基于深度学习的智能问答

- 作者：周小强 陈清财 曾华军

- 2016-08-03
- 文章来源于阿里巴巴-哈尔滨工业大学的合作项目：基于深度学习的智能问答。

1 引言

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 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。然而，现有的自动问答系统所面临的问题远没有完全解决。事实上，无论是业界应用还是学术研究，问句的真实意图分析、问句与答案之间的匹配关系判别仍然是制约自动问答系统性能的两个关键难题。

2 问答系统概述

问答系统能够更为准确地理解以自然语言形式描述的用户提问，并通过检索异构语料库或问答知识库返回简洁、精确的匹配答案。相对于搜索引擎，问答系统能更好地理解用户提问的真实意图，同时更有效地满足用户的信息需求。

2.1 问答系统的发展历程

问答系统最早的实现构想可以追溯到图灵测试。为了测试机器是否具备人类智能，图灵测试要求电脑能在 5 分钟内回答由人类测试者提出的一系列问题，且其达到超过 30%的回答让测试者误认为是人类所答。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及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。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。

问答系统处理的数据对象主要包括用户问题和答案。依据用户问题的所属数据领域，问答系统可分为面向限定域的问答系统、面向开放域的问答系统、以及面向常用问题集（frequent asked questions, FAQ）的问答系统。依据答案的不同数据来源，问答系统可划分为基于结构化数据的问答系统、基于自由文本的问答系统、以及基于问答对的问答系统。此外，按照答案的生成反馈机制划分，问答系统可以分为基于检索式的问答系统和基于生成式的问答系统。本文主要阐述基于检索式的问答系统的处理框架和相关研究。

2.2 问答系统的处理框架

不同类型的问答系统对于数据处理的方法存在不同。例如，相对于面向 FAQ 的问答系统的问句检索直接得到候选答案，面向开放领域的问答系统首先需要根据问题分析的结果进行相关文档、文本片段信息的检索，然后进行候选答案的抽取。虽然不同类型的问答系统对于系统模块的功能分工和具体实现存在差异，但依据数据流在问答系统中的处理流程，一般问答系统的处理框架中都包括问句理解、信息检索、答案生成三个功能组成部分，如图 2.1 所示。

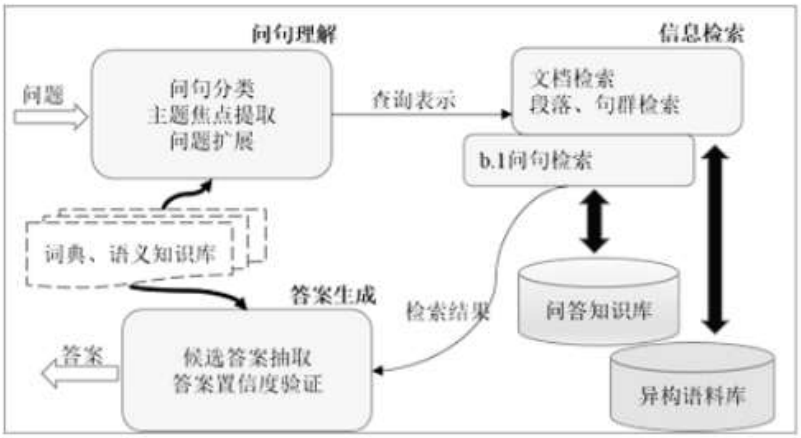

图 2.1

2.2.1 问句理解

问句理解是问答系统理解用户意图的关键一环，问句理解模块的性能直接制约着后续处理模块的效果。用户意图是一个抽象的概念，要想作为答案检索的依据，需要把它转换成机器可以理解的形式。用户的检索意图导致信息需求的产生，因此，研究中往往将信息需求作为用户意图的代表，根据问句的语义结构，可以从问题类别和问题内容两方面来表示。通常采用自然语言技术对问题进行深层次的理解，包括命名实体识别、依存句法分析、词义消歧等。

问句理解主要包括问句分类、主题焦点提取、问题扩展处理。问句分类是将用户提问归入不同的类别，使系统能够针对不同问题类型采用不同的答案反馈机制得到候选答案集合。问答系统通常使用机器学习算法训练问题分类器[3,4]来实现用户提问的分类。主题焦点提取主要完成用户问题的信息需求的精确定位，其中主题表示问句的主要背景或者用户的感兴趣的对象，焦点则是用户询问的有关主题的内容，通常是问句话题的相关信息或对话题起到描述性的作用，比如属性、动作、实例等等。问题扩展是将用户在提问中没有充分表达的意思补充出来，对问题中潜在的信息显化出来，从而提高答案检索的召回率。

2.2.2 信息检索

根据问句理解得到的查询表示，信息检索模块负责从异构语料库、问答知识库中检索相关信息，传递给后续的答案生成处理模块。对于基于不同的问答系统，系统的检索模型以及检索数据形式也不同。对于基于自由文本数据的问答系统，信息检索过程是一个逐渐缩小答案范围的过滤过程，主要包括文档检索和段落句群检索。对于基于问句答案对的问答系统，信息检索处理是通过问句检索得到与用户提问相似的候选问句，返回对应的候选答案列表。

首先，文档检索是根据问题理解的结果检索用户提问的相关文档集合。最简单的方法是直接用已有的检索系统（如 Smart，Lucene 等）对问题的非停用词进行全文索引，直接检索得到用户提问的相关文档集合，但是这种方法很难获得好的效果。通常问答系统中的文档检索模型包括布尔模型、向量空间模型、语言模型、概率模型等。布尔模型是最简单的一种检索模型，它把关键词组织成一个布尔表达式，使得文档中出现的关键词需要满足这个布尔表达式。向量空间模型把文档和查询都表示成向量，根据查询和文档对应向量的相似度（通常是两个向量夹角的余弦值）对文档进行排序。概率模型估计计算文档和查询相关的概率，并按照相关性概率对文档进行排序。语言模型是把查询和文档分别表示成语言模型，通过两个语言模型之间的 KL 距离来估计两者的相似度。其次，段落句群检索就是从候选文档集合中检索出最有可能含有答案的段落（自然段落或者文档片段），进一步过滤噪声信息，得到更为精确的答案相关信息。广泛使用的段落检索算法有三个：MultiText 算法[6]、IBM 的算法[7,8]和 SiteQ 算法[9]。Tellex[10]等人的实验结果表明基于密度的算法可以获得相对较好的效果。所谓基于密度的算法，就是通过考虑查询关键词在段落中的出现次数和接近程度来决定这个段落的相关性。相比之下，Cui[5]提出的检索算法通过把问句和答案都解析成语法树，从两者语法树的结构中找出一些相关性的信息。

问句检索的主要问题在于如何缩小用户提问与知识库中问句之间的语义鸿沟。近几年，研究人员采用基于翻模模型的方法计算从用户提问“翻译”到检索问句的翻译概率，从而实现相似性问句检索。例如，算法[11-14]都是把两个问句看作是不同表达方式的语句，计算两个问句之间的翻译概率。为了计算这种翻译的概率，就必须估计词与词之间进行翻译的概率。这种方法首先需要通过答案相似度计算得到同义或近义的问答对集合，该集合中的相似问题集合就构成了一个估计翻译概率的训练集，类似于机器翻译中多语言平行语料库。实验证明，这样做的效果会比语言模型，Okapi BM25 和空间向量模型都好。

2.2.3 答案生成

基于信息检索得到的检索信息，答案生成模块主要实现候选答案的抽取和答案的置信度计算，最终返回简洁性、正确性的答案。按照答案信息粒度，候选答案抽取可以分为段落答案抽取、句子答案抽取、词汇短语答案抽取。段落答案抽取是将一个问题的多个相关答案信息进行汇总、压缩，整理出一个完整简洁的答案。句子答案抽取是将候选答案信息进行提纯，通过匹配计算过滤表面相关，实际语义不匹配的错误答案。词汇短语抽取是采用语言的深层结构分析技术从候选答案中准确地提取答案词或短语。

答案置信度计算是将问题与候选答案进行句法和语义层面上的验证处理，从而保证返回答案是与用户提问最为匹配的结果。应用最广泛是基于统计机器学习的置信度计算方法。这种方法通常定义一系列词法、句法、语义以及其他相关特征（如编辑距离、BM25 等）来表示问题与候选答案之间的匹配关系，并使用分类器的分类置信度作为答案的置信度。例如 IBM Waston 中使用的答案融合和特征排序方法[15]，以及基于关系主题空间特征的多核 SVM 分类方法[16]。近几年，基于自然语言处理的问答匹配验证通常是使用句子的浅层分析获得句子的浅层句法语法信息，然后将问句与答案的句法树（短语句法树或依存句法树）进行相似性计算[17-20]。然而，问答系统的答案正确性更需满足问题和答案之间的语义匹配，比如问“苹果

6s plus 最新活动价多少”，如果回答“红富士苹果降到了 12 元”，就属于所答非所问。常用的方法是通过引入诸如语义词典（WordNet），语义知识库（Freebase）等外部语义资源进行问答语义匹配建模[21-23]，以此提高问句答案间的语义匹配计算性能。

传统问答系统中构建的机器学习模型基本属于浅层模型。譬如，问句分类过程中常用的基于支持向量机（SVM）的分类模型[24]，答案抽取使用的基于条件随机场（CRF）的序列标注模型[25]，以及候选答案验证过程中使用的基于逻辑回归（LR）的问答匹配模型[26]等。这种基于浅层模型研发的问答系统往往存在人工依赖性高，并且缺少对不同领域数据处理的泛化能力。人工依赖性主要表现在浅层模型的特征工程上，由于浅层模型缺乏对数据的表示学习的能力，于是在面对不同领域的问答数据以及不同的问答任务的情况下，研究人员不得不进行针对性的数据标注，并且需要依据研究人员的观察和经验来提取模型所需的有效特征，这也就造成了此类问答系统可移植性低的结果。

3 基于深度学习的相关问答技术

近年来，深度神经网络在诸如图像分类、语音识别等任务上被深入探索并取得了突出的效果，表现出了优异的表示学习能力。与此同时，通过深度神经网络对语言学习表示已逐渐成为一个新的研究趋势。然而，由于人类语言的灵活多变以及语义信息的复杂抽象，使得深度神经网络模型在语言表示学习上的应用面临比在图像、语音更大的挑战。其一，相比于语音和图像，语言是非自然信号，完全是人类文明进程中，由大脑产生和处理的符号系统，是人类文明智慧的高度体现。语言的变化性和灵活度远远超过图像和语音信号。其二，图像和语音具有明确的数学表示，例如灰度图像为数学上的数值矩阵，而且其表示的最小粒度元素都有确定的物理意义，图像像素的每个点的值表示一定的灰度色彩值。相比而言，以往的词袋表示方法会导致语言表示存在维数灾难、高度稀疏以及语义信息损失的问题。

当前，研究人员越来越对深度学习模型在 NLP 领域的应用感兴趣，其主要集中在对词语、句子和篇章的表示学习以及相关应用。例如，Bengio 等使用神经网络模型得到一种名为词嵌入（Word Embedding）或词向量的新型向量表示[27]，这种向量是一种低维、稠密、连续的向量表示，同时包含了词的语义以及语法信息。当前，基于神经网络的自然语言处理方法大都是基于词向量的表示基础上进行的。在此基础上，相关研究人员设计深度神经网络模型学习句子的向量表示，相关工作包括递归神经网络（Recursive Neural Network）、循环神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Network, CNN）的句子建模[28-30]。句子表示被应用于大量的自然语言处理任务上，并在一些任务上取得了较为突出的效果。例如机器翻译[31, 32]、情感分析等[33, 34]。从句子的表示到篇章的表示学习仍然较为困难，相关工作也较少，比较有代表性是 Li 等人通过层次循环神经网络对篇章进行编码，然后通过层次循环神经网络进行解码，从而实现对篇章的表示[35]。然而，NLP 领域涵盖了不同性质，不同层次的具体问题，这就需要针对不同问题的特点，设计深度模型学习到任务特定的本质特征。

问答领域所需解决的两个关键问题：一是如何实现问句及答案的语义表示。无论是对于用户提问的理解，还是答案的抽取验证，都需抽象出问题和答案的本质信息的表示。这不仅需要表示问答语句的句法语法信息，更需表示问句及答案在语义层面上的用户意图信息和语义层匹配信息。二是如何实现问句答案间的语义匹配。为了保证反馈用户提问的答案满足严格语义匹配，系统必须合理利用语句高层抽象的语义表示去捕捉到两个文本之间关键而细致的语义匹配模式。鉴于近几年卷积神经网络（CNN）和循环神经网络（RNN）在 NLP 领域任务中表现出来的语言表示能力，越来越多的研究人员尝试深度学习的方法完成问答领域的关键任务。例如问题分类（question classification），答案选择（answer selection），答案自动生成（answer generation）。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。

接下来内容安排：首先，分别介绍基于 CNN 和 RNN 的问答语句的语义表示方法；然后，介绍基于 DCNN 的两种语义匹配架构；最后，介绍基于 RNN 的答案自动生成方法。

3.1 基于深度神经网络的语义表示方法

3.1.1 基于卷积神经网络（CNN）的语义表示方法

基于 CNN 的语义表示学习是通过 CNN 对句子进行扫描，抽取特征，选择特征，最后组合成句子的表示向量。首先从左到右用一个滑动窗口对句子进行扫描，每个滑动窗口内有多个单词，每个单词由一个向量表示。在滑动窗口内，通过卷积（convolution）操作，进行特征抽取。这样，在各个位置上得到一系列特征。之后再通过最大池化（max pooling）操作，对特征进行选择。重复以上操作多次，得到多个向量表示，将这些向量连接起来得到整个句子的语义表示。如图 3.1 所示，基于 CNN 的句子建模的输入是词向量矩阵，矩阵的每一行的多个点的值在一起才有明确的物理意义，其代表句子中对应的一个词。词向量矩阵是通过将句子中的词转换为对应的词向量，然后按照词的顺序排列得到。该模型通过多层交叠的卷积和最大池化操作，最终将句子表示为一个固定长度的向量。该架构可以通过在模型顶层增加一个分类器用于多种有监督的自然语言处理任务上。

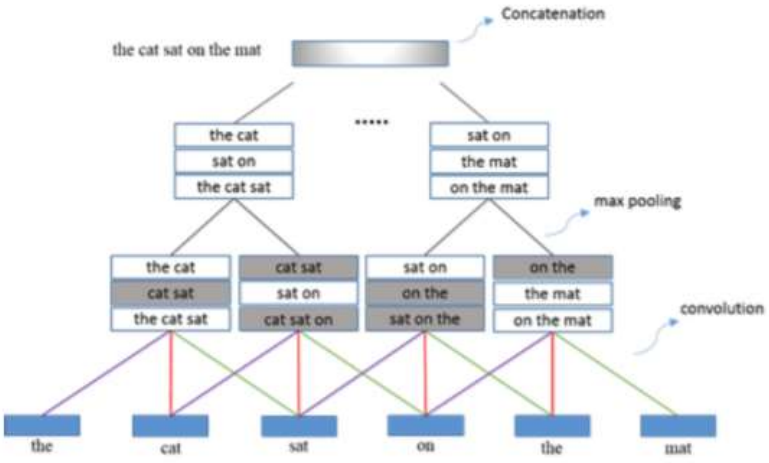

图 3.1 基于 CNN 的句子建模

基于 CNN 的句子建模可以表现为具有局部选择功能的“组合算子”，随着模型层次的不断加深，模型得到的表示输出能够覆盖的句内词的范围越广，最后通过多层的运算得到固定维度的句子表示向量。该过程的功能与“递归自动编码”的循环操作机制[33]具有一定的功能类似。对于只使用了一层卷积操作和一层全局最大池化操作的句子建模，称之为浅层卷积神经网络模型，这种模型被广泛应用于自然语言处理中句子级分类任务上，如句子分类[36]，关系分类[37]。但是，浅层的卷积神经网络模型不能对句子中复杂的局部语义关系进行建模，也不能对句子中深层次的语义组合进行很好的表示，并且全局最大池化操作丢失了句子中的词序特征，所以浅层的卷积网络模型只能对语句间的局部特征匹配进行建模。面对问答中复杂多样化的自然语言表示形式（如多语同现，异构信息，表情符号等），问答匹配模型[38-40]往往使用深层卷积

神经网络(DCNN)来完成问句和答案的句子建模 ,并将高层输出的问答语义表示传递给多层感知器(MLP ）进行问答匹配。

面对开放领域中的关系性推理问题，例如“微软公司的创始人是谁？”，往往通过引入外部语义知识推理得到问题的答案，此时单一的句子建模很难实现逻辑关系的语义表示。通常先需要对问题进行语义解析（Semantic Parse），然后针对问句实体、实体关系等不同类型的语义信息进行表示学习。Yih 将关系性问题拆分成实体集合和关系模板[41]，其中实体集合为问题中连续词语的子序列，关系模板为问句实体被特殊符号替换后的句子，针对实体集合和关系模板分别使用 CNN 进行句子建模，从而实现问句在实体及关系两个层面上的语义表示。Dong 提出多栏（Multi-Column）卷积神经网络模型[42]对关系推理性问题进行不同层面（词语表达层、实体关系层、语境信息层）的语义表示学习，并实现从关系知识库中抽取候选答案的多层面语义信息，最后与候选答案进行多层次匹配打分。

3.1.2 基于循环神经网络（RNN）的语义表示方法

基于 RNN 的句子建模是把一句话看成单词的序列，每个单词由一个向量表示，每一个位置上有一个中间表示，由向量组成，表示从句首到这个位置的语义。这里假设，每一个位置的中间表示由当前位置的单词向量以及前一个位置的中间表示决定，通过一个循环神经网络模型化。RNN 把句末的中间表示当作整个句子的语义表示，如图 3.2 所示。

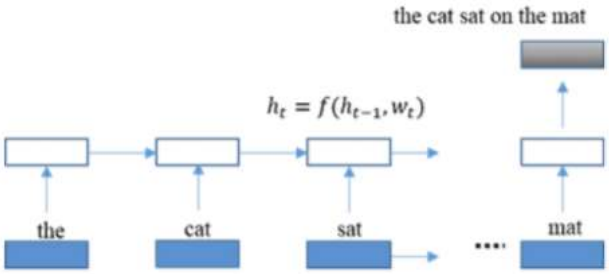

图 3.2 基于 RNN 的语句建模

RNN 与隐马尔可夫模型有相似的结构，但是具有更强的表达能力，中间表示没有马尔可夫假设，而且模型是非线性的。然而，随着序列长度的增加，RNN 在训练的过程中存在梯度消失（Vanishing gradient problem）的问题[43]。为了解决这个问题，研究人员对循环神经网络中的循环计算单元进行改善设计，提出了不同的变形，如常用的长短记忆（Long Short Term Memory, LSTM）[44, 45]和门控循环单元（Gated Recurrent Unit, GRU）[56]。这两种 RNN 可以处理远距离依存关系，能够更好地表示整句的语义。Wang 和 Nyberg [47]通过双向 LSTM 学习问题答案对的语义表示，并将得到的表示输入到分类器计算分类置信度。

此外，对于近几年的看图回答的任务（Image QA），研究人员通过整合 CNN 和 RNN 完成问题的图像场景下的语义表示学习。基本想法：模型在 RNN 对问句进行词语序列扫描的过程中，使用基于深度学习的联合学习机制完成“图文并茂”的联合学习，从而实现图像场景下的问句建模，用于最终的问答匹配。例如，Malinowski 等人[48]提出的学习模型在 RNN 遍历问句词语的过程中，直接将 CNN 得到的图像表示与当前词语位置的词向量作为 RNN 学习当前中间表示的输入信息，从而实现图像与问句的联合学习。相

比之下，Gao 等人[49]则是先用 RNN 完成问题的句子建模，然后在答案生成的过程中，将问句的语义表示向量和 CNN 得到的图像表示向量都作为生成答案的场景信息。

3.2 基于 DCNN 的语义匹配架构

问答系统中的语义匹配涉及到主要功能模块包括:问句检索 ,即问句的复述检测(paraphrase) ;答案抽取，即问句与候选文本语句的匹配计算；答案置信度排序，即问题与候选答案间的语义匹配打分。

3.2.1 并列匹配架构

第一种基于 DCNN 的语义匹配架构为并列匹配 [38-40]架构。这种架构的匹配模型分别将两句话输入到两个 CNN 句子模型，可以得到它们的语义表示 (实数值向量) 。之后，再将这两个语义表示输入到一个多层神经网络 ,判断两句话语义的匹配程度 ,从而判断给定的两句话和是否可以成为一对句子匹配对(问答对) 。这就是基于 DCNN 的并列语义匹配模型的基本想法。如果有大量的信息和回复对的数据，就可以训练这个模型。

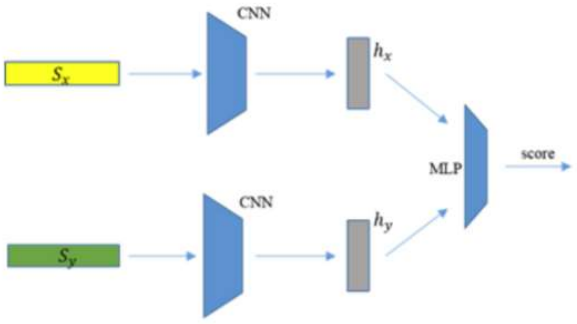

图 3.3 基于 DCNN 的并列匹配架构

从图 3.3 所示的并列匹配架构可以看出，这种匹配模型的特点是两个句子的表示分别通过两个独立的卷积神经网络 (CNN) 得到，在得到它们各自的表示之前，两个句子间的信息互不影响。这种模型是对两个需要匹配的句子从全局语义上进行匹配，但是忽略了两个句子间更为精细的局部匹配特征。然而，在语句匹配的相关问题中，两个待匹配的句子中往往存在相互间的局部匹配，例如问题答案对：

Sx: 好饿啊，今天去哪里吃饭呢。

Sy: 听说肯德基最近出了新品，要不要去尝尝。

在这一问答对中，“吃饭”和“肯德基”之间具有较强的相关性匹配关系，而并列匹配则是对句子两个句子全局的表示上进行匹配，在得到整个句子的表示之前，“吃饭”和“肯德基”之间并不会互相影响，然而，随着深度卷积句子模型对句子的表示层次不断深入，而句子中的细节信息会部分丢失，而更关注整个句子的整体语义信息。

3.2.2 交互匹配架构

第二种基于 DCNN 的语义匹配架构为交互匹配[39]架构。与并列匹配不同，交互匹配的基本想法是直接对两个句子的匹配模式进行学习，在模型的不同深度对两个句子间不同粒度的局部之间进行交互，学习得到句子匹配在不同层次上的表示，最终得到句子对固定维度的匹配表示，并对匹配表示进行打分。

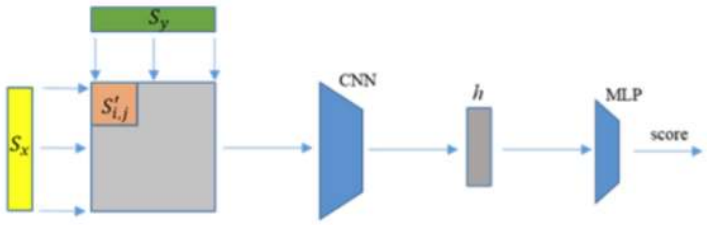

图 3.4 基于 DCNN 的交互匹配架构

如图 3.4 所示，交互匹配架构在第一层通过两个句子间的滑动窗口的卷积匹配操作直接得到了两个句子间较为底层的局部匹配表示，并且在后续的高层学习中采用类似于图像领域处理过程中的二维卷积操作和二维局部最大池化操作，从而学到问句与答案句子之间的高层匹配表示。通过这种形式，使得匹配模型既能对两个句子的局部之间的匹配关系进行丰富建模，也使模型能够对每个句子内的信息进行建模。很显然，交互匹配学习得到的结果向量不仅包含来自两个句子的滑动窗口的位置信息，同时具有两个滑动窗口的匹配表示。

对于问答的语义匹配，交互匹配可以充分考虑到问句与答案间的内部匹配关系，并通过二维的卷积操作与二维局部最大池化操作学习得到问句与答案间的匹配表示向量。在整个过程中，交互匹配更为关注句子间的匹配关系，对两个句子进行更为细致的匹配。

相比于并列匹配，交互匹配不仅考虑到单个句子中滑动窗口内的词的组合质量，而且同时考虑到来自两个句子组合间的匹配关系的质量。并列匹配的优势在于匹配过程中可以很好的保持两个句子各自的词序信息，因为并列匹配是分别对两个句子在顺序的滑动窗口上进行建模。相对而言，交互匹配的问答匹配过程是学习语句间局部信息的交互模式。此外，由于交互匹配的局部卷积运算和局部最大池化操作都不改变两个句子的局部匹配表示的整体顺序，所以交互匹配模型同样可以保持问句与答案的词序信息。总之，交互匹配通过对问句与答案的匹配模式进行建模，可以学习到两个句子间的局部匹配模式，而这种匹配模式在正常顺序的句子中具备很大的学习价值。

3.3 基于 RNN 的答案自动生成方法

与基于检索式的回复机制对比而言，基于生成式的答案反馈机制是根据当前用户输入信息自动生成由词语序列组成的答案，而非通过检索知识库中用户编辑产生答案语句。这种机制主要是利用大量交互数据对构建自然语言生成模型，给定一个信息，系统能够自动生成一个自然语言表示的回复。其中的关键问题是如何实现这个语言生成模型。

答案自动生成需要解决两个重要问题，其一是句子表示，其二是语言生成。近年来，循环神经网络在语言的表示以及生成方面都表现出了优异的性能 ,尤其是基于循环神经网络的编码-解码架构在机器翻译[31, 32]和自动文摘[51]任务上取得了突破。Shang[52]等人基于 CRU（Gated Recurrent Unit, GRU）[46]循环神

经网络的编码-解码框架，提出了完全基于神经网络的对话模型“神经响应机”（Neural Responding Machine，NRM），该模型用于实现人机之间的单轮对话（single-turn dialog）。NRM 是从大规模的信息对（问题-答案对，微博-回复对）学习人的回复模式，并将学到的模式存于系统的近四百万的模型参数中，即学习得到一个自然语言生成模型。

如图 3.5 所示，NRM 的基本想法是将输入的一句话看作一个单词表示的序列，通过编码器（Encoder），即一个 RNN 模型，将转换成一个中间表示的序列，再通过解码器（Decoder），是另一个 RNN 模型，将转换成一个单词的系列，作为一句话输出。由于 NRM 在编码部分采用一种混合机制，从而使编码得到中间表示的序列不仅能够实现用户语句信息的整体把握，同时还能充分保留句子的细节信息。并且在解码部分采用了注意力(attention)机制[31] ,从而使生成模型可以相对容易的掌握问答过程中的复杂交互模式。[52]中的实验结果表明基于生成式的问答机制与基于检索式的答案反馈机制各具特点：在表达形式个性化的微博数据上 ,生成式比检索式的准确率会高一些 ,检索系统的准确率是 70% ,生成系统的准确率是 76%。但是 ,生成式得到的答案会出现语法不通 ,连贯性差的问题 ,而检索式的答案来源于真实的微博用户编辑，所以语句的表述更为合理可靠。

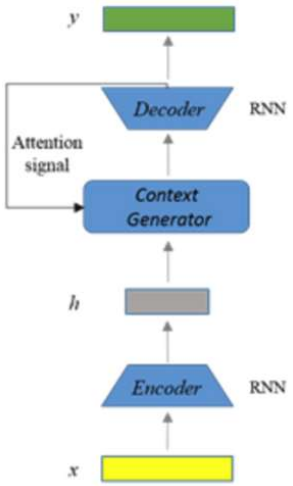

图 3.5 基于编码-解码结构的答案生成模型

目前，NRM 以及 Google 的 Neural Conversational Model（NCM）[53]主要还是在对复杂语言模式记忆和组合上层面上实现语言生成，尚无法在交互过程使用外界的知识。例如，在对“五一期间杭州西湖相比去年怎么样吗？”这样的句子，无法给出真实的状况（旅游人数的对比结果）相关的回复。虽然如此，但是 NRM 和 NCM 的真正意义在于初步实现了类人的语言自动反馈，因为此前的近几十年，研究人员不懈努力而生成的问答或对话系统（dialogue model），大都是基于规则和模板，或者是在一个较大的数据库中进行搜索，而这种两种方式并非真正的产生反馈，并且缺乏有效的语言理解和表示。这往往是由于模板/例子的数量和表示的局限性，这些方式在准确性和灵活性上都存在一定不足，很难兼顾语言的自然通顺和语义内容上的匹配。

4 结语

本文简单介绍了问答系统的发展历程、基本体系结构。并针对问答系统所需解决的关键问题，介绍了基于深度神经网络的语义表示方法，不同匹配架构的语义匹配模型，以及答案生成模型。当前深度学习在解决

问答领域中的关键问题取得了不错的效果，但是问答系统的技术研究仍然存在有待解决问题，比如，如何理解连续交互问答场景下的用户提问，例如与 Siri 系统交互中的语言理解。以及如何学习外部语义知识，使问答系统能够进行简单知识推理回复关系推理性问题，例如“胸闷总咳嗽，上医院应该挂什么科”。再者，随着最近注意（attention）机制、记忆网络（Memory Network）[54,55]在自然语言理解，知识推理上的研究推广，这也必将给自动问答的研究提供的新的发展方向和契机。

参 考 文 献

[1] Terry Winograd. Five Lectures on Artificial Intelligence [J]. Linguistic Structures Processing, volume 5 of Fundamental Studies in Computer Science, pages 399- 520, North Holland, 1977.

[2] Woods W A. Lunar rocks in natural English: explorations in natural language question answering [J]. Linguistic Structures Processing, 1977, 5: 521–569.

[3] Dell Zhang and Wee Sun Lee. Question classification using support vector machines. In SIGIR, pages 26–32. ACM, 2003

[4] Xin Li and Dan Roth. Learning question classifiers. In COLING, 2002

[5] Hang Cui, Min-Yen Kan, and Tat-Seng Chua. Unsupervised learning of soft patterns for generating definitions from online news. In Stuart I. Feldman, Mike Uretsky, Marc Najork, and Craig E. Wills, editors, Proceedings of the 13th international conference on World Wide Web, WWW 2004, New York, NY, USA, May 17-20, 2004, pages 90–99. ACM, 2004.

[6] Clarke C, Cormack G, Kisman D, et al. Question answering by passage selection (multitext experiments for TREC-9) [C]//Proceedings of the 9th Text Retrieval Conference(TREC-9), 2000.

[7] Ittycheriah A, Franz M, Zhu W-J, et al. IBM ’ s statistical question answering system[C]//Proceedings of the 9th Text Retrieval Conference (TREC-9), 2000.

[8] Ittycheriah A, Franz M, Roukos S. IBM’ s statistical question answering system—TREC-10[C]//Proceedings of the 10th Text Retrieval Conference (TREC 2001), 2001.

[9] Lee G G, Seo J, Lee S, et al. SiteQ: engineering high performance QA system using lexico-semantic pattern.

[10] Tellex S, Katz B, Lin J, et al. Quantitative evaluation of passage retrieval algorithms for question answering[C]// Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR ’ 03). New York, NY, USA: ACM, 2003:41–47.

[11] Jiwoon Jeon, W. Bruce Croft, and Joon Ho Lee. Finding similar questions in large question and answer archives. In Proceedings of the 2005 ACM CIKM International Conference on Information and Knowledge Management, Bremen, Germany, October 31 – November 5, 2005, pages 84–90. ACM, 2005.

[12] S. Riezler, A. Vasserman, I. Tsochantaridis, V. Mittal, Y. Liu, Statistical machine translation for query expansion in answer retrieval, in: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, Association for Computational Linguistics, Prague, Czech Republic, 2007, pp. 464–471.

[13] M. Surdeanu, M. Ciaramita, H. Zaragoza, Learning to rank answers on large online qa collections., in: ACL, The Association for Computer Linguistics, 2008, pp. 719–727.

[14] A. Berger, R. Caruana, D. Cohn, D. Freitag, V. Mittal, Bridging the lexical chasm: statistical approaches to answer-finding, in: SIGIR ' 00: Proceedings of the 23rd annual international ACM SIGIR conference on Research and development in information retrieval, ACM, New York, NY, USA, 2000, pp. 192–199.

[15] Gondek, D. C., et al. "A framework for merging and ranking of answers in DeepQA." IBM Journal of Research and Development 56.3.4 (2012): 14-1.

[16] Wang, Chang, et al. "Relation extraction and scoring in DeepQA." IBM Journal of Research and Development 56.3.4 (2012): 9-1.

[17] Kenneth C. Litkowski. Question-Answering Using Semantic Triples[C]. Eighth Text REtrieval Conference (TREC-8). Gaithersburg, MD. November 17-19, 1999.

[18] H. Cui, R. Sun, K. Li, M.-Y. Kan, T.-S. Chua, Question answering passage retrieval using dependency relations., in: R. A. Baeza-Yates, N. Ziviani, G. Marchionini, A. Moffat, J. Tait (Eds.), SIGIR, ACM, 2005, pp. 400–407.

[19] M. Wang, N. A. Smith, T. Mitamura, What is the jeopardy model? a quasisynchronous grammar for qa., in: J. Eisner (Ed.), EMNLP-CoNLL, The Association for Computer Linguistics, 2007, pp. 22–32.

[20] K. Wang, Z. Ming, T.-S. Chua, A syntactic tree matching approach to finding similar questions in community-based qa services, in: Proceedings of the 32Nd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ' 09, 2009, pp. 187–194.

[21] Hovy, E.H., U. Hermjakob, and Chin-Yew Lin. 2001. The Use of External Knowledge of Factoid QA. In Proceedings of the 10th Text Retrieval Conference (TREC 2001) [C], Gaithersburg, MD, U.S.A., November 13-16, 2001.

[22] Jongwoo Ko, Laurie Hiyakumoto, Eric Nyberg. Exploiting Multiple Semantic Resources for Answer Selection. InProceedings of of LREC(Vol. 2006).

[23] Kasneci G, Suchanek F M, Ifrim G, et al. Naga: Searching and ranking knowledge. IEEE, 2008:953-962.

[24] Zhang D, Lee W S. Question Classification Using Support Vector Machines[C]. Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2003. New York, NY, USA: ACM, SIGIR’03.

[25] X. Yao, B. V. Durme, C. Callison-Burch, P. Clark, Answer extraction as sequence tagging with tree edit distance., in: HLT-NAACL, The Association for Computer Linguistics, 2013, pp. 858–867.

[26] C. Shah, J. Pomerantz, Evaluating and predicting answer quality in community qa, in: Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR ’10, 2010, pp. 411–418.

[27] T. Mikolov, K. Chen, G. Corrado, J. Dean, Efficient estimation of word representations in vector space, CoRR abs/1301.3781.

[28] Socher R, Lin C C, Manning C, et al. Parsing natural scenes and natural language with recursive neural networks[C]. Proceddings of International Conference on Machine Learning. Haifa, Israel: Omnipress, 2011: 129-136.

[29] A. Graves, Generating sequences with recurrent neural networks, CoRR abs/1308.0850.

[30] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[C]. Proceedings of ACL. Baltimore and USA: Association for Computational Linguistics, 2014: 655-665.

[31] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate [J]. arXiv, 2014.

[32] Sutskever I, Vinyals O, Le Q V V. Sequence to Sequence Learning with Neural Networks[M]. Advances in Neural Information Processing Systems 27. 2014: 3104-3112.

[33] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions[C]. EMNLP 2011

[34] Tang D, Wei F, Yang N, et al. Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification[C]. Proceedings of the 52nd Annual Meeting of the Association for

Computational Linguistics (Volume 1: Long Papers). Baltimore, Maryland: Association for Computational Linguistics, 2014: 1555-1565.

[35] Li J, Luong M T, Jurafsky D. A Hierarchical Neural Autoencoder for Paragraphs and Documents[C]. Proceedings of ACL. 2015.

[36] Kim Y. Convolutional Neural Networks for Sentence Classification[C]. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha, Qatar: Association for Computational Linguistics, 2014: 1746–1751.

[37] Zeng D, Liu K, Lai S, et al. Relation Classification via Convolutional Deep Neural Network[C]. Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers. Dublin, Ireland: Association for Computational Linguistics, 2014: 2335–2344.

[38] L. Yu, K. M. Hermann, P. Blunsom, and S. Pulman. Deep learning for answer sentence selection. CoRR, 2014.

[39] B. Hu, Z. Lu, H. Li, Q. Chen, Convolutional neural network architectures for matching natural language sentences., in: Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, K. Q. Weinberger (Eds.), NIPS, 2014, pp. 2042–2050.

[40] A. Severyn, A. Moschitti, Learning to rank short text pairs with convolutional deep neural networks., in: R. A. Baeza-Yates, M. Lalmas, A. Moffat, B. A. Ribeiro-Neto (Eds.), SIGIR, ACM, 2015, pp. 373-382.

[41] Wen-tau Yih, Xiaodong He, and Christopher Meek. 2014. Semantic parsing for single-relation question answering. In Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics, pages 643–648. Association for Computational Linguistics.

[42] Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question Answering over Freebase with Multi-Column Convolutional Neural Networks. In Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics (ACL) and the 7th International Joint Conference on Natural Language Processing.

[43] Hochreiter S, Bengio Y, Frasconi P, et al. Gradient flow in recurrent nets: the difficulty of learning long-term dependencies[M]. A Field Guide to Dynamical Recurrent Neural Networks. New York, NY, USA: IEEE Press, 2001.

[44] Hochreiter S, Schmidhuber J. Long Short-Term Memory[J]. Neural Comput., 1997, 9(8): 1735-1780.

[45] Graves A. Generating Sequences With Recurrent Neural Networks[J]. CoRR, 2013, abs/1308.0850.

[46] Chung J, Gülçehre Ç, Cho K, et al. Gated Feedback Recurrent Neural Networks[C]. Proceedings of the 32nd International Conference on Machine Learning (ICML-15). Lille, France: JMLR Workshop and Conference Proceedings, 2015: 2067-2075.

[47] D.Wang, E. Nyberg, A long short-term memory model for answer sentence selection in question answering., in: ACL, The Association for Computer Linguistics, 2015, pp. 707–712.

[48] Malinowski M, Rohrbach M, Fritz M. Ask your neurons: A neural-based approach to answering questions about images[C]//Proceedings of the IEEE International Conference on Computer Vision. 2015: 1-9.

[49] Gao H, Mao J, Zhou J, et al. Are You Talking to a Machine? Dataset and Methods for Multilingual Image Question[C]//Advances in Neural Information Processing Systems. 2015: 2287-2295.

[50] Sun M S. Natural Language Procesing Based on Naturaly Annotated Web Resources [J]. Journal of Chinese Information Processing, 2011, 25(6): 26-32.

[51] Hu B, Chen Q, Zhu F. LCSTS: a large scale chinese short text summarization dataset[J]. arXiv preprint arXiv:1506.05865, 2015.

[52] Shang L, Lu Z, Li H. Neural Responding Machine for Short-Text Conversation[C]. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China: Association for Computational Linguistics, 2015: 1577-1586.

[53] O. Vinyals, and Q. V. Le. A Neural Conversational Model. arXiv: 1506.05869，2015.

[54] Kumar A, Irsoy O, Su J, et al. Ask me anything: Dynamic memory networks for natural language processing[J]. arXiv preprint arXiv:1506.07285, 2015.

[55] Sukhbaatar S, Weston J, Fergus R. End-to-end memory networks[C]//Advances in Neural Information Processing Systems. 2015: 2431-2439.