* 1. 研究意义及应用背景

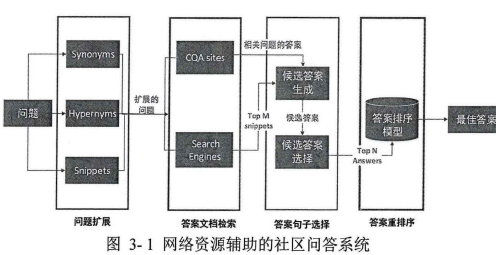
随着国家的快速发展和不断强大，公民在经济、政治、文化等各个方面享受越来越多的权益，与此同时，随着国家对“依法治国”观念的深入贯彻，越来越多的公民更愿意去学法，懂法、守法、提高自己的法律意识，维护与承担属于自己的权利与义务，因此，近年来，有越来越广泛的关于法律咨询的途径与渠道应运而生；法律咨询是指提供法律知识问题的解答，由于法律的复杂性，非专业人士在遇到法律问题时，往往需要通过有效渠道求助于相关人士，法律咨询可以分为现场口头咨询、当事人提供材料后给予咨询意见、正式的收费咨询、非正式的不收费的仅供参考的咨询，而法律问答社区就是通过网络资源共享，互联网信息搜素、开放式的交流互动等方式为法律知识提问者提供解答方案及咨询意见的有效途径，并且已经成为人们传授知识、交流经验的重要平台

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，社区问答不同于传统的问答系统，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；然而，不管是哪类社区问答系统，都会以不同的方式获取到问题对应的多个候选答案，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社区问答系统中的一个非常重要研究内容；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，这在当今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性能的好坏以及用户体验效果的好坏；

如图1-1所示，显示了一般社区问答系统的构建框架以及答案选择排序在社区问答系统中存在的意义及地位；



1.2国内外研究现状

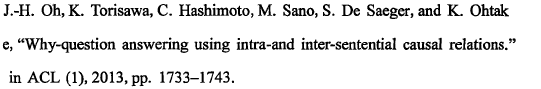
现阶段对于答案选择问题的研究大多是通过提取问句与候选答案的句法及语义信息，然后计算问题与候选答案的语义相关性，前期阶段主要是通过词汇、句法等特征生成问句和候选答案的句法逻辑表示，文献（1）（2）提出了一种新的基于条件随机场的概率模型，通过构造句法树来表示问句和候选答案的句法结构相关性；文献（3）为了克服简单词袋模型的缺点，首先对句子进行指代消解预处理，并提出将编辑距离与向量空间模型相结合，同时在候选句选取过程中，通过句法树树形结构匹配的方法将句子的句法信息融入到处理过程中，文献（4、5）提出了机器翻译的方法，该方法主要是通过衡量问句与候选答案中词、短语的对其情况来得出其句法匹配程度；文献（6）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；

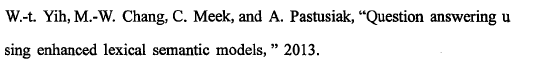
文献（10）中提出，利用问答对的文本信息及用户权威性信息等大规模的问答数据和Ranking SVM算法训练出一个答案排序模型，选取最佳答案，该模型可以有效的弥补问答社区给出的最佳答案的不足，给出更好的答案。

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，而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，

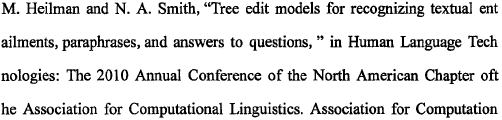
文献（23）提出将二元卷积神经网络分别应用于问题和候选答案，生成其分布式语义表示，计算出相似矩阵对问题和候选答案的语义相似度进行打分；文献（28）提出了结构更为复杂的卷积神经网络，并对问句和候选答案共用同一个网络的模型和分别对问句、候选答案建模的模型效果进行了对比分析，并证明了增加问句、候选答案之间的约束能提升模型表现。 Yu文献（24）等将答案选择问题看成二分类问题，提出了一种二元卷积模型将候选答案分类为正确问题及不正确问题；文献（25）针对非事实类问题提出基于bilstm的建模，获取句子的向量表示，利用余弦相似度衡量问句和候选答案的语义相关性，并且该文献中尝试将问句和候选答案通过共同的网络结构，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；

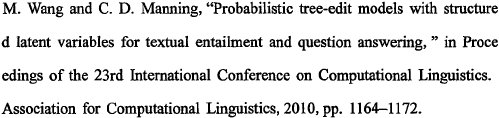
随着深度学习在自然语言处理领域的应用越来越成熟，越来越多的研究者开始尝试在不同的任务中引入Attention机制，以试图提升模型的效果；在事实类问答任务（33）、标题生成（32）、机器翻译（31 ）（30）等任务中，Attention机制对模型效果的提升起到了非常大的作用；文献（25）中通过计算候选答案在每一个时间步下的向量和问句向量的相关度来加强对问题较为重要的词的权重，减弱对问题不重要的词的权重，这使得网络输出端的某个节点能够将注意力集中在输入部分的某些特定词语片段，从而提升模型的效果；

9

8 

7 

1 

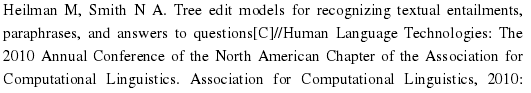
2 

3

5 



4 

6 

33 Advances in Neural Information Processing Systems. 2015:

32 Neural image caption generation with visual attention[J]. ar Xiv preprint ar Xiv:1502.03044, 2015.

31 Sequence to sequence learning with neural networks[C]//Advances in

30 neural information processing systems. Neural machine translation by jointly learning to align and translate[J]. ar Xiv preprint ar Xiv:1409.0473, 2014.

（11）joint learning of character word embedding

（27）answer sequencelearning with nerual network in。。。

（23）deep learning on answer setence selection

（24）Yu, L., K. M. Hermann, P. Blunsom, and S. Pulman (2014). Deep Learning for Answer Sentence Selection.

（25）LSTM-Based Deep Learning Models for Non-factoid Answer Selection(IBM)

5 

6 

（26）深度学习与多元特征相结合的答案选择排序研究

（27）问题检索与答案排序相互促进的社区问答系统

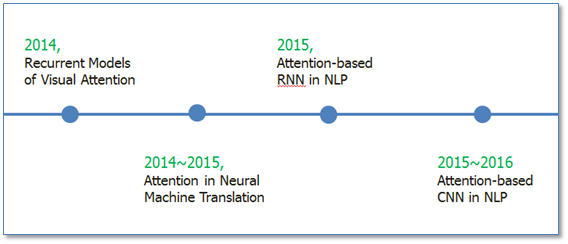
（10）非事实类问题的答案选取

 （28）Applying Deep Learning to Answer Seleciton (IBM)

（31）Effective Approaches to Attention-based Neural Machine Translation [2]

（30）Bahdanau等人在论文《Neural Machine Translation by Jointly Learning to Align and Translate》 [1]中，使用类似attention的机制在机器翻译任务上将翻译和对齐同时进行，他们的工作算是是第一个提出attention机制应用到NLP领域中

下图表示了attention研究进展的大概趋势。



* 1. 论文的主要研究内容

1）本文分析了基于深度学习的答案选择模型输入端词向量的计算，考虑到在汉语言中，组成词语的内部字往往拥有很丰富的内部信息的特点，提出了字和词语联合训练的词向量计算方法，并尝试将此方法应用于CBOW词向量模型中，对于此方法中字向量的确定，考虑到汉字具有很强的歧义性，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，从而为每一个汉字设立多重向量，并建立了多标准的字向量确定方法，即基于位置的字向量，基于簇的字向量，并尝试将两者结合，建立基于位置和簇的字向量；通过在数据集上的多次实验，得出基于字和词语联合训练的CBOW模型有更好的性能；

2）以词向量矩阵输入为基础，构建神经网络结构来对问题和候选答案进行句子对建模，分别构建了基于CNN、BILSTM的句子对建模模型；而不同的词语对于句子的语义特征表示有着不同的重要性，所以在BILSTM的基础上，融入Attention机制，首先应用的是基于问句和候选答案相似性的APN注意力模型，生成问句和候选答案之间的Attention矩阵，然后计算出问句和候选答案中每个词语的权重；但针对本课题研究的面向社区问答的答案选择排序问题，考虑到非事实类问题占主要部分及其相似性并不能反映重要性的固有特点，提出了基于问题和候选答案独立性的self-Attention模型，根据句子中每个词的重要性赋予其相应的权重，从而更好的表示问句和候选答案的语义特征；

3）以基于深度学习的答案选择排序模型为基础，分析并设计了法律社区问答系统，该系统主要包括问题问题扩展、问题检索、答案选择排序等模块，重点设计实现了答案选择排序模块，并对系统进行了测试，测试结果表明系统已经达到了功能要求；

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。

第三章，答案选择模型输入端的词向量计算方法。在原有的以词为单位训练词向量的方法上