清 华 大 学

综 合 论 文 训 练

题目：基于双向Attention机制的中文问题答案抽取算法研究

系 别：计算机科学与技术系

专 业：计算机科学与技术

姓 名：周建宇

指导教师：徐华 副教授

2017 年 6 月 7 日

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：学校有权保留学位论文的复印件，允许该论文被查阅和借阅；学校可以公布该论文的全部或部分内容，可以采用影印、缩印或其他复制手段保存该论文。

(涉密的学位论文在解密后应遵守此规定)

签 名： 导师签名： 日 期：

中文摘要

问答作为自然语言处理领域中最重要的研究方向之一，一直得到计算机科学界的高度关注。机器问答也被学界一直认为是下一代搜索引擎的发展趋势，高效、精准的自动问答对信息的高效获取和传播具有重要意义。

自计算机诞生依赖，对机器问答的研究就从未间断。机器问答的核心是自然语言处理，其发展方向也随自然语言处理技术的发展而不断更新。从早期的基于计算语言学的统计模型发展为如今的基于数据驱动的深度学习模型，问答效果也在不断提升。目前学界绝大部分研究都是基于英文问答的，而中文问答领域的研究与应用仍有很多不足。

问答的种类繁多，本文专注于根据文本并从中抽取问题答案（也称作机器阅读理解）的方法研究。本文借鉴了目前该领域应用效果最好的基于双向Attention机制的英文问答（阅读理解）算法，并将其加以改进和优化，以应用到中文问答场景。为了完成这一目标，本文的主要工作有：

1. 实现了基于双向Attention机制的英文问答算法。
2. 设计并实现了基于翻译机制的可应用于中文问答场景的双向Attention算法。
3. 设计并实现了基于中文训练语料库的中文双向Attention算法。
4. 对比了基于翻译与基于中文训练语料库的两种Attention算法在不同中文问答场景下的优劣并分别对两种算法进行了改进优化。
5. 实现了基于双向Attention算法的中文问答平台，该平台支持用户上传和编辑存在问题答案的文本，平台可基于该文本针对用户问题产生答案。

关键词：问题答案抽取；中文问答；双向Attention；机器阅读理解

ABSTRACT

As one of the most important research field for Natural Language Processing(NLP), Question Answering has always been a hot topic in computer science. Question Answering is also regarded as the next generation search engine. Offering precise answer effectively has a great significance on the effective acquirement and spread of information.

Ever since the birth of computer, the research for Question Answering has never been stopped. The core technique of Question Answering is NLP. As a result, the development of Question Answering is closely related to the improvement of NLP. From the initial statistical language computing method to the deep learning method, the performance of Question Answering task has been improved greatly. However, most research of Question Answering lies on English field, so there are still a lot of word needed to be done in Chinese Question Answering.

Question Answering is a broad field with various types. To be more specific, this paper aims at the research of extract answers based on given contexts, which is also known as Machine Reading. This paper learns from the most advanced method for English Machine Reading, which is known as the bi-directional attention flow, and propose some improvements, to make it better for Chinese Machine Reading. To achieve this goal, this paper has the following main contributions:

1. We implement an English Question Answering algorithm based on bi-directional attention flow mechanism.
2. We design and implement a Chinese Question Answering algorithm based on translation and bi-directional attention flow mechanism.
3. We design and implement a Chinese Question Answering algorithm based on original Chinese training corpus.
4. We compare these two Chinese Question Answering algorithm in different application scenario and propose improvements respectively.
5. We implement a bi-directional-attention-flow based Chinese Question Answering platform, which supports users upload and edit contexts and answer questions based on them.

Keywords：Answer Extraction; Chinese Question Answering; Bi-directional Attention Flow; Machine Reading

目 录

[中文摘要 I](#_Toc484338826)

[ABSTRACT II](#_Toc484338827)

[第1章 引 言 1](#_Toc484338828)

[1.1 研究背景 1](#_Toc484338829)

[1.1.1 问答概述 1](#_Toc484338830)

[1.1.2 问答发展历程 1](#_Toc484338831)

[1.1.3 问答系统与问题分类 2](#_Toc484338832)

[1.2 研究现状 3](#_Toc484338833)

[1.2.1 问答范式概述 3](#_Toc484338834)

[1.2.2 基于信息检索的问答范式 4](#_Toc484338835)

[1.2.3 基于知识库的问答范式 7](#_Toc484338836)

[1.3 本文主要贡献 8](#_Toc484338837)

[第2章 预备知识 9](#_Toc484338838)

[2.1 卷积神经网络 9](#_Toc484338839)

[2.2 循环神经网络 10](#_Toc484338840)

[2.3 长短期记忆神经网络 10](#_Toc484338841)

[2.4 Attention机制 10](#_Toc484338842)

[第3章 研究内容 11](#_Toc484338843)

[3.1 实现基于双向attention的英文问答算法 11](#_Toc484338844)

[3.2 研究基于双向attention与翻译相结合的中文问答算法 11](#_Toc484338845)

[3.3 研究基于双向attention与中文训练语料库的中文问答算法 11](#_Toc484338846)

[3.4 实现基于双向attention算法的中文问答平台 11](#_Toc484338847)

[第4章 算法详述 11](#_Toc484338848)

[4.1 模型概述 11](#_Toc484338849)

[4.2 字符编码层 12](#_Toc484338850)

[4.3 词语编码层 13](#_Toc484338851)

[4.4 短语编码层 13](#_Toc484338852)

[4.5 注意流层 13](#_Toc484338853)

[4.6 建模层 14](#_Toc484338854)

[4.7 输出层 15](#_Toc484338855)

[4.8 模型训练 15](#_Toc484338856)

[4.9 模型测试 16](#_Toc484338857)

[第5章 实验结果与分析 17](#_Toc484338858)

[5.1 中文embedding维度对准确率的影响 17](#_Toc484338859)

[5.2 是否使用pre-trained embedding对准确率的影响 17](#_Toc484338860)

[5.3 卷积神经网络 filter size对准确率的影响 17](#_Toc484338861)

[5.4 卷积神经网络卷积层数对准确率的影响 17](#_Toc484338862)

[5.5 相似度矩阵算法对准确率的影响 17](#_Toc484338863)

[第6章 问答数据集 17](#_Toc484338864)

[6.1 斯坦福问答数据集 17](#_Toc484338865)

[6.2 哈工大填空型中文阅读理解数据集 17](#_Toc484338866)

[6.3 哈工大填空式中文阅读理解数据集 17](#_Toc484338867)

[6.4 填空型问答数据的自动生成技术 17](#_Toc484338868)

[第7章 总结与展望 18](#_Toc484338869)

[7.1 本文工作的总结 18](#_Toc484338870)

[7.2 未来工作的展望 18](#_Toc484338871)

[插图索引 19](#_Toc484338872)

[表格索引 20](#_Toc484338873)

[参考文献 21](#_Toc484338874)

[致 谢 22](#_Toc484338875)

[声 明 23](#_Toc484338876)

[附录A 外文文献书面翻译 24](#_Toc484338877)

# 引 言

## 研究背景

### 问答概述

问答（Question Answering）是计算机科学领域的一个重要研究方向，与信息检索、自然语言处理等技术密切相关。问答的最终目标是构建一个能够自动回答人类以自然语言提出的各种问题的系统。

传统的问答系统的工作机制是根据问题，从一个结构化的数据库（通常是知识库）中抽取和组织答案。更一般的问答系统还能够从非结构化的知识文档语料中抽取答案。而常见的非结构化知识文档语料包括维基百科、新闻网页等。

问答领域的研究致力于自动回答多种多样的问题，包括事实类、列表类、定义类等等。而按照问答系统的知识获取方式，又可将问答系统分为封闭领域问答系统和开放领域问答系统两类。封闭领域问答系统着重于回答某个特定领域（如医疗领域）的各类问题，这类问答系统的任务相对来说比较简单，由于问题范围较窄，只需通过自然语言处理的方法从该特定领域的本体中挖掘答案即可。另一方面，封闭领域问答系统常常只接受特定种类的问题，如只接受请求描述类的问题而不接受询问步骤类的问题。而随着机器阅读的方法在问答系统中的应用，一些领域（如医疗）已经有了该领域的问答系统，如询问有关阿兹海默症的问答系统。

开放领域问答系统则几乎负责回答一切问题，其回答问题是基于本体于各类已经存在的知识的。这类系统通常有十分丰富的知识预料可供挖掘，但随着问题种类和范围的大大增加，回答问题的难度也越来越高。另一方面如何从庞大繁杂的知识库中高效快速搜寻组织答案也是一大考验。

### 问答发展历程

最早的问答系统当属BASEBALL和LUNAR，二者都为封闭领域问答系统。BASEBALL能够回答关于一年某个时间段内关于美国棒球联盟的问题。LUNAR则能够回答关于月球岩石的地理信息，这些信息由阿波罗探月计划收集。由于当时互联网还没有十分普及，问答领域也很窄，因此尽管当时的硬件资源落后，但BASEBALL（注释1）和LUNAR（注释2）在问题回答准确率上还算令人满意。接下来若干年里，封闭领域问答系统得到了长足发展，这类问答系统几乎都有共同的特征——以数据库或某种特定领域的知识系统为核心，将用户的询问转化为可供数据库查询的SQL（注释3）语句，最终根据SQL语句返回查询结果。

SHRDLU是第一个获得巨大成功的问答系统，它于60年代末70年代初由Terry Winograd开发。其与之前的问答系统最大的不同在于对机器人行为的模拟，可以说是最早的人机对话系统，它实现了早期的人机交互，人可用自然语言提问和发出指令，SHRDLU会依据人的指令做出相应动作或解答问题。当然，问题仅限于特定种类和特定领域，指令的种类也较少，但其意义是重大的。

到了70年代，问答系统更加集中于封闭领域的细化，问答的专业性越来越强，并有了知识库的概念。此时问答系统开始与专家系统对接，致力于针对特定问题产生更可靠且重复性较高的答案。专家系统与现代问答系统已经十分相似，只是内部工作机制不同。专家系统基于高度结构化组织的专家知识库，而现代问答系统则基于对海量非结构化自然语言语料库的统计学方法。

八十年代左右，计算语言学理论的不断完善极大促进了问答系统的发展，使其在自然语言理解方面的能力大大增强。其中的代表如由加州大学伯克利分校的Robert Wilensky的Unix Consultant(UC)系统。该系统负责回答有关Unix操作系统的各类问题。UC依赖与一个十分庞大完整的Unix知识库，几乎涵盖了包含Unix的一切知识，根据用户不同种类的询问，UC可尝试从相关知识点中抽取组成答案。

目前，针对特定领域的高度面向自然语言的问答系统也发展起来，如生命健康领域的EAGLi(注释4)系统。另一方面开放领域问答系统也加速发展，如微软小冰、苹果Siri、麻省理工问答系统Start、IBM的沃森。值得一提的是，在2011年，沃森参加问答类综艺节目《危险边缘》并击败了该节目两位最强选手Brad Rutter和Ken Jennings，堪称问答系统发展的一座里程碑。

### 问答系统与问题分类

目前主流的分类主要依据为问题答案的来源，主要分为“数据库问答”、“常问问题问答”（Frequently Asked Questions, FAQs）、“新闻问答”、“互联网问答”等。由于数据库数据存储组织的高效性，数据库问答系统首先发展起来，其依赖结构化的查询语句与用户进行交互，但用户使用该类问答系统的学习成本较高。FAQ问答系统在企业客服中应用十分普遍，其主要思想是将一些提问频率很高的问题答案统一整理、高效组织，依据用户问题与系统中已有问题的相似度给出系统中存在的答案，这类问答系统的优点是查询速度快，缺点是回答的问题数量比较有限。另一类重要的系统是新闻问答系统，该类系统之所以脱颖而出最主要的原因是数字新闻媒体的普及，如今每天互联网上涌现的海量新闻，其蕴藏的信息量是十分可观的，也是目前公认的作为开放领域问答系统的最好数据来源。关于互联网问答系统，其核心是利用搜索引擎，然后根据用户询问返回若干包含答案信息的相关文档并从中抽取答案，其表现第一依赖于搜索引擎的返回结果，第二依赖于对答案的精确检索，目前还面临很多挑战。

随着问答领域研究的不断深入，对问题的分类也不断细化，目前形成了包括“仿真陈述类问题”（Factoid Question）、“清单类问题”（List Question）、“定义类问题”(Definition Question)、“时间限制类问题”(Temporally Restricted Question)、“序列类问题”(Series of Question)在内等多类问题。其中最为普遍和基本的是“仿真陈述类问题”，这类问题询问有关一段预先给定语料的问题，并从该段语料中抽取若干文字片段组成答案。“清单类问题”顾名思义，即能回答诸如“请列举中国由哪些省份”一类的问题。“定义类”、“时间限制类”、“序列类”问题与字面意思相近，不再赘述。本文研究的问题类型为“仿真陈述类”，即回答一系列基于简单事实、并能用简短精炼的语言回答的问题。

## 研究现状

### 问答范式概述

现代问答系统按照回答问题的方法可分为**基于信息检索的问答**（IR-based question answering）和**基于知识库的问答**两种范式。本文的研究重点是“仿真陈述类”问答，因此下文重点均为两类范式在该类问答中的应用。

基于信息检索的问答范式也可以说是基于文本的。这种问答依赖的是互联网海量的文本数据。根据用户询问，利用信息检索技术从海量文档中抽取出与问题答案相关的文本段落。更具体地，这种方法会首先对用户以自然语言提出的问题进行分析，确定最可能的问题类型（通常是诸如人物、地点、时间等），再形成可供搜索引擎接受的询问（query）。搜索引擎根据询问会返回一个依据答案相关度排序的文档列表，最终系统会将抽取可能的候选答案文本并依据相关程度返回给用户。

第二种基于知识库的问答范式，我们则首先需要对用户询问进行一种形式化的语义表示，是的用户询问变成一种可计算的表达。形式化语义表示的方式多种多样，但其最终目标都是利用这种表示去进行数据库查询。数据库可以是多种多样的，如科学事实数据库或地理信息数据库。各种数据库都需要符合一定语法规则的、逻辑性较强的查询（如SQL语句）。

### 基于信息检索的问答范式

信息检索式问答的目标是从互联网文本中抽取小段文本作为问题答案返回给用户，它能够回答的问题大致如表xxx所示：

|  |  |
| --- | --- |
| 问题 | 答案 |
| 卢浮宫在哪儿？ | 法国巴黎 |
| 问答系统的英文说法是什么？ | Question Answering |
| 中国的流通货币名称是什么？ | 人民币 |
| 杏仁蛋白奶糖中用到的坚果是什么？ | 杏仁 |
| 吕思清演奏什么乐器而出名？ | 小提琴 |
| 中国的国土面积是多少？ | 960万平方公里 |
| 世界上海拔最高的山峰是什么？ | 珠穆朗玛峰 |

（表注释）

通过上表能够看出，此类问答范式比较适合回答的大部分问题都是“仿真陈述类”问题。

如前文所述，基于信息检索的问答范式回答问题的流程大致分为三步：对问题的解析，信息检索、对候选答案的再加工。问题解析部分的主要任务为**形成询问**和**答案类型确定**，询问主要由问题中的关键词构成，这些关键词能够提高搜索引擎检索效率和结果准确性。答案类型确定的主要作用是确定产生答案的命名实体类型，该步骤可以有效降低信息检索缓解的搜索空间。信息检索的主要功能是排序，从海量文档中将文档排序，并将文档内部的段落排序。对候选答案的再加工指的是从已排序的文档和段落中抽取最可能的答案片段的过程，该步骤将最终产生返回给用户的答案。具体流程如图xxx所示：



（图注）

问题解析步骤的第一个任务是形成询问。形成询问通常是从抽取问题中抽取一系列关键词，并在需要的情况下进行扩展得到。形成问询的另外一种方法是句子改写，Lin, J. (2007)（参考文献）提出了一系列改写规则，其核心是将疑问词去掉并改为待填空陈述句，这种方式可以最大程度在文档中匹配到与答案相关的文本。对于答案类型的解析，通常采用的是分层归类的方法。Li and Roth (2005)（参考文献）建立了一套标签式分类体系，在这种分层的标签分类体系下，每一个问题都会首先被赋予一个粗粒度的标签如人物，或是一种复合式细粒度的标签如人物：描述、人物：分组等，具体分类方法如表xxx所示。问题分类的方式很多，既可以通过既定规则、也可以通过监督式机器学习或者融合二者的方法。但现代问题分类的主流方法还是在已经经过人工标注的数据集上进行训练最终产生问题分类器的（Li and Roth, 2002 参考文献）。

信息检索的核心是搜索引擎，可以是面向一系列文档的检索系统，也可以是通用的互联网搜索引擎。若采用的是文档检索系统，则这一步首先进行粗粒度的文档相关度排序。但排序结果在前的文档未必就存在对问题的解答，这是因为系统最终要返回的是一小段文本答案，而不是整个文档，如此粗粒度的相关度排序有可能对接下来的答案抽取产生误导，因此在文档排序的基础上，我们还要进行更细粒度的排序，这通常是章节、段落或者是句子层面的排序。一种简单的方法是，我们采用某种分割算法，将一个文档划分为若干段落，再利用tf-idf(脚注)算法进行相关度排序。另外一种常用的计算相关度的方法是根据段落包含问题关键词的多少来决定，如果能够再更短的句子中包含更多的关键词（即关键词密度大），则相关度较高（Pasca 2003,Monz 2004参考文献）。还有一种比较常见的做法是采用N-gram overlap(Brill et al., 2002参考文献), 其思想是计算问题和段落文本的在n个词语中的最大匹配数。如果采用的是通用互联网搜索引擎（如谷歌），一种普遍的做法是直接将问题输入搜索引擎，依据搜索引擎返回的文档和关键词匹配结果直接抽取相关句子。

接下来最关键的是答案抽取。传统的基于规则的答案抽取方法主要有两类，分别是基于答案类型的抽取（answer-type pattern extraction）和基于N-gram tiling的抽取。基于答案类型的抽取是根据问题解析部分判断的答案所属类型，生成正则表达式，从而从段落中匹配出答案。例如，一个问题的答案是人物类，那么接下来就可以对于候选答案文本进行标签搜索，将所有标签为人物的实体全部提取出来，再利用正则表达式进行进一步匹配最终产生答案。N-gram tiling（Brill et al. 2002, Lin 2007 参考文献）方法主要应用于互联网搜索引擎检索返回的结果中。第一步对于返回结果中包含关键词的片段，我们赋予所有片段中的单词（unigram）、双词(bigram)、三词(trigram)一定的权重，权重与这些gram在所有包含关键词片段中出现的频率有关。接下来是给每一个gram打分，分数与gram跟问题类型的匹配程度有关。最后一步是将得分高的gram拼接起来组成候选答案，一种常见的做法是贪心，即按照得分由高到低依次将有overlap的N-gram拼接产生候选答案，并将候选答案递归拼接，直到产生最终答案，在此过程中会不断淘汰掉组合后得分低的候选答案。

而现在的趋势则是基于端到端的监督式机器学习直接抽取答案，这类方法比之前基于规则的方法在答案准确率上有较大提高且不需要引入大量人为规则，因此近年来受到学界追捧并逐渐发展成为一个相对独立的研究领域——机器阅读。基于神经网络的机器阅读中第一步也是最关键的步骤是字词编码（word/character embedding）。起初使用的是单纯的循环神经网络（Tom´aˇs Mikolov et al.2010参考文献Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model[C]//Interspeech. 2010, 2: 3.），这种网络结构十分适合对语义建模并具有一定的理解能力。随后一种特殊的循环神经网络——长短期记忆神经网络（Cheng J et al. 2015参考文献Cheng J, Dong L, Lapata M. Long short-term memory-networks for machine reading[J]. arXiv preprint arXiv:1601.06733, 2015.）因其对语言良好的记忆特性被广泛适用于字词编码和语义理解中。为了进一步提高答案抽取效果，一种是在结合LSTM（脚注）字词编码的基础上同时使用CNN（脚注）进行字符层面的编码（Zhang X et al. 2015参考文献Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. 2015: 649-657.），并将二者结果融合作为字词编码结果。另一种是结合语法分析树的语义编码（Liu R et al.2017参考文献Liu R, Hu J, Wei W, et al. Structural Embedding of Syntactic Trees for Machine Comprehension[J]. arXiv preprint arXiv:1703.00572, 2017.）。两种方法原理不同，但在实际应用中都取得了不错的效果。除了语义编码，在问题和候选语料相关度的计算方面也创造性地运用了一种叫做注意机制（Attention）的方法（Bahdanau D et al. 2014参考文献Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. arXiv preprint arXiv:1409.0473, 2014.），并在此基础上产生了问题到语料与语料文本到问题的双向注意机制（Xiong C et al. 2016参考文献Xiong C, Zhong V, Socher R. Dynamic Coattention Networks For Question Answering[J]. arXiv preprint arXiv:1611.01604, 2016.）。受双向注意机制的启发，在语义编码层面也产生了双向编码的方法（Seo M et al. 2016参考文献Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[J]. arXiv preprint arXiv:1611.01603, 2016.）。目前该方法已经在斯坦福问答数据集（Stanford Question Answering Dataset, SQuAD）上取得了很好的效果。

### 基于知识库的问答范式

基于知识库的问答是指在从数据库中查询答案的问答。这里的数据库通常是关系型数据库（relational database）或者是简单的RDF三元组（RDF triples）（脚注）数据库，目前知名度比较高的基于此类问答的应用有Freebase(脚注)（Bollacker et al., 2008参考文献）和DBpedia(脚注)（Bizer et al., 2009参考文献）。

一种简单的问答方法是填补三元组中的缺失项。如下面的RDF三元组：

|  |  |  |
| --- | --- | --- |
| **Subject** | **predicate** | **object** |
| 中华人民共和国 | 诞生时间 | 公元1949年 |

这样的一个三元组可以用来回答如“中华人民共和国是何时成立的？”或者“哪个国家于1949年成立？”一类的问题。我们能从该问题中挖掘出“……国家是何时诞生”这样的模式。更一般地，我们可以总结出更多常见的模式。若我们已经有大量已经标注过的问题数据，则也可以采用监督学习的方式来学习更多更复杂的模式（Zettlemoyer and Collins, 2005参考文献Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars.）。鉴于很难寻找大规模的训练语料库，也有许多采用半监督或非监督的方法来提取模式的（Faderet al., 2011参考文献Identifying relations for open information extraction.）。另外在扩大模式提取范围的基础上，还产生了同义模式扩充等方法（Berant and Liang 2014参考文献Semantic parsing via paraphrasing），用于最大程度地进行模式匹配。

## 本文主要贡献

本文的研究集中于基于信息检索类的问答，且着重于分析目前表现最好的基于神经网络答案抽取方法在中文语境下的应用。本文将首先实现基于双向注意机制的问答抽取算法（Seo M et al. 2016参考文献Seo M, Kembhavi A, Farhadi A, et al. Bidirectional Attention Flow for Machine Comprehension[J]. arXiv preprint arXiv:1611.01603, 2016.），并结合中文的语言特性对算法进行调整和优化，最终实现一个性能良好的中文问答抽取算法。本文主要选定了两条优化途径，一种是**基于翻译模式的中文问答**，该方法仍旧采用英文语料库进行模型训练，但在算法应用阶段会进行两次中英翻译，即将中文问题和翻译为英文并输入给系统，再将系统产生的答案翻译成中文返回给用户。另外一种则是直接**采用中文语料进行模型训练**，直接产生中文答案，省去了中间翻译环节。这两种方法各有利弊第一种方法中间需要两部翻译转换，增加了算法的开销，同时采用机器翻译具有一定的不准确性，可能会对问题理解产生偏差。第二种方法更为直观，理论上应该会取得更好的效果，但由于缺乏大规模中文训练语料库，因此本实验采用的是由英文翻译为中文的斯坦福问答语料库以及采用填空式生成技术产生的中文问答语料库，训练数据质量必然有所下降，导致对模型的性能产生影响。

本文探索目前主流的基于机器学习的问题答案抽取方法在中文场景下运用的可能性，并取得了一定成绩。同时我们也开发了一个小型的中文问答平台，供有兴趣的研究者测试并提出意见。

# 预备知识

## 卷积神经网络

在对气候变化下过去55年冬小麦生长趋势进行分析时，采用的气象数据取自北京气象站（No.54511）（表2.1）。

* + - * 1. 北京气象站基本信息

| 站点  Station | 站号  No. | 纬度  Latitude  (°) | 经度  Longitude  (°) | 高度  Altitude  (m) |
| --- | --- | --- | --- | --- |
| 北京 Beijing | 54511 | 39.8 | 116.5 | 31.3 |

气象资料包括：日期、气压、日平均气温、日最高气温、日最低气温、相对湿度、降水、日平均风速、日照时数。

因缺少辐射监测资料，由纬度、儒略日、日照时数，依世界粮农组织FAO（Food and Agriculture Organization）提供的方法，进行日平均净辐射和净短波辐射的推算。

1. 太阳的磁偏角

 （2‑1）

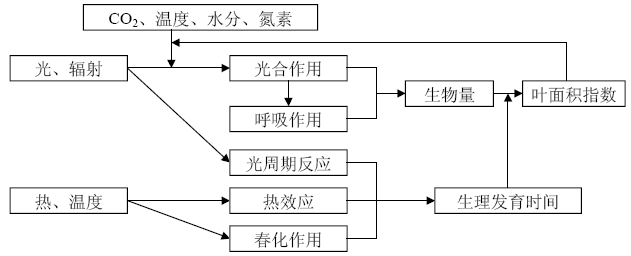
其中，J—在年内的天数。

1. 日落时角度

 （2‑2）

在该模型的研究中，依据CERES模型与WheatGrow模型的基本原理，分别建立了生理发育时间与干物质积累的模拟模型，并根据田间观测资料建立了冬小麦各部分干物质分配的子模型。

模型的基本框架如图2.1所示：



* + - 1. 冬小麦生长发育模型基本结构

## 循环神经网络

随着我国工业化程度提高、社会经济高速稳定发展，我国对纸及纸制品的消费量近年来呈现增速上升趋势。纸制品消费大幅增加，推动我国制浆造纸业产能快速扩张。[[1]](#footnote-1)

1995年至2001年间，纸浆消费总量平均年增长4.9%，2001年为2980万吨；此后纸浆消费增速大幅提高（实际反映的是全社会作为最终产品的纸及纸板消费量剧增），2001年至2004年间，年均增幅高达14.3%，2004年总消费量为4455万吨。造纸工业加速发展势头明显。

## 长短期记忆神经网络

## Attention机制

# 研究内容

## 实现基于双向attention的英文问答算法

## 研究基于双向attention与翻译相结合的中文问答算法

## 研究基于双向attention与中文训练语料库的中文问答算法

## 实现基于双向attention算法的中文问答平台

# 算法详述

## 模型概述

本文将实现一种基于双向注意机制（Bi-Directional Attention Flow）的神经网络结构。这种分层次的网络结构将在不同层次、不同粒度下对文本进行表示，详见见图xxx。其中包括字符、词语、短语在内的三个编码层，其主要作用是对问题和答案候选文本进行不同层次的表示。之后我们利用双向Attention层来产生一种对问题敏感的候选答案所在文本（上下文）表示（query-aware context representation）。这里我们对Attention机制的实现相比于之前主流的方法有了一些改进。首先我们不再将问题和上下文完全转化为单一向量后再计算相关度，而是在每一个生成向量的过程中就进行Attention计算，这样可以减少因为过早地产生编码向量而带来的信息损失。另外，我们采用了双向Attention计算，既计算从问题到上下文的Attention，也计算从上下文到问题的Attention。这样可以避免只进行前者的单项计算而产生的偏差。

该模型于2017年初在斯坦福问答数据集（Stanford Question Answering Dataset, SQuAD）取得了最高准确率，同时也在CNN/DailyMail等数据集上由良好的表现。模型核心为六层神经网络：

1. **字符编码层（Character Embedding Layer）**将用一个接受字符输入的卷积神经网络将问题和上下文中出现的所有词语映射到一个高维向量空间。
2. **词语编码层（Word Embedding Layer）**使用经过预训练的词语编码模型同样将所有词语映射到高维向量空间。
3. **短语编码层（Phrase Embedding Layer）**考虑到相邻若干词语间的作用关系，并结合前两层编码结果对词语编码进行优化表示。
4. **注意流层（Attention Flow Layer）**综合前三层的词语编码表示，并将二者融合产生基于上下文信息的问题表示。
5. **建模层（Modeling Layer）**利用循环神经网络并结合注意流层产生的问题表示对上下文再次扫描。
6. **输出层（Output Layer）**计算候选答案与问题的相关概率，并最终产生答案。

## 字符编码层

我们考虑对问题和上下文文本的形式化表示，令 和 分别表示上下文和问题中的单词，利用字符粒度编码的卷积神经网络（Kim et al.2014参考文献Yoon Kim. Convolutional neural networks for sentence classification），能够产生对词语语义高度抽象的单词向量。具体地，我们首先将26个英文字母和其他符号进行one-hot编码，并将其作为卷积神经网络（CNN）的一维输入，然后采用多层卷积操作和一次最大池化（max pooling）操作产生对单词的稠密向量表示，具体如图4.2.1所示。



（图注）

## 词语编码层

词语编码层的工作与字符编码层相同，均将所有单词映射到高维向量空间，只不过采用的方法有所不同。这里我们采用经过预训练的单词向量GloVe（Pennington et al., 2014参考文献Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation.）直接获得定长向量。

至此我们便获得了由两种不同方法产生的问题文本矩阵和以及上下文文本矩阵和。接下来的工作是融合，我们将上述四个矩阵通过一个两层的高速网络（Highway Network）(Srivastava et al., 2015参考文献 Rupesh Kumar Srivastava, Klaus Greff, and J¨urgen Schmidhuber. Highway networks. arXiv preprint arXiv:1505.00387, 2015.)，该高速网络的输出是经过融合的d维的问题文本矩阵 和d维的上下文文本矩阵，其中和分别代表问题单词数量和上下文单词数量。

## 短语编码层

该层将接受X和Q矩阵作为输入，利用长短期记忆神经网络(Long Short-Term Memory Network, LSTM)( (Hochreiter & Schmidhuber, 1997参考文献Sepp Hochreiter and Jurgen Schmidhuber. Long short-term memory)

## 注意流层

该层的输入是短语编码层产生的上下文矩阵H以及问题矩阵U，本层的目的是将二者融合，产生一种基于问题的上下文表示（query-aware context representation），该表示可表达为矩阵G。为了生成可计算矩阵G，我们需要分别计算从上下文到问题、从问题到上下文两个方向的attention，为了双向计算我们首先需要获得一个相似度矩阵（similarity matrix），该矩阵表达的是短语编码层生成的矩阵H和U中每一个词的相关关系。具体地，表达的是上下文中第t个单词和问题中第j个单词的相似度。相似度矩阵S的计算方法为：

(4-1)

其中是一个标量函数，表示上下文矩阵第t个单词所代表的向量，表示问题矩阵第j个单词所代表的向量。对于的解析式表达并没有一个明确的定义，通常我们可选取，其中，其元素具体数值可通过训练产生，代表基于矩阵元素的乘法。[;]表示向量按行拼接。

**从上下文到问题的attention(Context-to-Query Attention)**表征了对于上下文中的每一个单词，问题中哪一个单词与之相关度最高。我们令 表示上下文中第t个单词对于问题中所有单词的attention 权重，则直观地我们有, 且，相应地，我们接下来获得的基于问题的上下文表示矩阵，这样 即是一个规模的矩阵，该矩阵是基于问题的上下文表示。

**从问题到上下文的attention（Query-to-Context Attention）**表征了对于问题中的每一个单词，上下文中哪一个单词与之相似度最高，这也是该网络最关键的部分。与计算Context-to-Query Attention类似，相应的权重,其中函数是取矩阵中最大元素所在列的列向量。接下来我们就得到了基于上下文的问题表示，该向量将上下文中关于问题最重要的单词进行了加权求和，最终为了计算方便，我们将按列拼接T次，最终得到矩阵。

最后，结合短语编码层生成的矩阵H，我们最终可以得到对于问题敏感的上下文表示矩阵G，

(4-2)

对应与上下文中的第t个单词。对于函数，这里的处理是将其简单看做一若干有关向量的按行拼接，如。当然，一种更好的做法是将看作一个可训练的带参函数（如多层感知机），但简单的矩阵拼接再英文数据集上已经取得了不错的效果。

## 建模层

得到矩阵G后，建模层将进一步捕捉问题与上下文之间的交互关系，可以直观的理解成对带着问题对上下文的再次扫描。我们采用在机器阅读中应用广泛的双向LSTM（Bi-LSTM）扫描矩阵G, 并产生对回答问题最有帮助的矩阵表示，M的每一列代表一个单词，但此时的单词向量既包含上下文信息，也包含问题信息。

## 输出层

该层的结构功能依应用场景（问答、阅读理解）而定。此处应用于仿真陈述类问答，我们的目标是从所给上下文中抽取片段作为答案返回。因此我们要确定该片段的起止位置。我们首先计算片段开始位置的概率分布：

(4-3)

其中，是一个权重可训练矩阵。对于结束位置，我们将矩阵M再次通过一个双向的LSTM得到，接下来我们计算结束位置的概率分布：

(4-4)

至此我们只需选出概率最大的和中的元素直接将截取答案并返回即可。

## 模型训练

对于神经网络的模型训练我们首先要定义训练的损失函数，由于我们采用的是监督学习，将采用直观的概率分布损失之和作为损失函数，具体表达为：

(4-5)

这里表示该模型中所有可以训练的参数，N代表训练集的数据规模，和代表第i个样本的真正答案实际的起止位置。

最终我们选取答案文本范围为，使其的值最大。

## 模型测试

我们在斯坦福问答数据集（SQuAD）、哈工大填空式中文阅读理解数据集和微软机器阅读理解数据集（Microsoft Machine Reading Comprehension Dataset, MS-MARCO）上对模型进行了评测。

对于前两种数据集，我们采用Exact Match(EM) score和F1 score作为模型评测指标。EM score衡量模型预测的答案文本与实际答案文本的实际单词匹配率。F1 score是召回率和准确率的调和平均，具体计算方法为：

(4-6)

其中precision表示预测答案文本中正确的单词数与文本总单词数的比率，recall表示正确单词数和实际答案单词数的比率。

对于MS-MARCO数据集，其数据特性与前两种数据集不同，其数据全部从微软必应搜索引擎获取，所有问题和答案均来源于现实世界，答案也全部为人工手动填写，因此答案文本很可能并非候选文档中的文本片段。这里我们分别采用ROUGE-L（脚注）和BLEU1（脚注）两个指标衡量从候选文本片段选出文本片段作为答案，该片段应该与实际人工填写的答案具有最高的ROUGE-L和BLEU1值。通过这种方法，我们仍旧能够采用EM score和F1 score对模型进行评价。

# 实验结果与分析

## 实验参数设定

本文实验分为两大部分，一部分是研究基于翻译的中文问答抽取算法的表现，另一部分是研究基于中文语料训练的问答抽取算法的表现，选取的数据集分别是经过翻译的SQuAD和MS-MARCO数据集。两种方法采用的核心算法相同，但在数据处理和部分参数上略有不同。两部分实验采用的实验参数大致相同，并采用控制变量法研究部分参数对实验结果的影响。具体参数设定如表xxx所示。

表xxx

|  |  |  |
| --- | --- | --- |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |
|  |  |  |

## Embedding维度对准确率的影响

根据第四章介绍的神经网络模型，模型的前三层分别对问题和上下文进行不同粒度的编码（embedding），而词语编码层输出的向量（word embedding）维度从根本上决定了语义表达的精准度，也会持续影响到后续各层的效率，因此我们首先研究Embedding维度对模型效果的影响。

对于基于翻译机制的问答抽取算法，我们对字符编码层的过滤器（filter）数量进行调整，filter数量直接决定了字符编码层输出向量维度。前文也曾提到，原生的问答算法在词语编码层采用了经过预训练的GloVe，依据GloVe所具备向量维度：50、100、200和300，我们分别选取第二层输出维度（d）为50、100、200和300，分别得到模型在SQuAD和MS-MARCO数据集上的表现如图xxx所示。能够看到，模型在维度在d=100的情况下表现最好，在SQuAD数据集上EM值达到了70%，F1值达到了70%。分析原因不难看出，当d=50维度较低，可能无法将语义充分表达，而维度过高一方面会给计算增加负担，另一方面存储的冗余信息会使需要训练的参数量大大增加，导致模型不宜收敛，效果大大折扣。因此d=100是一个比较符合实际的结果。

（图注）

在d=100的情况下，我们也统计了不同上下文本长度对回答准确率的影响，如图xxx所示。随着文档长度的上升，准确率下降很快，在词数小于100时能够达到90%以上，而当词数超过300以后只能维持在60%左右，说明该模型对于回答需要浏览大量文档的问题效果仍有待提升。

（图注）

对于基于中文语料库训练的方法，我们将分别采用分词和分字的方法模拟算法第一层和第二层的单词输入，即一种方法利用分词工具把中文语料进行分词并将分词结果看作英文单词集合进行训练；另一种方法是直接将汉字按字直接分隔并将结果看作英文单词集合进行训练。值得注意的是，第二层我们无法获得经过预训练的单词向量，而是利用谷歌发布的词向量训练工具包word2vec（脚注<https://code.google.com/p/word2vec/>）（Mikolov T, Chen K, Corrado G, et al参考文献Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[J]. arXiv preprint arXiv:1301.3781, 2013.）在搜狗发布的中文全网新闻语料库SogouCA（脚注）上进行中文词向量训练。我们同样选取词向量输出维度(d)为50、100和200进行测试，效果如图xxx所示，同样地，在不同文档语料长度下的测试效果如图xxx所示。可以看到两种与翻译机制的表现类似，两种方法都是在d=100时表现最好，且均在文档语料长度（L）较小时准确率最高。另一方面也能够比较明显的发现基于中文语料训练的问答抽取算法的稳定性较好，在L较大时准确率优于翻译方法。

## 卷积神经网络 filter size对准确率的影响

## 卷积神经网络卷积层数对准确率的影响

## 相似度矩阵算法对准确率的影响

# 问答数据集

## 斯坦福问答数据集

## 哈工大填空型中文阅读理解数据集

## 哈工大填空式中文阅读理解数据集

## 填空型问答数据的自动生成技术

# 总结与展望

## 本文工作的总结

## 未来工作的展望



插图索引

图2.1 冬小麦生长发育模型基本结构 4

表格索引

表2.1 北京气象站基本信息 3

参考文献

1. 辛希孟. 信息技术与信息服务国际研讨会论文集: A集［C］. 北京: 中国社会科学出版社，1994.
2. 程根伟，1998年长江洪水的成因与减灾对策［M］//许厚泽，赵其国. 长江流域洪涝灾害与科技对策. 北京：科学出版社，1999：32-36
3. 李晓东，张庆红，叶瑾琳，罗云．气候学研究的若干理论问题 [J]．北京大学学报：自然科学版，1999，35（1）：101-106
4. 张志祥. 间断动力系统的随机扰动及其在守恒律方程中的应用[D]. 北京: 北京大学数学学院，1998: 5-10.
5. PACS-L：the public-access computer systems forum［EB/OL］ ． Houston, Tex: University of Houston Libraries, 1989［1995-05-17］. http://info.lib.uh.edu./;acsl.html.
6. Dubeck, L. (1990). Science fiction aids science teaching. *Physics* *Teacher, 28,* 316-318.

致 谢

致谢对象，原则上仅限于在学术方面对学位论文的完成有较重要帮助的团体和人士。

声 明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签 名：\_\_\_\_\_\_\_\_\_\_\_\_\_ 日 期：\_\_\_\_\_\_\_\_\_\_\_\_\_

附录A 外文文献书面翻译

调研阅读报告题目（或书面翻译题目）

写出至少5000外文印刷字符的调研阅读报告或者书面翻译1-2篇（不少于2万外文印刷符）。

参考文献（或书面翻译对应的原文索引）

1. 辛希孟. 信息技术与信息服务国际研讨会论文集: A集［C］. 北京: 中国社会科学出版社，1994.

附录A说明：“外文资料的调研阅读报告”或“书面翻译”二者择一；若是外文资料的调研阅读报告，请在文中对应“参考文献”；若是书面翻译请在文中对应“书面翻译对应的原文索引”。阅后删除此框及内容。

综合论文训练记录表

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **学生姓名** |  | **学号** |  | **班级** |  |
| **论文题目** |  | | | | |
| **主要内容以及进度安排** | **指导教师签字：**  **考核组组长签字：**  **年 月 日** | | | | |
| **中期考核意见** | **考核组组长签字：**  **年 月 日** | | | | |
| **指导教师评语** | **指导教师签字：**  **年 月 日** | | | | |
| **评阅教师评语** | **评阅教师签字：**  **年 月 日** | | | | |
| **答辩小组评语** | **答辩小组组长签字：**  **年 月 日** | | | | |

**总成绩：**

**教学负责人签字：**

**年 月 日**

1. 本节各类纸浆及纸制品产量、进口数量数据 [↑](#footnote-ref-1)