百度知道、知乎等被人们所熟知，其中，仅Yahoo! Answers ，其包含各类问题三亿多条以及答案5亿多条，这些被用户提出和发布的超大规模的数据，促进了社区问答系统的研究；此外，还有像Stack Overflow这样的IT技术问答网站在近几年来也越来越受到大众的青睐，其注重为用户提供某一类特定域问题解决方案，随着用户数量的空前增长，其积累了大量用户生成的高品质有效信息，这些具有多样性和重用性的数据资源，可以在被充分处理和利用后有效地满足人们对相关信息和知识的需求，社区问答相对于传统的问答系统，其主要解决的问题是以咨询建议及回答询问等非事实类问题描述为主，而不是局限于对应命名实体等较短答案的事实类问题描述；社区问答系统的构建需要大量的数据，近年来国际文本检索会议TREC及国际语义评估组织SemEval都举办了一系列关于社区问答的专业竞赛，TREC的实时问答竞赛注重于构建一个完整的社区问答系统，在没有海量问答资源库的前提下，参赛队伍借助网络资源去获取相关答案内容，并做相应的处理获取最佳答案，TREC则更注重于答案选择排序算法的创新，提供给参赛队伍已构建好的数据集并期望其尽可能的提升任务的准确性以及在不同数据集上的泛化性，这类竞赛吸引了国内外众多知名学者及研究人员的关注和参与；

无论哪一类问答系统，在经过问句分析，信息检索后都会得到一些不同形式的答案片段，需要对其进行答案抽取过程，才能将正确的答案返回给用户，但在社区问答以及基于问答对的问答系统中，类似的工作被定义为答案选择排序，其研究的对象为问题答案对（QA pair），如图1-1所示，是一般社区问答系统的构建框架，从图中可以看出答案选择排序在社区问答系统中的重要地位，

图1-1 社区问答系统的构建框架

系统会以不同的方式进行检索、匹配，获取到问题对应的多个候选答案，而如何对这些候选答案进行进一步的研究，帮助用户选取最能回答问题的最佳答案，即快速准确地找到排在前面的优质答案是问答系统中的一个非常重要的研究内容；候选答案选择排序的问题可以抽象为给定一个问题，如何从其纷乱复杂的诸多候选答案中找到确定正确答案的方法并返回给用户的问题，这在社区问答系统中是具有非常重要的实际意义的，直接影响着问答系统的性能以及用户体验的效果；

1.2国内外研究现状

现阶段对于答案选择问题的研究大多是通过提取问句与候选答案的句法及语义信息，然后计算问题与候选答案的语义相关性，前期阶段主要是通过词汇、句法等特征生成问句和候选答案的句法逻辑表示，文献（1）（2）提出了一种新的基于条件随机场的概率模型，通过构造句法树来表示问句和候选答案的句法结构相关性；文献（3）为了克服简单词袋模型的缺点，首先对句子进行指代消解预处理，并提出将编辑距离与向量空间模型相结合，同时在候选句选取过程中，通过句法树树形结构匹配的方法将句子的句法信息融入到处理过程中，文献（4、5）提出了机器翻译的方法，该方法主要是通过衡量问句与候选答案中词、短语的对其情况来得出其句法匹配程度；文献（6）问句与候选答案匹配问题转换为分类问题，通过定义一系列树编辑序列及特定于句法树的操作序列，生成问句和候选答案的编辑序列特征集，最后通过逻辑回归计算问句与候选答案的匹配程度；除了以上在基于句法分析方面的努力外，研究者们在通过基于语义方面的研究也取得了很大进展，文献（7）通过主题模型获得问句和候选答案的主题语义信息，通过计算两者的主题语义距离对答案进行排序；文献（8）提出了词汇语义模型，通过WordNet提取关系语义特征，得到问句和候选答案中词的对齐关系，然后使用潜在语义分析得到同义词的语义相近度，进而获得句子对语义特征集合；文献（9）通过捕捉句子内部和外部的因果关系，将其加入答案选择排序任务，以提高因果类非事实型问题答案选择的质量；文献（10）在不借助外部资源库而仅使用语言工具报的情况下，提取词汇相关特征，并使用SVM自动学习句子的浅层特征来对候选答案进行分类；

文献（10）中提出，利用问答对的文本信息及用户权威性信息等大规模的问答数据和Ranking SVM算法训练出一个答案排序模型，选取最佳答案，该模型可以有效的弥补问答社区给出的最佳答案的不足，给出更好的答案。

近年来，深度学习模型在自然语言处理的各项任务中有了越来越广泛的应用，因此，部分研究者也尝试应用深度学习技术，解决答案选择问题；相比传统的需要大量人工标注与特征资源的方法，基于深度学习的方法取得了更好的结果，此外，互联网用户为了交流信息而产生的大规模诸如微博回复、社区问答对的自然标注数据[50]，给训练深度神经网络模型提供了可靠的数据资源，并很大程度上解决自动问答研究领域的数据匮乏问题。深度学习的方法应用于自然语言处理各项任务的首要目标就是生成词语或句子的分布式表示，分布式表示相比于基于词汇的表示，克服了其数据稀疏性问题。同时能获取词语或句子更好的浅层语义信息，在生成句子的分布式语义特征表示之前，首先需要生成词向量（word-embedding），词向量计算作为答案选择任务在词层面的研究，得到了越来越多的关注，早期的词向量大多是利用google word2vec(5)中的词向量模型生成，此外，jeff（6）等人，提出了将全局词共现（Global matrix factorization）和局部词共现（Shallow Window-Based）两种词向量学习方法相结合的方式来生成词向量；近年来，有越来越多的研究根据不同的NLP任务对词向量模型做相应的改进以期更好的完成不同的任务，文献（12）（13）中提出了一种融合情感标签的word2vec-ST词向量生成方法，以此为词向量层的基础，完成情感分析人物，并在，并证明此方法大大提升提取句子语义信息及情感信息的效果；文献（11）中提出了一种字词联合训练的词向量生成方法,首次提出了将组成词语的汉字考虑进词向量的训练过程，并在人名日报数据集上验证了此方法相比传统词向量生成方法有更好的效果，

文献（23）提出将二元卷积神经网络分别应用于问题和候选答案，生成其分布式语义表示，计算出相似矩阵对问题和候选答案的语义相似度进行打分；文献（28）提出了结构更为复杂的卷积神经网络，并对问句和候选答案共用同一个网络的模型和分别对问句、候选答案建模的模型效果进行了对比分析，并证明了增加问句、候选答案之间的约束能提升模型表现。 Yu文献（24）等将答案选择问题看成二分类问题，提出了一种二元卷积模型将候选答案分类为正确问题及不正确问题；文献（25）针对非事实类问题提出基于bilstm的建模，获取句子的向量表示，利用余弦相似度衡量问句和候选答案的语义相关性，并且该文献中尝试将问句和候选答案通过共同的网络结构，并共享网络参数以获得更好的选取效果；文献（26）提出了以深度学习模型LSTM为基线系统，在其基础上以不同的形式添加问句和候选答案的词汇特征、BTM主题特征以及对齐特征，并在公共数据集上验证了其所构建模型的有效性；文献（27）中提出，将LSTM和CNN模型进行结合，利用CNN模型获取问句和各个候选答案的联合特征向量表示，并将其作为LSTM模型每个时间步的输入，以此计算并预测每个候选答案的匹配质量（e.g good，potential or bad）；可以发现，前期的主要研究多集中在以卷积神经网络为基础的模型，后期的主要研究多集中在以长短期记忆网络为基础的研究或是二者的联合，这主要是因为卷积神经网络最早在图像处理领域取得了鲜为人知的丰富研究成果，才有了其最先在自然语言处理领域的首次尝试，同时，这也充分证明了循环神经网络在自然语言处理各类问题中的巨大潜力；

随着深度学习在自然语言处理领域的应用越来越成熟，越来越多的研究者开始尝试在不同的任务中引入Attention机制，以试图提升模型的效果；在事实类问答任务（33）、标题生成（32）、机器翻译（31 ）（30）等任务中，Attention机制对模型效果的提升起到了非常大的作用；文献（25）中通过计算候选答案在每一个时间步下的向量和问句向量的相关度来加强对问题较为重要的词的权重，减弱对问题不重要的词的权重，这使得网络输出端的某个节点能够将注意力集中在输入部分的某些特定词语片段，从而提升模型的效果；

* 1. 论文的主要研究内容

1）构建基于深度学习的答案选择排序算法框架，明确答案选择排序问题的研究思路及核心任务；

1）针对基于深度学习的答案选择排序模型输入端词向量的计算问题，传统的词向量模型大多是以词为单位进行分析并训练，考虑在汉语言中，组成词语的内部字往往拥有很丰富的内部信息，并且同一个汉字在不同的组合或不同的语境中往往有着不同的字义等特点，本文提出字和词语联合训练的词向量计算方法，尝试将此方法应用于CBOW词向量模型中，并为每一个汉字设立多重向量，建立多标准的字向量确定方法，并尝试将基于位置的方法和基于簇的方法结合以确定字向量，提高词向量模型的可靠性和有效性；

2）以词向量矩阵输入为基础，构建神经网络结构对问题和候选答案进行句子对建模，分别构建基于CNN、BILSTM的句子对建模模型；由于不同的词语对于句子的语义特征表示有着不同的重要性，在句子语义特征向量生成过程中应被分配不同的注意力，所以，在BILSTM的基础上，融入Attention机制，根据问句和候选答案之间的关联性进行Attention计算生成句子中每个词语的权重；针对答案选择排序任务中非事实类问题占主要部分的实际情况，并且对于非事实类问题，问句和候选答案之间的相似性、关联性和正确性并没有必然联系的特性，提出基于问题和候选答案独立性的Attention模型self Attention，此模型利用一个附加的BILSTM结构来进行Attention计算，根据每个词语对于句子本身语义表达所起的作用来衡量词语的重要性并赋予其相应的权重，从而更好的表示问句和候选答案的语义特征，提高答案选择排序模型的性能。

1.3 论文的主要组织结构

第一章，绪论。主要阐述了本文的研究背景及选题意义，并对答案选择排序的国内外理论研究现状和应用现状进行介绍分析，并阐明了本论文的主要研究内容，给出论文的组织结构安排；

第二章，相关理论和技术，介绍了目前常用的答案选择排序方法，着重介绍了本文基于深度学习的方法的理论基础，包括word embedding、CNN卷积神经网络、BILSTM双向长短期记忆神经网络、Attention机制，最后，介绍了本课题研究数据的来源与获取；

第三章，答案选择排序模型输入端词向量的计算，首先，介绍字和词语联合训练的词向量计算方法，设立多标准的字向量确定方法，根据汉字在词语中出现位置的不同和组合方式的不同，分别设立基于位置的方法和基于簇的方法，然后将基于位置的方法和基于簇的方法结合确立字向量，将此种词向量计算方法应用于CBOW词向量模型，然后在研究数据集上对改进的模型进行实验验证，采用Hierarchical Softmax和梯度上升法更新模型参数训练模型，最后，从模型的复杂度和词语对相关度计算两个方面对模型性能进行了评估，确定了基于字和词语联合训练的词向量模型的可靠性和有效性，

第四章，基于深度学习的答案选择排序模型构建，首先，构建神经网络对问句和候选答案进行句子对建模，在BILSTM模型的基础上融合Attention机制进行Attention计算获得问句和候选答案中每个词的权重；然后，提出基于问句和候选答案独立性的self Attention模型，通过一个附加的BILSTM网络进行Attention计算，在不考虑问句和候选答案关联性的基础上赋予句子中每个词语不同的权重；最后，以最小化交叉熵代价函数为训练目标，反向传播及随机梯度下降为训练方法，在数据集上对模型进行训练，并对答案选择排序模型的性能和效果进行分析评价；

第五章，结论与展望，对本文的整体内容进行总结，在目前已经完成工作基础上，对后续的算法改进工作进行展望；