Computer Science and Application 计算机科学与应用, 2020, 10(12), 2457-2465
Published Online December 2020 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2020.1012261

Hans汉斯

机器阅读理解综述

殷　明

公安部第一研究所，北京
Email: 15911097538@163.com

收稿日期：2020年12月1日；录用日期：2020年12月24日；发布日期：2020年12月31日

摘　要

机器阅读理解是为了让机器能够真正理解人类语言，它是人工智能发展过程中不可或缺的步骤。由于自然语言的复杂性和多样性导致全面理解自然语言是智能化领域的难点问题之一。本文介绍了机器阅读理解相关的技术方法，主要分为基于规则的方法、基于机器学习的方法和基于深度学习的方法，并分类对机器阅读理解领域的相关代表性工作进行了详细的总结。随着深度学习在多个领域取得成果能够，本文重点介绍了基于深度学习的机器阅读理解方法。最后本文对机器阅读理解未来发展趋势进行展望。

关键词

机器阅读理解，深度学习，词向量，自然语言处理

A Review of Machine Reading Comprehension

Ming Yin

The First Research Institute of the Ministry of Public Security of PRC, Beijing
Email: 15911097538@163.com

Received: Dec. 1st, 2020; accepted: Dec. 24th, 2020; published: Dec. 31st, 2020

Abstract

Machine reading comprehension is to make the machine truly understand human language. Machine reading comprehension is an indispensable step in the development of artificial intelligence. Due to the complexity and diversity of the natural language, comprehensive understanding of the language is one of the difficult problems in the field of the intelligence. This paper introduces the related technologies and methods of machine reading comprehension, which are mainly divided into rule-based, machine learning-based and deep learning-based. We summarize the relevant

文章引用: 殷明. 机器阅读理解综述[J]. 计算机科学与应用, 2020, 10(12): 2457-2465.
DOI: 10.12677/csa.2020.1012261

representative work in detail. With the achievements of deep learning in many fields, this paper focuses on the machine reading comprehension method based on deep learning. Finally, the future development trend of machine reading comprehension is prospected.

Keywords

Machine Reading Comprehension, Deep Learning, Word Vector, Natural Language Processing

1. 引言

自然语言理解是人工智能的重要研究领域，是实现智能化过程中必须要解决的技术问题。而自然语言理解的最终目标是为了让机器能够理解人类语言，实现机器通过语言符号了解人类的需求，从而智能化地服务于人类。机器阅读理解就是测试机器全方位、大篇幅理解人类语言文本能力的任务。

传统自然语言理解较多的任务是面向单个句子内部的词性、语法、句法、情感以及语义等方面的单维度分析，然而人类实际的语言交流过程中是需要理解大范围、多语句、深层析的上下文信息才能顺利完成，因此需要让机器能够全方位、多上下文的理解人类的文字信息，实现机器阅读理解才能更好的实现智能化，对机器阅读理解的研究能为机器全方的位智能化提供技术基础。当前，机器阅读理解相关技术的实际应该也非常广泛。通过对搜索引擎用户上下查询的理解，可以帮助搜索引擎更好的理解用户当前查询的意图；通过对大规模文档资料(例如说明书、评测报告等)的阅读理解，减少人工筛选信息的工作量，提高智能客服的信息获取准确率；通过对公司累积地大规模文档，辅助企业自动建立行业知识库，更好的分析商业前景。

传统的自然语言理解任务由于多是面向单个和少量的句子，可以通过关键词、规则和简单的基于概率的统计模型很好的去解决，但是机器阅读理解是面向大篇幅的文本语言且要进行深层次的语义理解和推理，因此机器阅读理解是一项综合且复杂的任务，是人工智能领域长期以来的困难和挑战。机器阅读理解从上世纪 70 年代开始展开研究，先后经历基于规则的方法[1]、基于统计机器学习方法[2]和基于深度学习神经网络[3]三个阶段。

Lehnert [4]在 1977 年提出 QUALM 系统，QUALM 是机器阅读理解研究初期具有代表性的工作，该系统基于规则实现手动编码脚本实现问答过程。Hirschman [5]在 1999 年提出了一个包含 60 个故事的数据集，该数据集规定的任务是根据问题找到答案所在的句子。受限于当年的技术水平，并且机器阅读理解任务又十分复杂，机器阅读理解的研究一直停滞不前。

进入 21 世纪，统计学习方法兴起，机器阅读理解被转化为对段落、问题、答案三个维度进行统计机器学习编码解码的过程，其中最著名的工作就是 Richardson 等人[6]提出的 MCTest 数据集，基于该数据集研究人员提出了基于最大边缘学习的机器学习框架[7] [8] [9]。虽然机器学习方法一定程度能减少人工参与量，提高相关指标性能，但是相当有限。

近年来，深度学习在语音、图像和模式识别等领域取得了非常好的结果，在自然语言理解的各个子任务也逐步引入深度学习方法。在机器阅读理解任务上，为应对深度学习需要大规模的训练数据集，Hermann 等人[3]在 2015 年提出训练数据集 CNN/Daily Mail，该数据集包含约 126 万多条数据，基于

CNN/Daily Mail 进行训练预测，基于注意力机制的 LSTM 模型 THEATTENTIVE READER 被提出，该模型的取得的效果远超过往传统机器阅读理解方法。为了解决 CNN/Daily Mail 是完形填空形式的局限，Rajpurkar 等人[10]提出了科研领域非常著名的数据集 Stanford Question Answering Dataset (SQuAD)，该数据集是通过对 536 篇 Wikipedia 段落进行人工分析，从中抽取到了 107,785 个问题答案对。近两年来，相关研究学者基于 SquAD 提出了相关的神经阅读理解模型，机器阅读理解的效果在不断的被刷新。SQuAD 数据集的提出在机器阅读理解领域具有里程碑的意义，但是并不能代表着完成了机器阅读理解领域乃至人工智能的目标。为了更好地研究机器阅读理解，其他学者们也提出并构建了其他面向不同年龄、不同类型问题对象等方面的数据集，详见表 1。

Table 1. Machine reading comprehension data set summary
表 1. 机器阅读理解数据集汇总

问答类型	数据集	语言	问题数量	文章数量	问题来源	文章来源	网址
抽取型	TriviaQA [11]	英文	4 万	66 万	搜索日志	百科、网络文本	http://nlp.cs.washington.edu/triviaqa/
	NewsQA [12]	英文	10 万	1 万	众包	新闻	https://datasets.maluuba.com/NewsQA
	SQuAD [10]	英文	10 万	536	众包	百科	https://rajpurkar.github.io/SQuAD-explorer/
	SearchQA [13]	英文	14 万	6902 万	搜索日志	网络文本	https://github.com/nyu-dl/SearchQA
	DRCD [14]	中文	3 万	1 万	众包	百科	https://github.com/DRCKnowledgeTeam/DRCD
	CMRC 2018 [15]	中文	2 万	3507	众包	百科	https://hfl-rc.github.io/cmrc2018/
	CJRC [16]	中文	5 万	1 万	人工合成	裁判文书	http://cail.cipsc.org.cn/
多项选择	RACE [17]	英文	87 万	5 万	英语考试	英语考试	http://www.cs.cmu.edu/glai1/data/race/
	MCTest [6]	英文	2000	500	众包	虚假故事	http://research.microsoft.com/mct
完形填空	CNN/Daily Mail [3]	英文	140 万	30 万	人工合成	新闻	https://cs.nyu.edu/~kcho/DMQA/
	CBT [18]	英文	68 万	108	人工合成	儿童读物	http://fb.ai/babi/
	HLF-RC [19]	中文	10 万	2.8 万	人工合成	新闻、儿童故事	https://github.com/ymcui/Chinese-Cloze-RC
生成型	NarrativeQA [20]	英文	4.6 万	1500	众包	书籍、电影	http://deepmind.com/publications
	MSMARCO [21]	英文	10 万	20 万	搜索日志	网络文本	http://www.msmarco.org
	DuReader [22]	中文	20 万	100 万	搜索日志	网络文本	http://ai.baidu.com/broad/download?dataset=dureader
多跳推理	HotpotQA [23]	英文	11.3 万	-	众包	百科	https://HotpotQA.github.io

当前各种类型方法都没能达到实用的效果，各个数据集上的实验结果也不是很完美，与真实的人类水平还存在着差距，机器阅读理解的理论研究和落地使用仍然需要更多的探讨和拓展。本文接下来的部分将机器阅读理解各个类别的代表性研究成果进行详细介绍，并对未来进行总结展望。

2. 基于规则的方法

Lehnert 提出 QUALM 系统[4]，Lehnert 设计了一种关于问答系统的框架，该框架认为上下文对理解故事是非常重要的。QUALM 系统还强调实用性，利用策略模拟和脚本来构建系统理论框架的实践。因为 QUALM 系统构建的脚本和策略限定领域且规模小，所以很难推广普适的应用。20 世纪八九十年代，由于相关技术的落后以及机器阅读理解的复杂性，该领域一直缺乏研究，相关技术的研发进展缓慢。直到 1999 年，Hirschman 等人[5]提出了一个包含 60 个故事的数据集，该数据集规定的任务是根据问题找到答案所在的句子。该数据集的内容是来自小学生的阅读理解材料，每一个故事都包含人物、时间、地点以及发生的事件、原因等，同时针对每个故事有相关问题和回答。同时，Hirschman 等人提 DEEP READ 系统，该系统用词袋模型对问题和材料中的句子进行分析，并做信息抽取，通过模式匹配从材料中选择与问题匹配度最高的句子，以此作为相关问题的答案。2000 年，Hirschman 等人在 ANLP/NAACL 上组织举办了机器阅读理解比赛。参加比赛的队伍大部分都是基于词袋模型，使用规则进行匹配方法。Riloff 等人[24]中基于词汇和语义一致性人工定制规则，从而计算材料中语句和问题的匹配度。基于规则的系统大多限定领域可扩展性差，同时多数基于浅层的语义分析，查找正确答案的准确率很低，仅有 30%~40%。2010 年，Poon 等人[25]通过训练将海量文本信息进行结构化表示，构建了了一个基于规则的系统阅读理解系统，系统可以进行联合推断并且能够持续自主的学习。Berant 等人[26]提出了一种基于图匹配的方法。该方法通过借用类似于语义角色标注的方法，将文章段落、问题和答案分别转化成一个图结构，同过问题的图结构去文章段落中的图结构中做相似性的匹配进而找到答案，这种图结构的定义也需要人工的定义，因此比较难以推广。

众所周知，基于规则的方法多是制定规则然后进行模式匹配，不能动态扩展地处理句子的多样性带来的变化，又由于模式长度有限，基于窗口进行匹配，不能解决跨句子的依赖问题。

3. 基于机器学习的方法

21 世纪初期，研究者们将机器阅读理解转化为有监督学习问题来进行解决。机器阅读理解的内容被定义为{段落，问题，答案}三个维度，通过建立人类标记的训练数据集集，通过统计机器学习模型学习{段落，问题}与答案之间的内在联系，从而将答案映射到{段落，问题}上。这个时间出现的著名数据集是 MCTest [6]。MCTest 数据集是一个多项选择类型，包含 660 篇虚构故事，故事对应的问题是一道单选选择题，每个选择题有四个答案。基于 MCTest 数据集一系列机器学习模型先后被提出。Narasimhan 等人[7]在最大化模型概率的时候考虑句子之间的关系，通过设置隐变量来捕捉关联。Sachan 等人[8]提出一个隐变量结构叫做 answer-entailing，该结构代表着段落、问题和答案之间的关系。由于 answer-entailing 是隐含的，他们提出一个统一的最大边缘计算来学习隐变量。Wang 等人[9]综合句法、语义框架、共现关系和词向量等特征构建最大边缘计算框架进行机器阅读理解。

SQuAD 也是一个机器阅读理解领域经典的数据集，Rajpurkar 等人[10]在 2016 年发布，数据集包含 10 万多个包含段落、问题和答案的三元组，数据集的文章段落是维基百科中获取的，问题和答案是通过众包的方式构建。同时 Rajpurkar 等人[10]在考虑匹配、词频率、依存关系等特征的基础上，提出了一种基于 logistic 回归的基线方法。基线方法的实验结果准确率达到 40.4%和 F1 值达到 51.0%，该实验性能比基于简单的滑动窗口模式匹配模型要高一些。

相较于基于规则的方法，基于机器学习的方法能够取得一些性能上的提升，然而提升的效果并不显著，远不能达到使用的效果。基于机器学习的方法仍然存在一些问题。机器学习模型过度依赖已有的词法分析、句法分析和语义分析等工具，这些工具远不能达到较好的性能且泛华能力差，因此会造成错误层级传播；机器学习模型多是人工选择有限范围内的特征，很多远距离特征较少考虑，如果考虑也很难构建有效特征。

4. 基于深度学习的方法

2015 年机器阅读理解进入了基于深度学习模型的时代，Hermann 等人[3]提出了神经网络模型“The Attentive Reader”，此模型是有监督的基于 attention 机制的 LSTM 模型，实验中证明使用 CNN/Daily Mail 数据集，该数据集以新闻文章为段落，通过实体替换的方法形成完形填空的形式，实验结果表明准确率远超过往的自然语言处理方法，达到了 63.8%。Chen 等人[27]基于 CNN/Daily Mail 数据集提出一种巧妙的深度学习神经网络模型，并证实了深度学习神经网络模型能更好的识别词法匹配和释义，该阅读理解模型准确性提高到了 72.4%。

CNN/Daily Mail 包含有过多的噪声，为了解决噪声问题，Rajpurkar 等人[10]在 2016 年发布了 SQuAD 数据集。SQuAD 包含十万多问答对，数据规模大，数据质量较高，从而成为机器阅读理解模型主流的测试基准，研究人员基于该数据集提出很多阅读理解模型[28] [29] [30] [31]。随着模型的不断进行，基于深度神经网络学习模型的效果已经超过人类，为了进一步更好改善领域的发展，SQuAD 2.0 [32]被提出，此版本在已有基础上增加了 5 万多人类设计的无法回答问题，在 SQuAD 2.0 上进行阅读理解任务要识别到有些问题是文本段落中找不到答案的。随着词向量的预训练新模型的不断提出，在 SquAD 数据集基于 XLNet 预训练性能最好的单模型系统 F1 值已经达到了 95%以上[33]，已经超过人类的平均水平 91.2%。

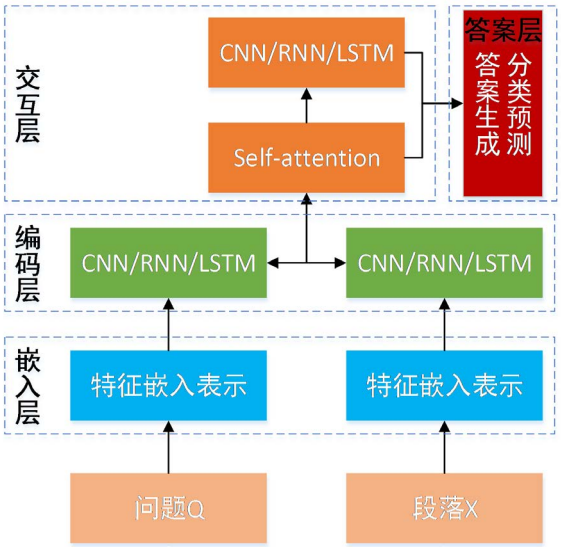

Figure 1. A general framework of reading comprehension methods based on deep learning
图 1. 通用基于深度学习的阅读理解方法框架

基于深度学习的阅读理解方法虽然各有自身特点，但是各个模型都是采用 4 层结构的框架，如图 1 所示。

1) 嵌入层：使用预训练的词向量将段落 X 和问题 Q 表示为多维向量作为模型的输入，还有一些模型方法将词性、命名实体等维度作为特征，有些模型还会引入 attention 机制融合各个维度的特征。

2) 编码层：使用神经网络模型将嵌入层输入的向量进行编码，进一步提取内部特征学习上下文的语义信息，最终通过 attention 机制输出经过感知的段落表示和问题表示。

3) 交互层：使用编码层输入的段落表示和问题表示对词语之间的相关性进行计算，融合获取新的表征，通过各种类型的 attention 机制形成最终的表示。

4) 答案层：根据数据集类型不同，进行最终答案的解码。

基于深度学习学习的阅读理解第一个关键步骤嵌入层，就是将段落和问题进行向量化，在深度学习之前经典的向量化模型是 One-Hot、TF-IDF 等，这些方法基本上不考虑单词间的关联，不能表达上下文的语义关联关系。为更好的表示文本的语义，字符集嵌入、静态不变向量转换成考虑上下文的动态向量、单词特征嵌入词向量等方法被相继提出。

在机器阅读理解领域，基于单词的分布式假设训练的词向量模型通常使用的 Word2Vec、GloVe 以及 Fasttext。Word2Vec 词向量[34]是在大规模的文本语料上利用神经网络模型训练获得的单词向量表示，是通过设置目标单词预测上下文或者上下文预测目标单词的目标进行训练。Word2Vec 词向量可以计算性的表示单词与单词指尖的相似性。为了弥补 Word2Vec 不能使用文本语料库的统计信息的问题，Pennington 等人[35]提出 GloVe 模型进行词向量的计算。GloVe 基于局部上下文串口和全局对数双线性回归模型，通过共现矩阵中的统计信息训练获得。为了解决 Word2Vec 和 GloVe 未考虑单词本身的结构和忽略单词形态学特征，Fasttext 词向量[36]模型在 Word2Vec 基础上考虑字词信息学习单词形态学特征。

为进一步学习上下文的关联性，研究人员从词的表征转移到上下文层面，也就是从单词映射组成句子的角度转换成句子到单词的映射，通过段落语句来学习单词的在语境下的语义表示，更好的学习同一个句子中的单词的上下文关联。上下文嵌入的经典模型有 CoVe、ELMO、BERT 以及 XLNet 等。

单词除了和语义相似性的单词存在关联，其和处在同一个句子里面的单词也应该是有关联的，单词的词向量要包含其上下文单词的信息，基于以上考虑 CoVe 词向量[37]被提出，CoVe 通过基于机器翻译编码解码的思想提出的 MT-LSTM 模型训练输出上下文向量。ELMO 上下文向量[38]不仅考虑语法、语义和上下文特征还考虑到多义词的现象，引入动态词向量理念，词向量训练生成采用耦合的双向 LSTM 语言模型获取双向语义信息，一个单词的向量是由整个输入句子的单词综合计算所得。ELMO 的双向模型不是同时进行，并不是真正的完全双向，而传统语言模型的原理导致完全双向的 Bi-LSTM 会存在自己预测自己的可能。BERT [14]使用双向 Transformer 编码模型[39]，采用遮蔽语言模型 MLM 和连续句子预测，BERT 是动态词向量，充分考虑字词句的关系表征，可以获取上下文相关的双向特征表示，增加了词向量的泛华能力。BERT 基于去噪自编码器模型进行预训练能够很好地建模并学习双向语境信息，但是 BERT 忽略了被 mask 位置依赖关系。因此，基于泛化的自回归预训练模型 XLNet 被提出[33]，XLNet 利用最大化所有可能的因式分解顺序的对数似然学习双向语境信息，使用用自回归模型克服 BERT 的缺点。同时，XLNet 为了更好的学习长句字依赖关系，引入双注意力流机制，将 Transformer 改为 Transformer-XL [40]。

编码层的作用是使用神经网络模型将嵌入层输入的向量进行编码，进一步提取内部特征学习上下文的语义信息。机器阅读理解的各种模型中最常用的是循环神经网络(RNN) [41]。循环神经网络充分考虑上下文的信息，上文信息会输入到下一轮的决策，它可以根据其记忆来处理任意时刻的输入序列，且支持输入序列长度可变。循环神经网络最大的问题是迭代过程中出现梯度消失从而导致不能收敛到最优值，相关 RNN 的变体被提出，例如长短期记忆网络(LSTM) [42]和门控循环单元(GRU) [43]。RNN 思想的编码层存在训练和预测速度慢的问题，针对此问题 Yu 等人[31]使用 CNN [44]来代替 RNN 模型，并用自 attention 机制[39]来补充 CNN 不能考虑全局的缺点，实验结果证明此方法能大大提高模型的处理速度。

交互层的作用是段落片段与问题之间单词权重交互，从而相互影响的融合编码的段落向量与问题向量。交互层通常采用 attention 机制，attention 机制首先分别计算段落中单词和问题中单词的相似度，然后进行权重归一化，最后加权求和获得相互影响的新的向量表示。编码层是进一步学习自己的内部结构，有了 attention 机制的思想，可以用 self-attention 机制进行阅读理解的编码，也就是自己和自己进行注意力机制学习，由此基于多头 self-attention 机制的提出 Transformer 模型，实验证明 Transformer 模型能很好的并优于 RNN、CNN 等的编码机制。Wang 等人[45]提出的 R-Net 模型，首次将自注意力机制运用于机器阅读理解任务。答案层作用是答案预测或生成，此层次需要根据数据集任务的目标不通来计算参数，本文就不再赘述。

基于深度学习的方法具有很好的灵活性，可以不依赖与任何的下游的任务的获取的语言特征，可以将所有的特征都统一到一个端到端的学习框架中。基于深度学习方法统一的学习框架可以减少传统机器学习人工构建特征体系耗费的人力物力，基于深度学习的方法还可以通过向量共享解决数据稀疏性的问题。

5. 总结与展望

机器阅读理解的目标是能像人一样在阅读一篇文章或一段话语片段之后，能多方位、多角度的理解里面包含和蕴含的知识。机器阅读理解领域先后发展了基于规则方法、基于机器学习方法和基于深度学些的方法。随着深度学习的快速发展，深度学习的神经网络在一些机器阅读理解数据集已经超越了人类识别的能力，机器阅读理解技术上已经取得了很大的进步并取得了一定的实际应用效果。

尽管机器阅读理解近五年来得到快速的发展，数据集的实验性能效果被不断刷新，但是机器阅读理解还是仅限于在数据集上进行优化，完全没有达到人类的阅读理解水平，目前机器阅读理解还仍处于探索阶段。未来机器阅读理解要在引入知识推理、提高模型的处理速度、增加模型的鲁棒性和泛化能力等方面进行研究和发展。

参考文献

[1] Schank, R.C. and Abelson, R.P. (1978) Scripts, Plans, Goals, and Understanding: An Inquiry into Human Knowledge Structures. *Language*, **54**, 779. https://doi.org/10.2307/412850

[2] Berant, J., *et al*. (2013) Semantic Parsing on Freebase from Question-Answer Pairs. *Proceedings of the* 2013 *Conference on EMNLP*, Washington DC, July 2013, 1533-1544.

[3] Hermann, K.M., *et al*. (2015) Teaching Machines to Read and Comprehend.

[4] Wg, L. (1977) The Process of Question and Answering. PhD Thesis, Yale University, New Haven.

[5] Hirschman, L., *et al*. (1999) Deep Read: A Reading Comprehension System. *Proceedings of the* 37*th Conference on ACL*, Maryland, June 1999, 325-332. https://doi.org/10.3115/1034678.1034731

[6] Richardson, *et al*. (2013) MCTest: A Challenge Dataset for the Open-Domain Machine Comprehension of Tex. In: *Proceedings of the* 2013 *Conference on Empirical Methods in Natural Language Processing*, Association for Computational Linguistics, Stroudsburg, 193-203.

[7] Narasimhan, K. and Barzilay, R. (2015) Machine Comprehension with Discourse Relations. *Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing*, Volume 1, 1253-1262. https://doi.org/10.3115/v1/P15-1121

[8] Sachan, M., *et al*. (2015) Learning Answer-Entailing Structures for Machine Comprehension. *Meeting of the Association for Computational Linguistics & the International Joint Conference on Natural Language Processing*, Volume 1, 239-249. https://doi.org/10.3115/v1/P15-1024

[9] Wang, H., *et al*. (2015) Machine Comprehension with Syntax, Frames, and Semantics. *Proceedings of the IJCNLP*, Beijing, July 2015, 700-706.

[10] Rajpurkar, P., *et al*. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. *Proceedings of the* 2016 *Conference on Empirical Methods in Natural Language Processing*, Austin, November 2016, 2383-2392. https://doi.org/10.18653/v1/D16-1264

[11] Joshi, M., *et al*. (2017) TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. *Proceedings of the* 55*th Conference on ACL*, Vancouver, July 2017, 1601-1611. https://doi.org/10.18653/v1/P17-1147

[12] Trischler, A., *et al*. (2017) NewsQA: A Machine Comprehension Dataset. *Proceedings of the* 2*nd Workshop on Representation Learning for NLP*, Vancouver, August 2017, 191-200. https://doi.org/10.18653/v1/W17-2623

[13] Dunn, M., *et al*. (2017) SearchQA: A New Q&A Dataset Augmented with Context from a Search Engine.

[14] Shao, C.C., *et al*. (2018) DRCD: A Chinese Machine Reading Comprehension Dataset.

[15] Cui, Y., *et al*. (2018) A Span-Extraction Dataset for Chinese Machine Reading Comprehension. *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*), Hong Kong, November 2019, 5883-5889. https://doi.org/10.18653/v1/D19-1600

[16] Duan, X., *et al*. (2019) CJRC: A Reliable Human-Annotated Benchmark DataSet for Chinese Judicial Reading Comprehension. In: *China National Conference on Chinese Computational Linguistics*, Springer, Cham, 439-451. https://doi.org/10.1007/978-3-030-32381-3_36

[17] Lai, G., *et al*. (2017) RACE: Large-Scale Reading Comprehension Dataset from Examinations. *Proceedings of the* 2017 *Conference on Empirical Methods in Natural Language Processing*, Copenhagen, September 2017, 785-794. https://doi.org/10.18653/v1/D17-1082

[18] Hill, F., *et al*. (2015) The Goldilocks Principle: Reading Children's Books with Explicit Memory Representations.

[19] Cui, Y., *et al*. (2016) Consensus Attention-Based Neural Networks for Chinese Reading Comprehension. *Proceedings of COLING* 2016, *the* 26*th International Conference on Computational Linguistics*: *Technical Papers*, Osaka, December 2016, 1777-1786.

[20] Kočiský, T., *et al*. (2017) The NarrativeQA Reading Comprehension Challenge. *Transactions of the Association for Computational Linguistics*, **6**, 317-328. https://doi.org/10.1162/tacl_a_00023

[21] Nguyen, T., *et al*. (2016) MS MARCO: A Human Generated Machine Reading Comprehension Dataset.

[22] He, W., *et al*. (2017) DuReader: a Chinese Machine Reading Comprehension Dataset from Real-world Applications. *Proceedings of the Workshop on Machine Reading for Question Answering*, Melbourne, July 2018, 37-46. https://doi.org/10.18653/v1/W18-2605

[23] Yang, Z., *et al*. (2018) HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. *Proceedings of the* 2018 *Conference on Empirical Methods in Natural Language Processing*, Brussels, October-November 2018, 2369-2380. https://doi.org/10.18653/v1/D18-1259

[24] Riloff, E. and Thelen, M. (2000) A Rule-Based Question Answering System for Reading Comprehension Tests. *Proceedings of the* 2000 *ANLP/NAACL Workshop on Reading Comprehension Tests as Evaluation for Computer-Based Language Understanding Systems*, Volume 6, 13-19. https://doi.org/10.3115/1117595.1117598

[25] Poon, H., *et al*. (2010) Machine Reading at the University of Washington. *NAACL HLT First International Workshop on Formalisms & Methodology for Learning by Reading*, Los Angeles, June 2010, 87-95.

[26] Berant, J., *et al*. (2014) Modeling Biological Processes for Reading Comprehension. In: *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Association for Computational Linguistics, Stroudsburg, 1499-1510. https://doi.org/10.3115/v1/D14-1159

[27] Chen, D., Bolton, J. and Manning, C.D. (2016) A Thorough Examination of the CNN/Daily Mail Reading Comprehension Task. *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 2358-2367. https://doi.org/10.18653/v1/P16-1223

[28] Wang, S. and Jiang, J. (2016) Machine Comprehension Using Match-LSTM and Answer Pointer.

[29] Chen, D., *et al*. (2017) Reading Wikipedia to Answer Open-Domain Questions. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 1870-1879. https://doi.org/10.18653/v1/P17-1171

[30] Wang, W., *et al*. (2017) Gated Self-Matching Networks for Reading Comprehension and Question Answering. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 189-198. https://doi.org/10.18653/v1/P17-1018

[31] Yu, A.W., *et al*. (2018) QANet: Combining Local Convolution with Global Self-Attention for Reading Comprehension.

[32] Rajpurkar, P., Jia, R. and Liang, P. (2018) Know What You Don't Know: Unanswerable Questions for SQuAD. *Proceedings of the* 56*th Annual Meeting of the Association for Computational Linguistics*, Volume 2, 784-789. https://doi.org/10.18653/v1/P18-2124

[33] Yang, Z., *et al*. (2019) XLNet: Generalized Autoregressive Pretraining for Language Understanding.

[34] Mikolov, T., *et al*. (2013) Efficient Estimation of Word Representations in Vector Space.

[35] Pennington, J., Socher, R. and Manning, C. (2014) Glove: Global Vectors for Word Representation. *Conference on Empirical Methods in Natural Language Processing*, Doha, October 2014, 1532-1543. https://doi.org/10.3115/v1/D14-1162

[36] Bojanowski, P., *et al*. (2017) Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics*, **5**, 135-146. https://doi.org/10.1162/tacl_a_00051

[37] Mccann, B., *et al*. (2017) Learned in Translation: Contextualized Word Vectors.

[38] Peters, M., *et al*. (2018) Deep Contextualized Word Representations. *Proceedings of the* 2018 *Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*, Volume 1, 2227-2237. https://doi.org/10.18653/v1/N18-1202

[39] Vaswani, A., *et al*. (2017) Attention Is All You Need.

[40] Dai, Z., *et al*. (2019) Transformer-XL: Attentive Language Models beyond a Fixed-Length Context. *Proceedings of the* 57*th Annual Meeting of the Association for Computational Linguistics*, Florence, July 2019, 2978-2988. https://doi.org/10.18653/v1/P19-1285

[41] Williams, R. and Zipser, D. (2014) A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation*, **1**, 270-280. https://doi.org/10.1162/neco.1989.1.2.270

[42] Hochreiter, S. and Schmidhuber, J. (1997) Long Short-Term Memory. *Neural Computation*, **9**, 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

[43] Cho, K., *et al*. (2014) Learning Phrase Representations Using RNN Encoder-Decoder for Statistical Machine Translation. *Proceedings of the* 2014 *Conference on Empirical Methods in Natural Language Processing* (*EMNLP*), Doha, October 2014, 1724-1734. https://doi.org/10.3115/v1/D14-1179

[44] Lecun, Y. and Bengio, Y. (1995) Convolutional Networks for Images, Speech, and Time-Series. In: Arbib, M.A., Ed., *Handbook of Brain Theory & Neural Networks*, MIT Press, Boston, 255-258.

[45] Wang, W., *et al*. (2017) Gated Self-Matching Networks for Reading Comprehension and Question Answering. *Proceedings of the* 55*th Annual Meeting of the Association for Computational Linguistics*, Volume 1, 189-198. https://doi.org/10.18653/v1/P17-1018