

神经机器阅读理解研究综述

摘 要: 最后写

文献标识码: A 中图分类号: TP391

Overview of Studies on Neural Machine Reading Comprehension

Abstract: Machine Reading Comprehension aims to make machines comprehend the natural language documents, which is an important research direction in the field of natural language processing. With the development of deep learning technology and release of large scale datasets, the research on the field of machine reading comprehension has made great breakthroughs. With the emergence of pre-trained model in natural language processing in recently years, it promotes the development of machine reading comprehension once again. This paper mainly make a survey from four aspects over the development of machine reading comprehension in recently years: to introduce the definition of machine reading comprehension tasks and its corresponding datasets; to analyze classical model in the field of machine reading comprehension which based on attention mechanism or reasoning structure as well as the currently popular pre-trained model; to discuss more complicated machine reading comprehension tasks; to summarize the existing problem and look into the future research trend about machine reading comprehension.

Key words: machine reading comprehension; natural language processing; pre-trained model; attention mechanism; reasoning structure

1 引言

自然语言处理是人工智能的重要分支,是实现人工智能的核心技术,主要研究如何处理、分析以及应用自然语言。教会机器阅读文本并且理解人类语言是自然语言处理领域的重要任务,机器阅读理解(Machine Reading Comprehension, MRC)的目标就是利用自然语言处理技术使得计算机能够像人一样阅读并且理解文章,它有着很多应用场景,如搜索引擎中的智能问答,电商领域的智能客服以及对话系统等。

早在 20 世纪 70 年代,学者们就已经意识到机器阅读技术是测试计算机理解人类语言的关键方法,如 Lehnert^[1] 构建的 QUALM 系统。但是系统非常小并且局限于手工编码,很难推广到更大的领域。1999 年出现首个自动阅读理解测试系统 Deep Read,该系统以故事为基础衡量阅读理解任务,利用词袋模型和人工编写规则进行模式匹配,准确率可以达到 40% 左右,但是由于依靠手工规则,模型泛化能力很差。总之由于没有合适的文本表示方法,数据集的规模又比较小,而且传统方法更多的是在句子级别粒度上回答问题,核心是特征提取,包括的特征主要有依存句法、词频共现和语篇关系等。但是由于特征需要人工制定,鲁棒性差。而且只能从文本中提取浅层特征不能对文本推理,因此达不到真正的阅读理解,因此早期的 MRC 系统难以获得期望的性能也不能实际的应用。

随着深度学习的兴起以及 NLP 领域技术的发展,为了弥补传统 MRC 技术的缺陷,2015 年 DeepMind 研究员 Hermann 等人^[2]提出使用神经网络模型解决 MRC 任务。他们提出一种新颖且代价小的解决方案用来构建 MRC 领域大规模的监督训练数据,基于此方案他们构建了规模比以往数据集都要大的阅读理解数据集 CNN&Daily Mail,并且提出两个基于神经网络和注意力机制构建的模型(Attentive Reader, Impatient Reader),模型在 CNN&Daily 数据集上效果远超过传统方法的效果。这项工作可以视为机器阅读理解领域的奠基性工作。此后越来越多的学者在这两个模型的基础上构建效果更好的基于神经网络的机器阅读理解模型,也简称为神经机器阅读理解模型。

然而 CNN&Daily Mail 数据集的构造方式使得数据集有一定的噪声,此外数据集的问题是自动生成而且答案仅仅是原文中的某个实体名词,对模型的理解能力要求不高。为了解决这些限制问题,Rajpurkar 等人^[3]发布了大规模数据集 SQuAD。SQuAD 数据集的问题由众包工人基于文章人工生成,并且问题的答案来源于原文中某一片段,而不像 CNN&Daily Mail 数据集仅仅是某个实体单词。由于数据集规模大、质量高,易于评估且具有很好的应用价值等特点获得了广泛的关注度,极大地推动了 MRC 领域的发展,很多经典的神经机器阅读理解模型(如 Match-LSTM^[4], DCN^[5], BiDAF^[6], R-Net^[7])都是在 SQuAD 数据集上构建出来的。

为了使得阅读理解任务更加的贴近真实场景,微软亚洲研究院发布了来源于真实应用场景下的数据集 MS MARCO^[8],问题收集自搜索引擎日志,答案由人工生成。在 MS MARCO 发布后,机器阅读理解领域涌现出很多的高质量、大规模、不同来源的数据集(如 RACE^[9], NarrativeQA^[10], CoQA^[11]等)。每一个新的阅读理解任务的提出都会伴随着相关数据集的发布,每一个数据集也都会在原有数据集的基础上增加适当的难度,目的是使得数据集可以真正的考察模型的阅读理解能力。

自 2018 年随着 GPT^[47]、BERT^[48]等预训练模型的出现,刷新了众多机器阅读理解任务上模型的最佳性能,甚至在某些数据集上模型的表现超过人类水平。

本文的目的是对从 2015 年来机器阅读理解领域的研究任务、相关数据集以及模型做综述。整体安排如下:第二章概述机器阅读理解任务并且简介各个任务相关的数据集;第三章介绍神经机器阅读理解模型,包括预训练模型出现之前的经典机器阅读理解模型以及基于预训练模型的机器阅读理解模型;第四章我还没想好名字;第五章讨论机器阅读理解的应用、面临的主要问题以及未来的研究方向;第六章对全文作总结。

2 机器阅读理解任务概述

机器阅读理解(MRC)任务是为了使得计算机具有对自然语言文本理解的能力,像人类一样阅读并且理解一篇文章。MRC 可以用一个三元组 $\langle D, Q, A \rangle$ 来描述,其中 D 代表文章(Document)¹, Q 表示问题(Question), A 表示答案(Answer),即给定一篇文章 D 和一些与文章 D 相关的问题 Q ,要求模型通过阅读 D 之后给出 Q 的正确答案 A ,建模给定 D 和 Q 的条件下预测 A 的概率: $P(A|D, Q)$ 。

MRC 任务已经从早期的需要阅读一篇文章,答案是单个单词的阅读理解任务,发展到需要阅读多篇文章,并且答案需要从多篇文章中推理出来的阅读理解任务,使得阅读理解任务更加的贴近真实场景。本文按照 Chen^[12]提出的分类方式,根据答案形式的不同将阅读理解任务概括为 4 类任务:填空式、多项选择式、抽取式和自由答案式。下面对这四种类型任务分别进行叙述并介绍相关的数据集。

¹本文中文章(Document)和段落(Passage)是同样的概念

2.1 填空式

填空式阅读理解是指给定一篇文章 D 和一个与文章相关的问题 Q ， Q 是通过删除掉句子中某一个单词构成，要求模型根据 D 能够正确的填写出 Q 缺失的单词 a ，且 $a \in D$ 。填空型数据集的一个样例见表 1。

典型的填空式数据集是由 Google DeepMind 和牛津大学发布于 2015 年的 CNN&Daily Mail^[13] 数据集，这是第一个较大规模的完形填空型阅读理解型数据集。从 CNN²中收集 93k 篇文章，从 Daily Mail³上收集 220k 篇文章。删去句子中的一个命名实体单词，以此作为问题，构建了 (文章-问题-答案) 的三元组形式的语料库作为填空式的阅读理解任务。

相关的数据集还有 Hill 等人^[14] 发布了 (Children’s Book Test, CBT) 数据集，语料库来源于儿童读物 (Gutenberg⁴工程)。与 CNN&Daily Mail 的不同之处是删除的单词不局限于命名实体，还可能删除句子中的名词、动词和介词。CLOTH^[15] 数据集是收集自中国中学生的英语考试试题中的完型填空题型。每一个问题是由专家为了测验学生英语水平而精心设计的，空白处位置的单词通常会考察学生的词汇、语法以及推理能力，因此 CLOTH 数据集相比于 CNN&Daily 和 CBT 更具有挑战性。

表 1 CNN&Daily Mail^[13] 数据集的一个样例
Table 1 An example of CNN&Daily Mail dataset

文章:	What was supposed to be a fantasy sports car ride at Walt Disney World Speedway turned deadly when a Lamborghini crashed into a guardrail. The crash took place Sunday at the Exotic Driving Experience, which bills itself as a chance to drive your dream car on a race-track. The Lamborghini’ s passenger, 36-year-old Gary Terry of Davenport, Florida, died at the scene, Florida Highway Patrol said. The driver of the Lamborghini , 24-year-old Tavon Watson of Kissimmee, Florida, lost control of the vehicle, the Highway Patrol said. (...)
问题:	Officials say the driver, 24-year-old Tavon Watson, lost control of a_____
答案:	Lamborghini

2.2 多项选择式

多项选择式这类任务是对于给定的文章 D 和问题 Q 以及多个候选答案 $A = \{A_1, A_2, \dots, A_n\}$ ，从 A 中选择正确的答案，建模概率： $P(A_i|D, Q)$ ，其中 $A_i \in A$ 。相关的数据集如 MCTest^[16] 和 RACE^[9]。多项选择型数据集的一个样例见表 2。

MCTest 是一个早期提出来的多项选择式数据集，问题形式为四选一。但是由于其规模比较小仅仅包含 500 篇故事文章很难利用神经网络模型来学习，因此 MCTest 通常用来作为验证集或测试集。

RACE⁵数据集是从中国中学生英语考试试题中的阅读理解题型建立的数据集。共有将近 2.8 万篇文章以及 10 万个问题，这些问题和候选答案都是由专家生成的，此外 RACE 数据集涵盖多个领域如新闻、故事、广告、传记等等，由于其类型的多样性因此可以更好的评估机器的阅读理解能力。

²www.cnn.com

³www.dailymail.co.uk

⁴<https://www.gutenberg.org/>

⁵www.cs.cmu.edu/glail/data/race/

表 2 RACE^[9] 数据集的一个样例
Table 2 An example of RACE dataset

文章:	Runners in a relay race pass a stick in one direction. However, merchants passed silk, gold, fruit, and glass along the Silk Road in more than one direction. They earned their living by traveling the famous Silk Road. .. The Silk Road was made up of many routes, not one smooth path. They passed through what are now 18 countries. The routes crossed mountains and deserts and had many dangers of hot sun, deep snow and even battles...
问题:	The Silk Road became less important because_____
选项:	<p>A. it was made up of different routes</p> <p>B. silk trading became less popular</p> <p>C. sea travel provided easier routes</p> <p>D. people needed fewer foreign goods</p>
答案:	C

2.3 抽取式

抽取式阅读理解任务可以看做是填空式任务的扩展，不像填空式任务仅仅要求答案是原文中的某个单词，抽取式任务要求模型从原文中抽取一段连续的文本作为答案，且答案的长度不固定。这类阅读理解任务是 MRC 领域较为流行的研究方向，主要原因在于从数据集的构建、评测指标以及应用价值等角度上看抽取式阅读理解是最合适的。给出文章 D 和问题 Q ，问题的答案是 D 中的一段连续的单词构成。可以表示为 $P(A|D, Q)$ ，其中 $A = \{t_i, t_{i+1}, \dots, t_{i+k}\} (1 \leq i \leq i+k \leq n)$ ， n 代表 D 中单词的个数， k 代表答案的长度。抽取式阅读理解任务的样例见表 3。

SQuAD^[3] 是抽取式阅读理解的代表性数据集，也是 MRC 领域最为广泛使用的数据集之一，它的提出极大地推动了 MRC 领域的发展。数据集由众包工人根据维基百科上面的文章给出问题，答案来源于文章中某段连续的文本，长度并不固定。SQuAD 含有 536 篇文章，总计十万多个问题-答案对。类似的数据集如 NewsQA^[17]，区别在于 NewsQA 文章来源于 CNN 新闻并且在 NewsQA 中某些问题是没有答案的，这也使得后来 Rajpurkar 等人在 SQuAD 版本上又增加了五万个不可回答的问题构建了数据集 SQuAD 2.0^[18]。

之前的数据集都是给定文章后，由人工构造出与文章相关的问题和答案，但是现实世界中人们通常是先提出问题然后搜寻相关的文章再找到答案。基于这一思想，Joshi 等人^[19] 首先从 trivia 上收集大量的问题-答案对，然后为每一个问题从网页上或者维基百科上搜索出相关的文章，这些文章就是答案的依据。最后构建出包含 65 万多个（问题-答案-文章）三元组的抽取式阅读理解数据集 TriviaQA，这种由问题找文章的数据集构造方式使得问题和文章在句法和词汇上都有着较大的差异性，这使得数据集难度更高。

目前大部分的抽取式数据集对模型的推理能力要求不高，主要原因在于回答这些数据集的问题往往只需要集中于某个句子的上下文或者在一篇段落上推理即可。为了提高模型的多步推理能力，Yang 等人^[20] 发布了 HotpotQA 数据集，每一个问题对应多个段落，问题的答案往往需要在多个段落上逐步推理才能获得，同时要求模型预测回答问题所必须的线索句子（supporting facts）。类似的数据集还有 WIKIHOP^[21]，每一个样本来源于知识库中的三元组，要求模型从多篇维基百科文章中推理然后从多个候选中选出正确的答案。

表 3 SQuAD 1.1^[3] 数据集的一个样例
Table 3 An example of SQuAD 1.1 dataset

文章:	In 1870, Tesla moved to Karlovac, to attend school at the Higher Real Gymnasium , where he was profoundly influenced by a math teacher Martin Sekulić. The classes were held in German, as it was a school within the Austro-Hungarian Military Frontier. Tesla was able to perform integral calculus in his head, which prompted his teachers to believe that he was cheating. He finished a four-year term in three years, graduating in 1873.
问题:	Why did Tesla go to Karlovac?
答案:	attend school at the Higher Real Gymnasium

2.4 自由答案式

抽取式任务这种从原文中抽取一段文本作为答案的约束并不能回答所有问题需要的答案，而且从文章中概括提炼出问题的答案也是更加的符合人们的阅读方式的。于是研究人员开始转向自由答案式阅读理解任务，这类任务的答案是自由形式的，不局限于文章中的某些单词，语法上往往是更加的灵活。可以表示为 $P(A|D, Q)$ ，其中 $A \subseteq D$ 或 $A \not\subseteq D$ 。自由答案式任务的一个样例见表 4。

MS MARCO^[8] 是自由答案式阅读理解数据集的代表之一，由微软通过在必应搜索引擎的日志上收集用户提出的问题，段落是来源于必应搜索引擎的返回的 10 个最相关的查询段落，人工的从这些选择出来的段落中概括提炼出答案。如果不能从给出的 10 个段落中推理出答案，那么这个问题就标记为无答案的问题。DuReader^[22] 是与 MS MARCO 类似的一个中文机器阅读理解数据集，类似于 MS MARCO 数据集的构造方式，问题和文章取自百度搜索和百度知道。

表 4 MS MARCO^[8] 数据集的一个样例
Table 4 An example of MS MARCO dataset

文章 1:	Rachel Carson' s essay on The Obligation to Endure, is a very convincing argument about the harmful uses of chemical, pesticides, herbicides and fertilizers on the environment.
.....	
文章 5:	Carson believes that as man tries to eliminate unwanted insects and weeds; however he is actually causing more problems by polluting the environment with, for example, DDT and harming living things
.....	
文章 10:	Carson subtly defers her writing in just the right writing for it to not be subject to an induction run rampant style which grabs the readers interest without biasing the whole article.
问题:	Why did Rachel Carson write an obligation to endure?
答案:	Rachel Carson writes The Obligation to Endure because she believes that as man tries to eliminate unwanted insects and weeds; however he is actually causing more problems by polluting the environment.

2.5 评估方法

对于不同的 MRC 任务有不同的评估指标。对于填空式任务与多项选择式任务都是属于客观题型，用准确率就可以衡量模型的性能。例如对于测试集中的所有问题 $Q = \{Q_1, Q_2, \dots, Q_m\}$ ，其中 m 代表问题的个数。如果模型预测出来的 m 个答案中有 n 个是正确的，那么模型的准确率是 n/m 。

抽取式任务属于半客观题型，通常用精确匹配 EM (Exact Match) 和 F1 分数来评估模型。EM 评估指标可以看做是准确率的扩展，就抽取式任务来讲，EM 要求预测出来的所有单词要和标准答案的所有单词要完全一致，EM 值才为 1，否则为 0。F1 值的计算方式是一种模糊匹配，它是精确率和召回率之间的调和平均数。精确率是指模型预测的答案中有多大比例的单词是标准答案中的单词。召回率是指标准答案中的单词有多大比例在预测答案中出现。

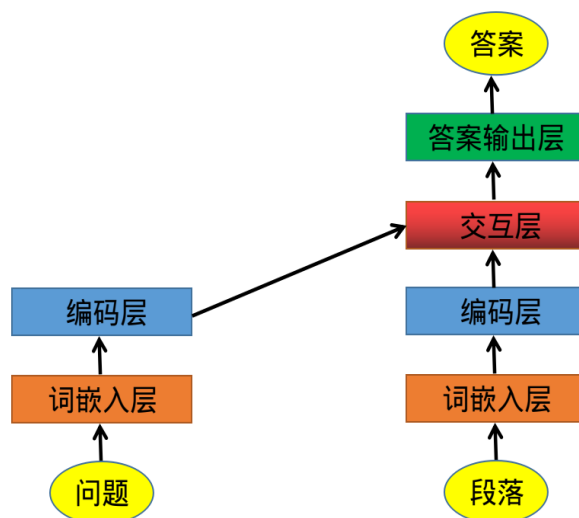
对于自由答案式任务由于其答案形式不固定，一般采用单词水平的匹配率作为评分标准，常用标准用 ROUGE-L^[23] 和 BLEU^[24]。ROUGE-L 用来计算标准答案和预测答案的最长公共子序列 (Longest Common Subsequence, LCS)，BLEU 最初用于评估翻译性能，在应用到 MRC 任务上主要用来衡量预测答案和真实答案之间的相似性。表 5 列举了本章介绍的所有数据集及相应的评估方法。

表 5 MRC 常用数据集对比，Acc 代表准确率
Table 5 Comparison of common dataset in MRC

数据集	发布时间	文章来源/数量	文章类型	问题来源/数量	答案类型	评估指标
CNN&Daily Mail ^[13]	2015	新闻/ 3×10^5	单段落型	人工合成/ 1.4×10^6	填空式	Acc
CBT ^[14]	2015	儿童读物/108	单段落型	人工合成/ 6.8×10^5	填空式	Acc
CLOTH ^[15]	2016	英语考试	单段落型	英语考试/ 1×10^5	填空式	Acc
MCTest ^[16]	2013	儿童读物/500	单段落型	众包/ 2×10^3	多项选择式	Acc
RACE ^[9]	2018	英语考试/ 5×10^4	单段落型	英语考试/ 8.7×10^5	多项选择式	Acc
SQuAD ^[3]	2016	维基百科/536	单段落型	众包/ 1×10^5	抽取式	EM/F1
TriviaQA ^[19]	2017	网页搜索/ 6.6×10^5	多段落型	搜索日志/ 4×10^4	抽取式	EM/F1
SQuAD 2.0 ^[18]	2018	维基百科/536	单段落型	众包/ 1.5×10^5	抽取式	EM/F1
NewsQA ^[17]	2017	新闻/ 1×10^4	单段落型	众包/ 1×10^5	抽取式	EM/F1
HotpotQA ^[20]	2018	维基百科	多段落型/	众包/ 1.13×10^5	抽取式	EM/F1
WIKIHOP ^[21]	2018	维基百科	多段落型/	自动生成，需多步推理	抽取式	EM/F1
MS MARCO ^[8]	2016	搜索引擎/ 2×10^5	多段落型	搜索日志/ 1×10^5	自由答案式	ROUGE-L/BLEU
DuReader ^[22]	2018	搜索引擎/ 1×10^6	多段落型	搜索日志/ 2×10^5	自由答案式	ROUGE-L/BLEU
NarrativeQA ^[10]	2017	小说和电影剧本/ 1.5×10^3	多段落型	众包/ 4.6×10^4	自由答案式	ROUGE-L/BLEU

3 神经机器阅读理解模型

Hermann 等人^[2]所设计的两个基于神经网络和注意力机制的模型可以看作是 MRC 领域的奠基性工作，开创了神经机器阅读理解模型。填空式阅读理解任务本质上可以认为是抽取式阅读理解任务的简化形式，而后续的很多任务如对话形式、开放领域形式、多段落形式的阅读理解任务也都是在抽取式任务的形式上设计模型。抽取式阅读理解任务是机器阅读理解领域的核心问题，本章主要以抽取式任务的 MRC 模型为出发点，安排如下（最后写）。



3.1 经典 MRC 模型

想让机器能够阅读理解文本需要解决以下几个问题：

1. 如何将段落和问题这种文本形式的无结构数据表示为计算机可以处理的形式；
2. 如何根据问题检索出段落中与问题最相关的部分；
3. 如何从检索出来的文章片段中归纳得到答案。

在预训练模型出现之前，用于 MRC 任务的深度学习模型的整体框架主要包括如下几个层：词嵌入层、编码层、交互层、答案输出层，如图 1 所示。词嵌入层的作用是将段落和问题嵌入到低维的向量空间中，用每一个向量表示每一个单词。编码层的作用是编码段落和问题中单词的语义信息，使得每一个单词可以关注到它的上下文。交互层的作用是将段落的语义信息与问题的语义信息融合，让模型学习到段落中与问题最相关的部分。答案输出层的作用是从段落中查找出问题的答案。

3.1.1 词嵌入层

如何将文本有效的表示成计算机可以处理的形式同时可以有效地利用单词之间的语义一直是 NLP 领域的重点问题。早期的 one-hot 形式编码用一个二值向量表示单词，但是存在数据稀疏并且随着单词个数的增加出现维度灾难的问题，此外这种形式的编码也不能够表示出单词之间的语义关系。

Rumelhart 等人^[25]最早提出分布式表示的概念，分布式表示是将单词嵌入到一个低维向量空间中用一个低维度的稠密向量表示，因此这种表示方式也叫词嵌入。语义相近的单词在向量空间中距离也相近，这种词表示方法解决了 one-hot 编码的很多问题。Bengio 等人^[26]最早将深度学习的思想融入到语言模型中提出神经网络语言模型（Neural Network Language Model, NNLM）模型，模型的第一层映射矩阵就是学习到的词向量，Mikolov 等人^[27]受到这种思想的启发提出 Word2Vec。Word2Vec 提出两种模型 CBOW 和 Skip-gram 来学习单词的分布式表示，CBOW 使用中心词的上下文来预测这个单词而 Skip-gram 利用中心词来预测其周围的单词。但是无论是 CBOW 还是 Skip-gram 都只是考虑了单词局部上下文的信息，GloVe^[28]利用单词共现矩阵考虑了全局统计信息。

大量实验表明利用 Word2Vec 或者 GloVe 预训练好的词向量作为下游任务文本的词表征来初始化下游任务模型的第一层可以显著地提升模型的效果。除了词嵌入方法外，还有很多细粒度的嵌入方式。如 Seo

等人^[6]提出在词嵌入的基础上结合单词的字符嵌入，以缓解 NLP 领域常见的 OOV (out-of-vocabulary) 问题。Chen 等人^[29]提出引入单词的语义特征来增强嵌入表示，如段落单词与问题单词之间的完全匹配特征、词性特征以及单词的命名实体特征等。

然而 Word2Vec 和 GloVe 训练出来的词向量是静态的词向量，即训练好模型后一个单词的表示向量就是固定的，没有考虑上下文的信息，因此无法解决多义词问题。为了解决这个问题，Peters 等人^[30]提出一种动态的基于上下文的词嵌入模型 ELMo，每一个单词的词向量都是根据它所在的上下文语义表示的，很好的解决了一词多义的问题。ELMo 使用两层带有残差连接的双向 LSTM 来训练语言模型，前向和反向语言模型的目标函数如公式 (1) 和 (2) 所示。

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_1, t_2, \dots, t_{k-1}) \quad (1)$$

$$p(t_1, t_2, \dots, t_N) = \prod_{k=1}^N p(t_k | t_{k+1}, t_{k+2}, \dots, t_N) \quad (2)$$

最后的目标函数是最大化联合的前向和后向最大似然：

$$L(\Theta) = \sum_{k=1}^N (\log p(t_1, \dots, t_{k-1}; \Theta)) + \log p(t_{k+1}, \dots, t_N; \Theta) \quad (3)$$

在做阅读理解任务时，将文章和问题输入到模型中，每一层都会得到句子的语义表示，将每一层的特征加权求和作为下游模型的输入。

从早期的 one-hot 形式编码到分布式表示技术最后到基于上下文的词嵌入技术，每一种技术的出现都证明了一个好的文本表示方法可以极大地提升模型的性能。

3.1.2 编码层

这一层的目的是在词嵌入层的基础上通过对词嵌入层的输入文本做特征提取，进一步获得句子层面的语义信息。NLP 领域最为常用的特征提取器是基于循环神经网络 (RNNs) 的变体如 LSTM^[31] 和 GRU^[32] 等，因为这种循环结构适合处理文本这类序列数据，绝大部分的 MRC 模型编码层都是利用 RNNs 作为特征提取器。但是这种序列式的结构不能并行计算，训练耗时，更重要的是由于梯度消失所以不能解决单词之间长距离依赖问题，使得其特征提取能力始终受限。Vaswani 等人^[33]提出了一种用于机器翻译的 seq2seq 结构的模型 transformer，舍弃了常用的循环神经网络结构，完全的基于自注意力机制构建模型，实验表明 transformer 的特征提取能力强于循环神经网络而且可以并行计算加快训练。Yu 等人^[34]提出一种网络模型 QANet，不像之前的那些模型几乎都是用 RNNs 来做编码器，QANet 提出一种新颖的编码结构，利用卷积结合 transformer^[33] 中的多头注意力结构。整个结构的思想是先利用卷积建模局部特征的交互，再用自注意力机制建模全局交互，实验结果表明这种架构不仅加快训练速度同时在 SQuAD 数据集上模型性能优于那些利用 RNNs 作为编码器的模型。关于 transformer 的细节介绍见 3.2.1 节。

3.1.3 交互层

交互层是整个网络模型中关键的一层，目的就是让段落的语义信息与问题的语义信息融合，达到对段落更深层次的理解，而交互层中最常用的方法就是注意力机制 (Attention)。

注意力机制可以被视为是一个查询向量 (query) 和一组键值对向量 (key-value pairs) 的映射过程。整个过程首先是利用函数 f 衡量 query 和 key 之间的相似度，生成一个权重分数向量，然后将权重分数向量

归一化（通常利用 softmax 函数）后对 value 加权求和，得到的结果就是 query 对 key-value pairs 的注意力。具体计算公式形式如下：

$$\alpha_i = \text{softmax}(f(Q, K_i))$$

$$\text{Attention}(Q, K, V) = \sum_{i=1}^n \alpha_i V_i \quad (4)$$

其中 Q 表示 query 的向量表示， (K_i, V_i) 代表 key-value pairs 向量表示的第 i 个值，函数 f 常采用计算方式有内积、二次型函数、前馈神经网络、双维度转换函数，分别见如下公式：

$$f(p_i, Q) = p_i^T Q \quad \text{内积} \quad (5)$$

$$f(p_i, Q) = p_i^T W Q \quad \text{二次型函数} \quad (6)$$

$$f(p_i, Q) = v^T \tanh(W p_i + U Q) \quad \text{前馈神经网络} \quad (7)$$

$$f(p_i, Q) = p_i^T W^T U Q \quad \text{双维度转换函数} \quad (8)$$

在 NLP 领域中 $K = V$ ，即对于两个序列，其中一个序列为另一个序列的每一个位置生成一个权重值，这个值代表当前位置的单词对另一个序列的重要性。如果是自注意力（self attention），那么 $Q = K = V$ ，目的是计算序列中某个单词和其它单词之间的相关性从而增强自身的语义表示。Bahdanau 等人^[35] 最早将注意力机制应用在机器翻译领域，获得了极大的反响，为 NLP 领域的其它任务的模型提供了启发式的思想。

MRC 模型做注意力运算有两个方向：从问题到段落（Question-to-Context, Q2C），从段落到问题（Context-to-Question, C2Q）。Q2C 的注意力是指将问题看做是 Q ，段落看做 K, V 。定义 $C = [c_1, c_2, \dots, c_n] \in R^{n \times d}$ 代表段落的表示向量，其中 n 代表段落单词的个数， d 代表向量维度， $Q \in R^d$ 代表整个问题的表示向量 Q2C 的注意力计算步骤如下：

$$\alpha_i = \text{softmax}(f(c_i, Q))$$

$$\text{Attention}(C, Q) = \sum_{i=1}^n \alpha_i c_i \quad (9)$$

计算得到的 $\text{Attention}(C, Q)$ 也叫段落感知的问题表示（context-aware query representation）。C2Q 注意力的计算方式类似，段落看作是 Q ，问题看做 K, V ，计算得到的 $\text{Attention}(C, Q)$ 也叫问题感知的段落表示（query-aware context representation）。

公式（9）是将问题压缩成一个固定维度的向量，得到的注意力权重 α 也是一维的，因此也称为一维注意力。一维注意力方法所关注的是问题序列的整体对文章的注意力，没有考虑问题序列的不同单词之间对文章的关注程度差异。与其相对应的是二维注意力，即对于问题序列中的每一个单词都会和段落做注意力计算，得到的注意力权重是二维的。

此外还可以将注意力机制分为 one-hop 和 multi-hop 形式。其中 one-hop 也叫“单跳结构”，是指仅仅通过一次计算得到注意力权值然后加权求和得到注意力结果。与之对应的是 multi-hop，也叫“多跳结构”，one-hop 形式下仅仅只做一次交互计算，而对于复杂的问题通常是不能在一个句子中找出答案，需要在多个句子甚至多篇文章中多次推理，每一步推理过程中都会变换注意力关注的对象。

鉴于目前多数模型交互层所使用的注意力机制较为复杂，很难按照上述形式完全的区分开每一个模型，本文按照 Liu 等人^[36] 的思路按照注意力计算的方向等划分各个模型。

单向注意力 Hermann 等人^[2] 最早利用神经网络模型并且融入注意力机制做 MRC 任务。文中提出两种不同的单向注意力机制模型 Attentive Reader 和 Impatient Reader，均是计算问题到文章（Q2C）的注意力，且注意力的运算方式采用前馈神经网络（公式 4）。Chen 等人^[37] 在 Attentive Reader 的基础上利用双线性项（公式 3）取代原有的前馈神经网络计算方式。类似的，Kadlec 等人^[38] 利用内积运算作为注意力计算方式。

双向注意力 单向注意力所能交互的信息有限，而双向注意力则可以达到两个方向的互补，提供更加全面的交互信息。Xiong 等人^[5] 提出 Dynamic Co-attention Network (DCN) 模型，在交互层中采用协同注意力机制，协同注意力同步的计算 C2Q 以及 Q2C 两个方向的注意力，最后融合两个方向的注意力作为交互层的输出。Seo 等人^[6] 提出 (Bidirectional Attention Flow, BiDAF) 模型。同样计算两个方向（C2Q 和 Q2C）的注意力，但是与之前模型不同的是 BiDAF 将之前的段落语义表示和交互层计算得到的问题感知的段落语义表示一起流向后面的层，这样一定程度上避免了过早的对段落语义信息概括而导致信息的损失。模型的简化实验表明 C2Q 方向的注意力对模型的重要性大于 Q2C 方向的注意力，一种可能的原因是由于问题序列的长度小于段落文本的长度所以计算得到的段落感知的问题语义向量的信息不够充分。

单跳结构 单跳结构是指段落与问题之间的交互仅仅计算一次。要么是将问题整体压缩为一个向量与段落计算一次注意力，如 Attentive Reader^[2]，AS Reader^[38] 等，或者问题与段落的整体表示采用并行化的计算方式，如 DCN^[5]，BiDAF^[6] 等。

多跳结构 单跳结构不能实现多步推理的效果，多跳结构可以视为单跳结构的堆叠，目的是通过多次计算段落与问题的交互信息加深模型对段落和问题的理解，从而达到多步推理的目的。实现多步推理这种机制通常有以下几种方式：

第一种方式是基于之前时间步所计算得到的问题感知的段落表示，计算下一时间步的段落和问题交互，以一种序列式的方式计算注意力，如 Impatient Reader^[2]，这种方式类似于人在阅读过程中不断的在问题和文章之间做关注。

第二种方式是利用 RNNs 这种基于上一时刻隐藏状态更新下一时刻隐藏状态的循环特性来达到多步推理，如 Wang 等人^[4] 提出的 Match-LSTM 模型，序列式的阅读段落，利用 LSTM 存储每个时间步计算得到的问题感知的段落表示。具体的，计算段落当前时刻的单词与问题的注意力，得到的问题感知的段落表示与当前时刻段落单词的向量表示拼接作为 LSTM 的输入。类似的模型如 RNet^[7]，IA Reader^[39] 等

第三种方式通过堆叠多个计算注意力的层数达到多步推理的目的，如 Dhingra 等人^[40] 提出的 Gated Attention Reader (GA Reader) 模型，类似于 IA Reader 模型，同样采用 BiGRU 作为编码模块实现多跳结构。在每一步的推理过程中，首先通过 BiGRU 得到问题的向量表示，然后与段落的每一个单词做注意力的计算得到问题感知的段落表示作为下一层 BiGRU 的输入，这种处理过程类比于带着问题反复的阅读文章，每一次都加深对文章的语义理解。

之前的模型在交互层利用注意力机制融合段落和问题时都是利用句子的高层级别的语义信息而忽略了句子在低层次级别的语义信息，如单词级别的词嵌入等。Huang 等人^[41] 提出 FusionNet 模型，将每一个单词在第一层到后面所有层的向量表示拼接成一个向量，原文中称为单词历史 (history of word)，因为它包含了一个单词所有层的语义编码。为了解决维度问题同时不损失单词的历史信息，FusionNet 提出全关注注意力机制的概念：利用段落和问题的单词历史计算得到注意力权重，然后对问题的某一层表示向量加权求和。这种机制使得两个输入向量可以互相关注到对方的历史信息同时压缩维度，文中对注意力权重的计算方式如下：

$$\alpha_i = \text{ReLU}(Up_i)^T D \text{ReLU}(Uq_j) \quad (10)$$

其中 $p_i \in R^d$ 和 $q_j \in R^d$ 分别代表段落第 i 个单词和问题第 j 个单词的单词历史, U 和 D 是训练的参数。Hu 等人^[42] 认为在多层架构中, 当前层的注意力计算并没有直接考虑到之前层计算得到的注意力信息, 这可能导致两个不同但是相关的问题: (1) 多层注意力分布集中在相同的文本上导致注意力冗余; (2) 多层注意力未能集中在文本的重要部分造成注意力缺乏。针对这两个问题他们提出强化助记阅读器 (Reinforced Mnemonic Reader, RMR) 模型, 利用重关注机制, 通过直接利用之前层计算的注意力信息来微调当前层注意力分布的计算。

除了上面介绍的一些模型外, 还有一些基于上述模型中注意力机制改进的模型, 如 Wang 等人^[7] 提出的 RNet 模型, 在 Match-LSTM 的基础上添加门控机制, 将问题感知的段落表示与段落单词的向量表示拼接后通过一个门控单元。通过添加门控机制使得模型决定哪部分作为重要的语义信息输出。这种机制类似于人在阅读过程中要忽略段落中那些与问题无关的信息, 凸显出重要的信息才能更加准确的找到答案。此外 RNet 利用自注意机制对段落表示进行自注意力计算, 尽管利用基于注意力机制的循环神经网络可以存储中间过程计算得到的问题感知的段落表示, 但是两个距离较远的单词之间的交互信息由于梯度消失等原因会变得更弱。因此通过自注意机制可以使得段落中每一个单词关注到其余所有的单词, 使得模型对段落达到更深层次的理解。

Shen 等人^[43] 提出一种动态决定推理次数的模型 ReasoNet, 不同于 IA Reader, GA Reader 等模型, 在整个推理过程中有着固定的推理次数。这种固定推理次数的缺点就是不考虑问题的复杂性。ReasoNet 模型利用一个终止门产生二值值输出出来动态的决定是否继续推理。ReasoNet 模型大致分为外部记忆单元模块、内部控制器模块、终止门模块以及答案输出模块。将段落和问题通过 Bi-GRU 编码后的语义表示作为外部的记忆单元 M , 利用内部控制器 (采用 GRU) 当前时刻的状态与 M 做二维注意力匹配, 得到注意力结果输入到内部控制器中更新内部控制器的状态。终止门模块以当前时刻内部控制器的状态作为输入来判断是否需要继续推理。由于产生了二元离散输出值, 使得模型不能用梯度下降法训练, 因此模型引入强化学习机制训练。

表 7 对比了本节介绍的经典的 MRC 模型在注意力机制设计上的差异。其中 Q2C 代表问题到段落注意力, C2Q 代表段落到问题注意力, Bidirectional 代表双向注意力, self-attention 代表对段落做自注意力运算, one-dim 代表一维注意力, two-dim 代表二维注意力, one-hop 代表单跳结构, multi-hop 代表多跳结构。

3.1.4 答案预测层

这是整个模型架构的最后一层, 用来输出预测的答案。MRC 任务按照答案形式的不同大致分成四类, 因此这一层的设计需要考虑到答案形式。下面介绍各个模型在四类不同的 MRC 任务上输出层的设计。

1) 填空式: 这类任务答案的形式是预测问题中缺失的单词, 而且缺失的答案来源于文章中。Hermann 等人^[2] 最早提出将问题的表示与问题感知的段落表示拼接成一个向量然后映射到整个词典中预测那个缺失的单词。这种方法存在的一个问题就是不能够确保预测的单词一定是段落中的词汇, 这就使得模型的预测准确率受到影响。指针网络 (Pointer networks^[44]) 模型由 seq2seq 模型演变而来, 主要就是为了解决输出源自于输入的问题, 实现方式是利用计算的注意力权重分布直接输出预测结果, 而这种机制正适合填空式任务以及抽取式任务。Kadlec 等人^[38] 提出 AS Reader 模型正是受到指针网络的启发, 对于计算得到的注意力权重分布, 将其中相同单词的注意力权重相加, 最后输出具有最大权值的单词最为答案。填空式任务下, 模型的损失函数可以写为 $L(\theta) = -\frac{1}{N} \sum_{i=1}^N \log P_{y_i}$ 。其中 θ 为模型参数, N 代表样本数目, y_i 表示第 i 个样本中标准答案的位置。

2) 多项选择式: 这类任务是从多个候选答案选项中选择正确的选项。处理这种任务最简单的一种方式

表 6 基于注意力机制的模型对比

Table 7 Comparison of models based on attention mechanism

模型	注意力方向	注意力维度	推理模式
Attentive Reader ^[2]	Q2C	one-dim	one-hop
Impatient Reader ^[2]	Q2C	two-dim	multi-hop
Stanford Reader ^[37]	Q2C	one-dim	one-hop
AS Reader ^[38]	Q2C	one-dim	one-hop
IA Reader ^[39]	Q2C	one-dim	multi-hop
GA Reader ^[40]	C2Q	two-dim	multi-hop
Match-LSTM ^[4]	C2Q	two-dim	multi-hop
DCN ^[5]	Bidirectional	two-dim	one-hop
BiDAF ^[6]	Bidirectional	two-dim	one-hop
ReasoNet ^[43]	Bidirectional	two-dim	multi-hop
R-Net ^[7]	C2Q+self-attention	two-dim	multi-hop
RMR ^[42]	Q2C+self-attention	two-dim	multi-hop
QANet ^[34]	Bidirectional	two-dim	one-hop

就是计算模型输出后的段落表示和选项之间的相似程度，相似程度最高的作为预测的选项，从而将多项选择式任务变为句子之间的语义匹配问题。Wang 等人^[45] 提出将问题、段落、选项一起放在模型中做交互计算，交互层输出一个向量作为输出层的输入，输出层采用简单的输出维度是 1 的全连接层，输出的值代表模型对这个选项的打分值，其它的选项类似的处理，值最高的选项作为预测的答案。多项选择式任务下，模型的损失函数可以写为 $L(\theta) = -\frac{1}{N} \sum_{i=1}^N \sum_{j=1}^m \log P_{y_j^i}$ 。其中 θ 为模型参数， N 代表样本数目， m 代表选项个数， y_j^i 表示第 i 个样本中第 j 个选项是正确答案。

3) 抽取式：这类任务是从文章中提取出来一段连续的单词作为答案，虽然类似于填空式任务输出来源自输入的性质，但是不像填空式任务仅仅只是预测一个单词。因此填空式任务答案输出层的设计不能直接用来作为抽取式任务的答案预测层。由于提取文本的长度不固定，使得这一任务更具有挑战性。Wang 等人^[4] 同样受到指针网络的启发提出了两种基于指针网络的输出模型，第一种是序列式模型，利用指针网络以一种序列式的形式预测答案的每一个位置，处理过程类似于 seq2seq 模型的解码过程。第二种是边界式模型，不同于序列式模型那样序列的生成答案的每一个位置，由于要预测的答案是一段连续的文本，因此可以利用指针网络仅仅预测答案的起始位置和终止位置。所预测答案的概率是预测这两个位置概率的乘积，这种方式相比于第一种更加的简单而且实验结果表明更加高效。

边界式模型的这种设计思想也被后来很多 MRC 模型采纳。尽管边界式模型简单有效，但是边界式模型有可能陷入局部极值的情况从而提取错误的文本片段。为了处理这个问题，Xiong 等人^[5] 提出一种动态迭代的指针网络作为解码端，利用上一次预测的答案的起始位置和终止位置以及解码端当前的状态来重新评估下一次预测答案的起始位置和终止位置。多次迭代后选取所有迭代次数中概率最大的情形作为预测答

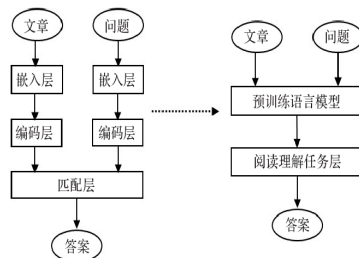


图1 机器阅读理解模型结构对比

案。抽取式模型的损失函数可以写为 $L(\theta) = -\frac{1}{N} \sum_{i=1}^N \log P_{y_i^s}^S + \log P_{y_i^e}^E$ 。其中 θ 为模型参数， N 代表样本数目， y_i^s 表示第 i 个样本中标准答案的起始位置在文章中的位置， y_i^e 表示第 i 个样本中标准答案的终止位置在文章中的位置。

4) 自由答案式：这类任务的答案形式已经不再是原文中某段文本，而是需要根据文章和问题生成符合语法规则的文本。这类任务对答案生成模块的能力要求较高。处理生成任务典型的架构是 seq2seq 模型，将段落看做是 encoder 端的输入，decoder 端根据词汇表中的单词生成答案。还有一类经典的处理生成任务的架构是 See 等人^[46] 提出的指针生成网络模型 (Pointer-Generator Network, PGNet)，最早用在文本摘要领域，模型结合了 seq2seq 的生成机制以及指针网络的拷贝机制，使得模型既能从词典中生成单词又能在原文中拷贝单词，实验结果表明该模型的效果优于传统的 seq2seq 模型。

表 8 对比了经典的 MRC 模型在 SQuAD^[3] 数据集上的表现⁶。

表 7 模型在 SQuAD^[3] 数据集上的对比 (acc 代表准确率)

模型	EM/F1
Match-LSTM ^[4]	64.7/73.7
DCN ^[5]	66.2/75.9
BiDAF ^[6]	68.0/77.3
ReasoNet ^[43]	70.6/79.4
R-Net ^[7]	72.3/80.7
RMR ^[42]	73.2/81.8
QANet ^[34]	76.2/84.6

3.2 基于预训练模型的 MRC 模型

预训练模型近年来 NLP 领域获得了极大的关注度，基于预训练模型的方法在 NLP 几乎所有的任务上都要优于之前的模型。预训练方式源自于迁移学习的概念：首先在其它相关任务上预训练模型，使得模型已经学习到一些知识，然后在目标任务上做进一步优化，实现模型所学知识的迁移。对于 NLP 领域来讲，预训练过程就是在大量的文本数据上学习到通用的语言表示。在应用到下游任务时，预训练所学习到的知

⁶统计数据源自 Yu 等人^[34]

识提供了一个很好的初始化点，从而加快模型的收敛并且提高模型的性能。此外预训练也对模型起到正则化的作用，使得模型避免在数据不充分的数据集上过拟合。

前面介绍的 ELMo^[30] 就是一个预训练模型，但是预训练的层次比较浅，对模型的提升效果也是有限的。目前流行的几个预训练模型如，GPT^[47]，BERT^[48] 以及基于 BERT 改进的预训练模型 RoBERTa^[49]，UNILM^[50]，ALBERT^[51] 等，从 MRC 模型结构的角度看这些预训练模型相当于将上述模型通用结构的编码层和交互层融合在一起，在编码的同时进行段落与问题的交互，这些预训练模型刷新了 MRC 领域多个任务的最佳性能。

3.2.1 Transformer

鉴于目前几乎所有的预训练模型都采用 transformer^[33] 结构或者其变体作为模型的特征提取器，因此本节首先介绍 transformer 结构。Transformer 是由 Vaswani 等人提出了一种用于机器翻译的序列到序列 (seq2seq) 结构。Encoder 端由六个相同的层堆叠而成，每一层有两个子层，第一个子层采用多头 (multi-head) 自注意力机制，第二个子层采用前馈神经网络 (Feed-Forward Network, FFN) 构成。之所以用自注意力机制是因为它既可以捕获句子中每一个单词的全局依赖关系而不受距离影响又可以并行计算。对比公式 (1)，自注意力机制下 $Q = K = V$ ，transformer 中采用的计算方式如下：

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (11)$$

其中 $\sqrt{d_k}$ 代表张量维度。此外 transformer 采用的是多头 (multi-head) 自注意力机制，将 Q, K, V 三个张量线性映射成多份，每一份之间做注意力的运算最后拼接。

$$\begin{aligned} \text{head}_i &= \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \\ \text{Multi-head}(Q, K, V) &= \text{concat}(\text{head}_1, \text{head}_2, \dots, \text{head}_h)W^o \end{aligned} \quad (12)$$

其中 h 代表头的数目，是一个超参数， W_i^Q, W_i^K, W_i^V, W^o 都是训练参数。

采用 multi-head 的目的是让模型联合关注序列中不同位置单词的不同表示子空间的信息，可以类比于卷积神经网络中利用多个卷积核做特征提取，目的同样是使得不同的卷积核关注的不同的特征。此外每一个子层都利用层正则化 (layer normalization^[52]) 和残差连接 (residual connection^[53]) 机制。Decoder 端与 Encoder 端类似，区别在于每一层额外添加了 encoder-decoder 注意力。

Transformer 的 encoder 端和 decoder 端都可以做特征提取器如 BERT^[48] 用 encoder 端特征提取，GPT^[47] 用 decoder 端特征提取，实验证明在大规模数据集上 transformer 的特征提取能力要强于基于 RNN 变体的编码器，目前几乎所有的 NLP 预训练模型都是利用 transformer 作为特征提取器。

3.2.2 预训练模型

OpenAI^[47] 提出一种生成式预训练模型 (Generative Pre-Trained, GPT)，使用多层的 transformer^[33] 的解码端作为特征提取器。模型采用两阶段的训练方式，先利用大规模无监督语料库训练前向语言模型，然后在下游任务的少量监督数据集上微调模型，因此 GPT 也是一种半监督学习方式，预训练阶段的目标函数就是标准的单向语言模型的目标函数，见公式 (1)。GPT 是 NLP 领域首先提出的一种基于微调 (fine-tune) 的通用式网络结构，不仅仅在 MRC 领域，在很多其它的 NLP 领域都取得了很大的进步。

GPT 属于自回归语言模型，自回归语言模型的缺点就是由于自回归的性质使得它不能同时利用一个单词的上下文信息预测这个单词 Devlin 等人 [48] 提出 BERT 预训练模型，与 GPT 最本质的不同在于预训练方式上采用的降噪自编码方式，随机掩盖掉一些单词，在输出层获得掩盖位置的概率分布，让模型根据

掩盖位置的上下文预测这个单词，这种机制也叫掩码语言模型（Masked Language Model, MLM）或双向语言模型。MLM 的目标函数为：

$$L(\Theta) = \sum_{i=1}^N \log P(t_i | t_1, t_2, \dots, t_{i-1}, t_{i+1}, \dots, t_N) \quad (13)$$

除了 MLM 任务外，还利用下一个句子预测（Next Sentence Prediction, NSP）任务使得模型在诸如文本蕴含、问答这类需要判断两个句子关系的下游任务表现更好。BERT 的预训练过程实质上是一个多任务学习的过程，通过 MLM 和 NSP 两个任务提高了预训练模型的语义表达能力，其中 MLM 任务用来学习句子中词与词之间的语义关联而 NSP 任务用来学习两个句子之间的逻辑关系。BERT 在 SQuAD^[3] 数据集上的效果超过了人类的水平，在其它的 NLP 任务上也都有提升。

BERT 开启了 NLP 领域预训练模型的时代，此后很多更加强大的预训练模型相继提出。这些预训练模型从结构上主要分为两大类：基于 BERT 的改进模型和 XLNet。基于 BERT 的改进模型主要是针对 BERT 预训练阶段的两个任务 MLM 和 NSP 做改进，如 Liu 等人^[49] 提出 RoBERTa 模型，使用动态掩码替换 BERT 的静态掩码同时去除 NSP 任务；Li 等人^[50] 提出 UNILM 模型扩展了 BERT 预训练的任务，由于双向语言模型的性质使得 BERT 在生成任务上效果不好，UNILM 同时训练单向语言模型（包括从左到右和从右到左）、双向语言模型以及 Seq2Seq 语言模型，使用掩码机制来解决不同的语言模型约束问题；Lan 等人^[51] 提出 ALBERT 模型，利用句子顺序预测（Sentence Order Prediction, SOP）任务改进了 BERT 的 NSP 任务；除了在预训练阶段改进训练任务外，Liu 等人^[54] 提出 MT-DNN 模型在微调阶段引入了多任务学习机制，使用多个任务来微调模型参数使得模型具有更好的泛化型。

BERT 模型以及基于 BERT 结构改进的模型在预训练阶段都采用降噪自编码的思想，所带来的问题就是预训练过程与微调过程不匹配，预训练使用的掩码机制在下游任务微调时不会被使用，导致两个过程存在数据分布的偏差。自回归语言模型虽然不存在这个问题，但是缺点是不能同时利用上下文信息。Yang 等人^[55] 提出的 XLNet 模型是一种可以同时获得上下文信息的自回归语言模型，通过使用排列语言模型，引入双向自注意力流和循环机制兼顾了自回归和自编码语言模型的优点。

表 11 从预训练任务、模型采用的特征提取器等详细对比了本文介绍的所有预训练模型。表 12 对比了几个预训练模型在两个常用的 MRC 数据集上的表现。

3.2.3 迁移预训练模型

预训练模型具有强大的文本表征能力，但是在应用到机器阅读理解任务上还需要根据具体的数据集特点设计不同的微调网络结构。以 BERT 系列的预训练模型为例，预训练阶段和微调阶段除了输出层外共享相同的体系结构，在做抽取式阅读理解任务时，输出层最简单的设计方式是仅利用两个向量与段落的向量表示计算答案在段落中起始位置和终止位置的概率。对于其它形式的阅读理解任务，往往根据任务的特点在输入层和输出层设计相应的结构，相关的工作可以参照^[56, 57, 58]。另一种迁移预训练模型的方式是基于特征形式的迁移，将具体任务的数据送进预训练模型，预训练模型的特征表示作为下游模型的输入，训练时固定预训练模型的参数。

虽然设计一个强大的预训练模型具有更好的泛化型和应用价值，但是预训练模型所消耗的计算资源是巨大的，因此如何利用预训练模型结合具体任务改进模型的结构是至关重要的。

表 8 预训练模型对比
Table 11 Comparison of pre-trained model

模型	任务	模型结构	介绍
ELMo ^[30]	前向 LM 反向 LM	LSTM	拼接两个单向语言模型的语义信息， 基于特征形式迁移
GPT ^[47]	前向 LM	Transformer	首次采用预训练+微调形式
BERT ^[48]	MLM+NSP	Transformer	利用掩码语言模型（MLM）和下一句预测（NSP）共同作为训练任务
MT-DNN ^[54]	MLM+NSP	Transformer	预训练过程与 BERT 一致，微调阶段采用多任务学习
XLNet ^[55]	PLM	Transformer-XL	采用排列语言模型（PLM），双向自注意力流 模型以自回归方式训练但是基于上下文预测
RoBERTa ^[49]	MLM	Transformer	采用动态掩码机制，去除 NSP 任务
UNILM ^[50]	前向 LM+ 反向 LM +MLM +Seq2SeqLM	Transformer	同时训练多种语言模型， 采用掩码机制解决不同语言模型的约束问题
ALBERT ^[51]	MLM+SOP	Transformer	对比 BERT 采用矩阵分解和共享参数减少模型的参数量， 同时用句子顺序预测任务（SOP）取代下一个句子预测任务（NSP）

表 9 预训练模型对比

模型	SQuAD 2.0 ^[18]	RACE ^[9]
	EM/F1	Acc
GPT _{v1} ^[47]	-	59.0
BERT _{large} ^[48]	80.0/83.1	72.0
XLNet ^[55]	86.4/89.1	81.8
RoBERTa ^[49]	86.8/89.8	83.2
ALBERT ^[51]	88.1/90.9	86.5

4 起什么名字？

以上介绍了典型的基于神经网络的机器阅读理解模型通用结构，这些模型所实验的数据集也大多是 SQuAD^[3] 等抽取式阅读理解数据集。Weissenborn 等人^[59] 专门设计了一个不包含交互层的网络模型 FastQA，仅仅在编码层的输入中额外添加两个特征：binary 和 weighted。其中 binary 特征表示段落中的单词是否出现在问题中，weighted 特征表示段落中的单词与问题的相似度。FastQA 仅仅添加了这两个额外的特征就在 SQuAD 数据集上达到很好的效果，优于很多之前介绍的基于复杂交互机制的模型。FastQA 的提出甚至质疑设计复杂交互机制的必要性，但同时也反映出 SQuAD 等抽取式数据集难度不高，简单的从一篇段落中抽取某一片段作为答案确实很难考察模型的阅读理解与推理能力，而且这种限制也不符合真实场景中的阅读理解。

基于以上的问题，为了能够真实的考察模型的阅读理解与推理能力，研究人员在原来的阅读理解任务基础上提出更加复杂的问题与任务，如要求模型判断问题是否可以根据给定的段落找出答案；要求模型在多篇段落中逐步推理寻找答案等。这些任务的提出是为了使得相应的阅读理解任务更加的贴近真实场景下

人们的阅读理解形式。我们将这些任务看作是阅读理解任务下新的研究问题。本章主要介绍三个问题：带有无答案问题的阅读理解，多段落阅读理解，对话型阅读理解。

4.1 带有无答案问题的阅读理解

早期的 MRC 数据集全都有一个共同的特点就是默认每一个问题都可以在给定的文本中找到答案，然而一段文本所包含的知识是有限的，因此有下述两点是需要考虑的：（1）这段文本不能回答那些与文本表达内容无关的问题；（2）某些问题可能与文本内容类似但是问题含义与文本含义不同。这两种问题都属于无答案问题，不能从给定的文章中找到问题的答案。目前流行的带有无答案问题的数据集如 SQuAD 2.0^[18]，在 SQuAD 的基础上增加了五万多个无答案问题。一个样例如表 9 所示，文章中 Bald Eagle Protection Act 指的是 1940 treaty 的名字，并不是问题所问的 1937 treaty 的名字。如果不考虑问题的可答性，仅仅依

表 10 SQuAD 2.0 的一个样例
Table 9 An example of SQuAD 2.0

文章：	Other legislation followed, including the Migratory Bird Conversation Act of 1929, a 1937 treaty prohibiting the hunting of right and gray whales, and the Bald Eagle Protection Act of 1940 . These later laws had a low cost to society—the species were relatively rare—and little opposition was raised.
问题：	What was the name of the 1937 treaty?
看似合理的答案：	Bald Eagle Protection Act

靠单词相似度匹配的模型很容易给出看似合理的答案。对于这类任务模型必须区分出哪些问题是无答案的，对于无答案的问题模型不能再给出“貌似合理”的答案。在 3.2 节所介绍的模型里，很多模型在 SQuAD 数据集上表现很好然而在 SQuAD 2.0 数据集上效果显著下降，这说明很多模型只是基于浅层的语义匹配来寻找答案而不是真正的理解了文章的含义。

对于带有无答案问题的阅读理解任务，模型要分为两个模块：（1）答案抽取模块；（2）判别无答案问题模块。答案抽取模块用来预测出答案在文章中的位置，第三章所介绍的 MRC 模型大部分都可以作为答案抽取模块，判别无答案问题模块用来判断当前问题是否可以回答。Clark 等人^[60] 尝试在原有的答案抽取模块的基础上额外添加一个专门用来预测无答案情况的网络层，此时损失函数定义如下：

$$L_{joint} = -\log\left(\frac{(1-\delta)e^z + \delta e^{\alpha_a \beta_b}}{e^z + \sum_{i=1}^{l_p} \sum_{j=1}^{l_p} e^{\alpha_i \beta_j}}\right) \quad (14)$$

其中 z 表示模型预测该问题是不可回答问题的分数，如果问题是可以回答的那么 $\delta = 1$ ，反之 $\delta = 0$ 。 α 和 β 分别表示输出层预测的文章中每一个单词作为答案起始位置和终止位置的概率， a 和 b 分别代表标准答案在文章中的起始位置和终止位置。

由公式 (17) 可以看出预测的答案跨度分数 α_a, β_b 和判断无答案问题的分数 z 是共同归一化的。Hu 等人^[61] 认为两个分数共同归一化会出现冲突，如果模型过分信任预测的答案跨度分数那么就会在预测的无答案问题时产生较低的分数。此外之前的模型并没有验证答案抽取模块预测的答案跨度的合理性。为了解决以上问题，他们提出 Read+Verify 架构。其中 Read 模块就是指答案抽取模块 + 判别无答案问题模块，Verify 模块用来进一步验证是否答案抽取模块预测的答案跨度所在的句子（原文中称为 answer sentence）

就是标准答案所在的句子。为了解决上面提到的冲突问题，在 Read 模块中额外增加了两个辅助损失函数：

$$L_{indep-span} = -\log\left(\frac{e^{\tilde{\alpha}_a \tilde{\beta}_b}}{\sum_{i=1}^{l_p} \sum_{j=1}^{l_p} \tilde{\alpha}_i \tilde{\beta}_j}\right) \quad (15)$$

$$L_{indep-unknown} = -(1 - \delta) \log \sigma(z) - \delta \log(1 - \delta(z)) \quad (16)$$

其中 $L_{indep-span}$ 代表答案抽取模块的损失函数，而此时的答案抽取模块是独立的预测答案片段而不考虑问题是否可以回答， $\tilde{\alpha}_a$ 和 $\tilde{\beta}_b$ 表示的就是答案抽取模块所预测出来的答案跨度。 $L_{indep-unknown}$ 代表判断问题无答案的损失函数，同样它是独立于答案抽取模块的。 σ 代表 sigmoid 函数。最后整个 Read 模块的损失函数定义为：

$$L_{Read} = L_{joint} + \gamma L_{indep-span} + \lambda L_{indep-unknown} \quad (17)$$

γ 和 λ 是两个超参数。实验表明去掉 $L_{indep-unknown}$ 后模型在判断无答案问题上的准确率显著下降，证明了上述提出的冲突确实存在。对于 Verify 模块，他们采用三种结构。第一种将预测出来的答案片段连同问题以及 answer sentence 连接成一个句子送入预训练模型 GPT^[47] 中预测无答案的概率。第二种采用交互式结构，通过注意力机制计算它们之间的关联。第三种结构是前两个结构的结合，具体的就是将前两个结构的输出张量拼接，实验证明这种混合结构使得模型效果更好。关于处理无答案问题阅读理解任务的其它相关模型可以参考^[62, 63]。

4.2 多段落阅读理解

目前大多数的研究热点集中于单段落阅读理解，仅需要从一段落上寻找答案，这对模型的阅读理解能力要求不高，此外真实场景中人们往往是从多篇段落中寻找与问题相关的答案最后互相比对得出最准确的答案。多段落阅读理解任务，一个问题 Q 会对应着多个相关的段落 (D_1, D_2, \dots, D_K) ，模型需要从这 K 个段落中寻找最佳答案 A ，建模概率

$$P(A|D_1, D_2, \dots, D_K, Q)$$

多段落阅读理解也可以认为是开放领域 (Open-domain) 问答的一种形式。Open-domain 问答目的是从广泛的领域资源 (如维基百科，网页搜索等) 寻找问题的答案而不仅仅在某段文本中，这更贴近于真实场景但同时具有相当大的难度。Chen 等人^[29] 提出利用检索 + 阅读 (Retrieve+Read) 的模式处理 open-domain 问答，先利用检索模块 (Document Retriever) 从维基百科中获取 5 个与问题最相关的段落，然后利用阅读器 (Document Reader) 预测出答案所在的位置。其中 Document retriever 采用基于 TF-IDF 权重的词袋向量模型，用来比较问题和文章的关联程度并且在此基础上用 bigram 哈希优化。

目前典型的多段落数据集如 MS MARCO^[8]、TriviaQA^[19]，中文的有 DuReader^[22]。以 MS MARCO 数据集为例，数据集样例见表 4。MS MARCO 由微软亚洲研究院发布，问题和文章来源于必应搜索，答案由人工生成，因此数据集接近真实应用场景而且答案不在局限于文章中。每个问题对应 10 个由必应搜索引擎返回的文本段落，其中包含答案依据用来回答问题的段落用 is_select=1 标记为 1。

多段落阅读理解的难点在于模型需要阅读多篇文章，更重要的是每篇文章都有可能包含与标准答案类似的句子，而有些是错误的答案，模型需要排除众多的干扰项选择最正确的答案。想要达到这个目的，要求模型能够解决跨段落实体消歧和跨段落答案验证等问题。

Tan 等人^[64] 提出 S-Net 模型，先通过片段抽取模块提取出一段文本作为答案的预测依据，然后利用生成模块生成答案。其中片段抽取模块采用多任务学习策略，除了预测文本片段之外还添加一个段落排名任务，将标记为 is_select=1 的段落视为正例。答案生成模块采用 seq2seq 模型，其中 encoder 端的输入是问

题和文章的向量表示, 同时将片段抽取模块的输出作为额外的特征拼接到文章中。实验证明 S-Net 在 MS MSCOCO 数据集上的效果要显著地优于 R-Net^[7], ReasoNet^[43] 这些单独做片段抽取任务的模型。

由于不同的段落都有可能包含与标准答案类似的句子, 但是有些答案并不是正确的。基于这个问题, Wang 等人^[65] 提出一种模型使得来自不同段落的候选答案在基于它们所在的上下文内容里互相验证对方的正确性。具体的就是将每一篇段落中预测出来的答案与其它段落预测的答案做交互验证。这样做的原因是经观察发现, 相比于错误的答案正确答案中的单词往往会在多个段落中重复出现, 因此通过交互验证可以凸显出正确答案。最后模型在 MS MARCO 数据集上的效果优于 S-Net。关于解决多段落阅读理解问题的相关工作可以参考^[66, 67, 68]。

4.3 对话型阅读理解

无论是单段落阅读理解还是多段落阅读理解任务, 它们都属于单轮对话问答, 即问答的形式只有一轮, 后面的问题与前面的问题和答案无关, 每一个问题都是互相独立的。而在真实场景中人们是通过多轮对话形式来交流的, 每一轮的问题和答案都是基于前面的问答情况。所以针对对话型阅读理解问题, 在回答当前轮的问题时不仅需要考虑文章还需要考虑前几轮的问题和答案。具体可以表示为: 给定 $Q_i, D, Q_{i-1}, \dots, Q_{i-k}$ 以及 A_{i-1}, \dots, A_{i-k} 要求模型给出 A_i 。其中 Q_i, A_i 表示第 i 轮的问题和答案, D 表示文章, Q_{i-1}, \dots, Q_{i-k} 和 A_{i-1}, \dots, A_{i-k} 分别表示前 k 轮的问题和答案。即建模概率:

$$P(A_i | D, Q_i, Q_{i-1}, \dots, Q_{i-k}, A_{i-1}, \dots, A_{i-k}) \quad (18)$$

目前典型的对话型阅读理解数据集有 CoQA^[11] 以及 QuAC^[69]。不同之处在于 CoQA 数据集的答案形式较为简单, 类似于 SQuAD^[3], 但是包含有 yes/no 以及无答案问题, 此外还有一定比例的问题是自由答案形式。对话型阅读理解数据集的一个样例见表 10。

其中每一个 Q_i 和 A_i 代表问题和对应的答案, 每一个 R_i 表示给出这个答案的依据, 用来训练模型。测试集中是没有答案依据的。从图中可以清楚地看到 Q_2 和 Q_3 中的 she 指代的是 Q_1 的答案 A_1 , 而 Q_4 的 How many? 以及 Q_5 的 Who? 所问的是 Q_3 中的 visitors。显然单独靠当前轮的问题是无法回答的, 对于这类任务模型必须解决指代消解问题以及如何利用对话历史信息。

Reddy 等人^[11] 首先提出将前几轮的问题与答案结合到段落中, 从而能够利用历史的对话信息回答当前轮的问题。Choi 等人^[69] 利用 BiDAF++^[60] 模型在 QuAC 数据集上进行实验, 为了利用历史的对话信息, 在文章中设置一个标记向量用来标记段落中的单词是否是历史答案, 将问题的轮次作为额外特征添加在问题向量上。

Huang 等人^[70] 认为上述的方法只是简单的添加之前轮的问题和答案, 而忽略了在回答之前轮问题时模型对整篇文章的推理过程状态。因此他们提出一种带有流机制的模型 FlowQA, 目的是将模型处理每一轮的问答过程下的对文章的语义理解状态流向下一轮的问答过程。FlowQA 模型整体上利用双向循环神经网络编码文章, 利用单向循环神经网络编码对话历史, 对比之前的模型, FlowQA 能够集成更加深层次的对话历史状态。

值得注意的是尽管 CoQA 数据集有部分答案是自由答案形式的, 但是上面的模型大多是利用片段抽取式的做法在 CoQA 数据集上实验, 主要原因在于生成式模型的效果往往不如抽取式模型的效果好, 因为生成式模型对答案生成模块要求较高。因此如何提高模型的答案生成能力是值得进一步研究的方向。由于预训练模型 UNILM^[50] 改进了 BERT 的训练任务, 增加了自回归语言模型以及 seq2seq 语言模型使得其在生成式任务上的效果很好, 在 CoQA 数据集上远远的超过于 Reddy 等^[11] 提出的生成式基准模型。关于如何迁移预训练模型在对话型阅读理解任务上的相关工作可以参考^[56, 71, 57]。

表 11 CoQA^[11] 数据集的一个样例
Table 10 An example of CoQA

文章	Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80. Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.
第一轮	<p>Q_1: Who had a birthday?</p> <p>A_1: Jessica</p> <p>R_1: Jessica went to sit in her rocking chair. Today was her birthday and she was turning 80.</p>
第二轮	<p>Q_2: How old would she be?</p> <p>A_2: 80</p> <p>R_2: she was turning 80.</p>
第三轮	<p>Q_3: Did she plan to have any visitors?</p> <p>A_3: Yes</p> <p>R_3: Her granddaughter Annie was coming over</p>
第四轮	<p>Q_4: How many?</p> <p>A_4: Three</p> <p>R_4: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.</p>
第五轮	<p>Q_5: Who?</p> <p>A_5: Annie, Melanie and Josh</p> <p>R_5: Her granddaughter Annie was coming over in the afternoon and Jessica was very excited to see her. Her daughter Melanie and Melanie's husband Josh were coming as well.</p>

5 讨论

本章主要讨论机器阅读理解模型的应用以及目前面临的主要问题和未来的发展方向。

5.1 机器阅读理解应用

智能客服 客服机器人是一种基于自然语言处理技术的模拟人类进行对话的服务程序，通过文字或语音与客户进行多轮交流。将用户的查询需求看做问题，通过阅读产品说明文档，利用机器阅读理解模型回答用户对产品的相关问题。

辅助决策 将机器阅读理解模型引入专业性较强的领域，如医疗病例报告、法律裁判文书等，可以帮助用户更好的决策。以医疗为例，Suster 等人^[72] 发布了关于医疗的阅读理解数据集 CliCR，包含着大量的病例报告，同时设计了包括确诊疾病、治疗用药等问题，这些与医生的日常工作息息相关。模型如果对数据集中的问题与文本有较深的理解就可以用于临床辅助医生诊断。

智能问答 搜索引擎的目标是根据用户的查询返回相关度较高的网页，机器阅读理解在搜索引擎中的一个重要应用就是智能问答。将网页内容视为文章，从网页文本中抽取或者生成最相关的答案，避免用户主动从网页中寻找答案。

5.2 面临的主要问题

缺乏推理能力 如前面所述，基于注意力机制的匹配模型大多是浅层的语义匹配模型，基于多跳结构的推理模式还过于单一，这些均没有形成深层次的阅读理解模型。Jia 等人^[73] 在 SQuAD 数据集的基础上设计对抗样本来测试 MRC 模型，在文章中加入人工设计的句子，这些句子与问题相似但是与答案无关以此用来误导模型。这些误导的句子对人来说没有影响然而实验证明几乎所有模型的准确率都显著下降，这表明模型很大程度上基于关键词匹配方式，离真正的阅读理解水平还很远。

答案生成技术不足 生成答案的技术还需要进一步提升，回顾目前机器阅读理解领域的数据集以及相应的模型，大多集中于片段抽取式问答且模型准确度很高，而对于自由答案型这种需要生成答案的模型效果很差，有些模型直接将生成式问题转为抽取式问题效果反而优于生成式的做法，主要原因在于生成答案模块的效果不够好。

缺乏可解释性 目前绝大多数的机器阅读理解模型基于深度学习，模型缺乏可解释性。而某些场景下的阅读理解任务可解释性就尤为重要，如上面提到的医疗、法律领域，模型做出判断后一定要给出医学解释或法律条文。

5.3 未来的研究方向

多模态融合 目前机器阅读理解主要集中于非结构化的文本领域，而还有许多其它结构、不同模态的数据如表格、视频、音频、图片等，相关的研究方向如数据库问答，视觉问答等。多模态阅读理解模型是未来的机器阅读理解发展方向之一。

引入外部知识 目前很少有机机器阅读理解模型融合外部知识，都是直接根据给定的文档回答相关的问题，而人在阅读一篇文章的时候对这篇文章的理解程度和他已经掌握的知识水平有很大关系。因此如果将外部知识源融入模型中，模型的性能大概率会显著提高。

结合其它 NLP 技术 MRC 任务涉及到 NLP 领域的许多关键任务，如文本生成、序列标注、句法分析等。此外 MRC 任务中的某些问题可以通过引入其它的技术解决，如通过指代消解技术可以更好的处理段落中代词所指代的实体，或利用实体识别技术区分出命名实体单词，这些都有助于提高答案的准确性。

6 总结

最后写。。