

文章编号: 1000-5641(2019)05-0036-17

面向自动问答的机器阅读理解综述

杨 康, 黄定江, 高 明

(华东师范大学 数据科学与工程学院, 上海 200062)

摘要: 人工智能正在深彻地变革各个行业。AI与教育的结合加速推动教育的结构性变革, 正在将传统教育转变为智适应教育。基于深度学习的自动问答系统不仅可帮助学生实时解答疑惑、获取知识, 还可以快速获取学生行为数据, 加速教育的个性化和智能化。机器阅读理解是自动问答系统的核心模块, 是理解学生问题, 理解文档内容, 快速获取知识的重要技术。在过去的几年里, 随着深度学习复兴以及大规模机器阅读数据集的公开, 各种各样的基于神经网络的机器阅读模型不断涌现。这篇综述主要讲述3方面的内容: 介绍机器阅读理解的定义与发展历程; 分析神经机器阅读模型之间的优点及不足; 总结机器阅读领域的公开数据集以及评价方法。

关键词: 人工智能; 智适应教育; 深度学习; 机器阅读理解

中图分类号: TP391 文献标志码: A DOI: 10.3969/j.issn.1000-5641.2019.05.003

A review of machine reading comprehension for automatic QA

YANG Kang, HANG Ding-jiang, GAO Ming

(School of Data Science and Engineering, East China Normal University,
Shanghai 200062, China)

Abstract: Artificial Intelligence (AI) is affecting every industry. Applying AI to education accelerates the structural reform of education and transforms traditional education into intelligent adaptive education. The automatic Question Answer system, based on deep learning, not only helps students to answer questions and acquire knowledge in real-time, but can also quickly gather student behavioral data and accelerate personalization of the educational process. Machine reading comprehension is the core module of an automatic Question Answer system, and it is an important technology to understand student problems, document content, and acquire knowledge quickly. With the revival of deep learning and the availability of large-scale reading comprehension datasets, a number of neural network-based machine reading models have been proposed over the past few years. The purpose of this review is three-fold: to introduce and review progress in machine reading comprehension; to compare and analyze the advantages and disadvantages between

收稿日期: 2019-07-29

基金项目: 国家自然科学基金(U1711262, 11501204)

第一作者: 杨 康, 男, 硕士研究生, 研究方向为基于机器阅读的自动问答技术。

E-mail: kyang1@163.com.

通信作者: 黄定江, 男, 研究员, 研究方向为机器学习与人工智能及其在计算金融等跨领域中大数据的解析和应用。E-mail: djhuang@dase.ecnu.edu.cn.

various neural machine reading models; and to summarize the relevant datasets and evaluation methods in the field of machine reading.

Keywords: Artificial Intelligence; intellectual adaptation education; deep learning; machine reading comprehension

0 引言

随着人工智能与大数据的飞速发展,深度学习技术正在深彻地变革着教育行业。自动问答系统能够理解学生问题,并从海量文档中快速获得答案,实时解答学生疑惑。机器阅读理解是自动问答系统的核心模块。它能理解用户问题,理解文档,精确返回给用户答案。同时在教育领域,可视化基于注意力的神经机器阅读理解模型可以辅助学生在文档中迅速定位与问题最相关的部分,帮助学生过滤无效文档,培养学生分析问题、信息抽取、快速聚焦重要信息的能力。

自然语言处理(Natural Language Processing, NLP)是实现智能、人机交互的重要基石,机器阅读理解则被视为自然语言处理领域皇冠上的明珠。早在20世纪70年代,学者们意识到机器阅读技术是测试计算机程序理解人类语言的关键方法,由于没有合适的文本表示方式以及传统机器学习模型有限的拟合能力,机器阅读理解发展缓慢。2015年后,随着大规模机器阅读数据集的出现以及神经网络的复兴,机器阅读领域快速发展。现存的机器阅读数据集按任务形式可划分为两大类:填空式机器阅读数据集,如CNN/Daily Mail, MCTest等;段落抽取式机器阅读数据集,如SQuAD, DuReader等。本文是一篇机器阅读理解综述,在对国内外机器阅读模型进行研究以及数据集调研之后,着重介绍填空式、段落抽取式机器阅读模型,从不同的角度阐述模型间的优点以及不足。

本文共4节,结构安排如下:第1节介绍机器阅读的定义与早期研究。第2节介绍近年提出的基于神经网络的机器阅读模型,比较各类模型的优点以及不足。第3节总结机器阅读数据集以及评估指标。第4节对机器阅读领域进行总结以及展望。

1 机器阅读理解的定义与发展历程

1.1 机器阅读理解的定义

教计算机智能程序理解人类语言是长期且有挑战的AI完全目标,但不禁会产生这样的疑问,什么样才算是真正的理解人类语言?或者说理解人类语言意味着什么?这里引用文献[1]对机器阅读理解的描述,通过表1所示的中国成语故事程门立雪来阐述机器阅读理解的定义。

在过去的几十年中,NLP社区尝试着从不同层面来理解人类语言,如下所示:

词性标注 要求计算机程序标注一句话中所有词的词性。成语故事程门立雪的第一句“在宋代,杨时喜欢研究学问”,“杨时”是专有名词,“学问”是普通名词,“研究”是动词,“在”属于介词。

命名实体识别 要求计算机程序识别出一句话中的人名、地名等实体,如“早期他在颍昌师从程颢,学到了不少知识”,“颍昌”是地名,“程颢”是人名。

句法解析 要求计算机程序给出句子中词与词之间的关系以及句子结构,如“杨时到洛阳请教另一位理学家程颐”,“杨时”是主语,“到”是谓语,“洛阳”是宾语等。

共指消解 要求计算机智能程序指出句子里人名与代词之间的共指关系。例如“宋代时，杨时喜欢研究学问。早期他在颍昌师从程颢，学到了不少知识”，第二句中的“他”指向杨时。

表 1 程门立雪成语故事

Tab. 1 A idiom story

在宋代，杨时喜欢研究学问。早期他在颍昌师从程颢，学到了不少知识。程颢死后，杨时到洛阳请教另一位理学家程颐（程颢的弟弟）。他到程颐家时，程颐在屋里睡觉。为了不打扰程颐，他就侍立在程颐家门口。程颐醒来后发现门外的雪已下了一尺多深。程门立雪由此而来。

根据以上材料回答以下问题：

1 杨时喜欢做什么？

研究学问。

2 早期杨时的老师是谁？

程颢。

3 谁侍立在程颐家门口？

杨时。

给定成语故事以及三个问题，回答第一个问题时，需要词性标注以及句法解析来分析“宋代时，杨时喜欢研究学问”。其中杨时是主语，喜欢是谓语，学问既是宾语又是名词。结合问题以“什么”结尾，可以轻易推测出答案。

回答第二个问题时，定位到“宋代时，杨时喜欢研究学问。早期他在颍昌师从程颢，学到了不少知识”。不仅需要句法解析、词性标注等技术，还需要共指消解指出句子中的“他”指代杨时，命名实体识别技术识别出“程颢”是人名，这与问题意图相符。

回答第三个问题时，需要运用上述 4 种自然语言处理技术来解决问题。

以上 4 种 NLP 任务对自然语言进行不同层面的分析，但是否存在一种任务形式能够综合评估以上 4 种 NLP 任务？文献 [1] 认为机器阅读理解是评估智能程序对语言不同层面理解程度的最佳方式。正如我们使用阅读理解的方法（给测试者一段话，让测试者回答反映文本不同层面的问题）去测试人类是否理解文章一样，对于计算机来说，机器阅读理解也扮演着相同的角色。为了回答问题，计算机首先要从不同层面理解文本，才能正确回答问题。由于问题是从文档的各个层面设计而来的，因此机器阅读理解是评估语言理解的最合适方法。

1.2 机器阅读理解的发展历程

早在 20 世纪 70 年代，Lehnert 等人^[2]认为机器阅读理解是测试计算机程序理解人类语言的重要方法。由于没有合适的文本表征方法，这一时期的机器阅读技术发展非常缓慢。随着硬件性能的提升、深度学习的复兴、大规模机器阅读数据集的出现使得神经机器阅读技术迅速发展，机器阅读理解再次引起人们的关注。本节介绍早期机器阅读模型，2010—2015 年的统计模型，以及目前发展迅速的基于神经网络的机器阅读模型。

1.2.1 早期系统

20 世纪 70 年代，研究人员意识到机器阅读理解是测试计算机程序理解人类语言的重要评估方法。在 1977 年，Lehnert 等人^[2]提出了 QUALM 系统。同时设计一套问答规则，主要用来指导研究人员解决实际问答问题。由于人类语言的复杂性，这些早期系统并不能达到理解人类语言的期望，在 1980 年至 1990 年期间，这种研究工作基本处于停滞状态。20 世纪 90 年代末，Hirschman 等人^[3]收集了 3—6 年级小学阅读材料，提出了第一个机器阅读数据集，该数据集要求机器阅读系统能够挑选出包含正确答案的句子。这一时期的工作包括 Hirschman 等人提出

的基于词袋、浅层语义的 Deep Reader System, Rilofd 等人^[4]提出的基于单词、语义以及规则的 QUARC 系统。由于缺乏合适的文本表征方法以及强大拟合能力的模型, 此阶段机器阅读任务的正确率只有 30%~40%, 模型性能远未达到人类期望。

1.2.2 基于机器学习的机器阅读理解

2013 年至 2015 年, 研究人员将机器阅读理解任务视为监督学习问题, 以(问题, 文档, 答案)三元组的形式收集数据集, 希望训练一个将(问题, 文档)映射成答案的统计模型。

Richardson 等人^[5]提出机器阅读数据集 MCTest 和 2 种统计模型。MCTest 收集 660 个虚构故事, 每个故事配有 4 个单选题, 每个单选题配有 4 个选项。同时 Richardson 等人提出两种统计机器阅读模型, 一种基于滑动窗口, 它衡量问题中词、答案与滑动窗口之间的加权距离, 使用距离作为特征来预测答案。另一种将问题与候选答案组成假设, 利用文本蕴含系统判断假设与文档之间的蕴含关系, 对候选答案进行排序。文献 [6-8] 提出一系列机器学习模型, 这些模型构造大量来源于句法解析、共指消解和词嵌入的文本特征, 构建在最大边界学习框架之上。这一时期基于机器学习的机器阅读理解模型在 MCTest 数据集上的准确率在 63%~70% 之间。

与早期基于规则的模型相比, 机器学习模型取得很大的进步。但模型性能提升十分有限, 且存在以下缺点:

(1) 机器学习领域, 缺乏合适的文本、词的表征方法。机器学习领域使用词袋模型、TFIDF 模型对文本进行表征, 但这种方法缺失序列之间的位置信息, 并且词汇表很大时, 会存在特征高度稀疏问题。

(2) 机器学习模型严重依赖基础自然语言处理技术, 如语义角色标注、句法解析等。这些 NLP 工具从特定领域数据训练得来, 处理不同领域数据时, 会存在泛化误差以及累积误差等问题, 导致模型性能下降。

(3) 现有机器阅读数据集虽有标签, 但规模太小。MCTest 数据集仅有 1 480 个训练样本, 不足以训练良好的统计机器学习模型。

1.2.3 基于深度学习的机器阅读理解

由于传统的机器学习模型不能很好地处理文本表征问题, 并且模型的复杂度不高, 对于机器阅读理解这种 AI 完全问题, 模型总是处于欠拟合状态。随着深度学习的再次兴起, 基于深度学习的机器阅读模型有效缓解以上 2 个问题, 促进了机器阅读理解的发展。

对于机器学习模型的文本表征问题, 文献 [9-11] 提出使用浅层神经网络训练词向量模型, 词向量模型将词嵌入到低维向量空间, 相同的词在低维向量空间中位置相近, 词表征问题得到有效缓解。在词的上下文表征中, 相同的词在不同上下文中应该有不同的表征, 即需要一个编码器, 这个编码器能够存储并编码词的上下文信息。文献 [12-13] 提出使用循环神经网络 (Recurrent Neural Network, RNN) 作为编码器。词向量模型以及编码器的引入使得文本表征问题得到有效处理。

面对复杂阅读理解任务时, 传统机器学习方法将复杂任务分为多个子任务, 用独立的机器学习模型处理子任务。由于这些模型没有 100% 的准确率, 当多个模型一起运作时, 难免会产生累积误差, 导致系统性能下降。在深度学习时代, 端到端的神经网络架构使得模型各个模块进行自适应学习, 将文本表征、文本理解、文本推理进行有效结合, 避免累积误差的产生。

2015 年, DeepMind 研究员 Hermann 首次提出大规模有监督填空式机器阅读理解数据集 CNN/Daily Mail。同时, 他们开发的基于注意力的神经网络阅读器在性能上碾压传统模型。神经机器阅读模型具有强大的文本表征能力以及推理能力, 促进机器阅读领域的进步。此后各式各样的模型被提出, 如 Standford Attention Reader, Attention Sum Reader 等。

数据集是神经网络模型的基石, 大规模机器阅读数据集的出现促进神经机器阅读模型的发

展。代表性的工作有斯坦福大学的 SQuAD 数据集^[14], 微软研究院的 MSMARCO 数据集^[15], 华盛顿大学的 TRIVIAQA^[16], NewsQA^[17], NarrativeQA^[18], RACE^[19], SearchQA^[20], DuReader^[21]等。根据数据集的复杂程度不同, 不同特征抽取能力以及推理能力的模型也被开发出来。在第 2 节, 我们将详细介绍基于神经网络的机器阅读模型。第 3 节我们分析各种数据集的特点、数据集评估指标以及模型性能对比。

2 模型介绍

大规模数据集的出现、硬件性能的提升以及联结主义的复兴促进神经机器阅读的发展。由于不同类型的阅读理解问题考察人类对文章不同层面的理解, 因此, 机器阅读数据集也可以分为 2 类: 填空式数据集, 从众多候选答案中挑选出正确答案; 段落抽取式数据集, 从文档中抽取小部分子集作为答案。虽然都采用答案抽取的方式来解决这两种任务, 但这两种任务在难度上是递增的, 同时填空式问题考察模型对于局部上下文的理解, 从局部上下文来预测缺失词。段落抽取式任务要求模型根据问题把控全文, 理解全文不同部分之间的语义关系。本节安排如下: 2.1 节定义问题形式; 2.2 节详细介绍基于注意力的模型; 2.3 节介绍基于多轮推理的模型; 2.4 节介绍其他模型; 2.5 节总结各类模型的优点以及不足。

2.1 问题定义

虽然机器阅读任务形式众多, 但基本上都可以理解为 3 元组 (问题 q , 文档 d , 答案 a) 建模的形式, 即给定问题 q 和文档 d , 模型对答案的条件概率

$$P(a|d, q) \quad (a \in C) \quad (1)$$

进行建模, 其中 C 为候选答案集。不同的区别在于填空式机器阅读的答案是一个实体或者动词, 存在一个候选答案集, 模型需要从候选答案集中挑选出正确的候选答案。段落抽取式任务的答案是文档 d 的子集, 需要从文档中预测答案的起始位置与结束位置。Taylor 等人^[22]首次提出填空式的阅读评测任务 (Cloze style Questions), 从一个句子中去掉一个实体词形成问题。让机器阅读系统阅读文档与问题, 预测空缺词, 从而评测系统的阅读能力。CNN/Daily Mail 数据集是填空式机器阅读任务中具有代表性的数据集。表 2 显示 CNN/Daily Mail 的一个样本, 模型需要将问题中的 **X** 替换成人名 **Oisin Tymon**。Rajpurkar 等人提出的 SQuAD 是段落抽取式任务中具有代表性的数据集。表 3 显示 SQuAD 数据集中的一一个样本, 可以看出不同问题所要抽取的答案长度也不尽相同。

表 2 Daily Mail 数据集中的一个样本

Tab. 2 An example of Daily Mail dataset

Context

The BBC producer allegedly struck by Jeremy Clarkson will not press charges against the “Top Gear” host, his lawyer said Friday. Clarkson, who hosted one of the most-watched television shows in the world, was dropped by the BBC Wednesday after an internal investigation by the British broad-caster found he had subjected producer **Oisin Tymon** “to an unprovoked physical and verbal attack.”...

Query

Producer **X** will not press charges against Jeremy Clarkson, his lawyer says.

Answer

Oisin Tymon

表 3 SQuAD 数据集中的样本

Tab. 3 An example of SQuAD dataset

In 1870, Tesla moved to Karlovac, **to attend school at the Higher Real Gymnasium**, where he was profoundly influenced by a math teacher **Martin Sekulic**. 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.

1. In what language were the classes given? German
2. Who was Tesla's main influence in Karlovac? Martin Sekulic
3. Why did Tesla go to Karlovac? attend school at the Higher Real Gymnasium

2.2 基于注意力的模型

受到人视觉系统能从宽阔的视野中重点关注感兴趣区域的启发, 计算机视觉领域最早提出注意力机制。直到 Bahdanau 等人^[23]将注意力机制引入神经机器翻译, 显著提升了模型的翻译性能, 注意力机制才在 NLP 领域被广泛使用。Hermann 等人^[24]将注意力机制引入机器阅读理解模型中, 显著提升了模型的准确率。

本节将详细介绍注意力机制在机器阅读领域的应用。基于注意力机制将模型分为两类: 一维匹配模型与二维匹配模型。

2.2.1 一维匹配模型

多数一维匹配模型(一维匹配的含义在下文说明)是为解决填空式机器阅读任务。文献[24-26]分别提出 Attentive Reader, Standford Attentive Reader 和 Attention Sum Reader 三种一维匹配模型。

Hermann 等人^[24]开发了 CNN/Daily Mail 数据集并提出 Attentive Reader。模型架构如图 1 所示, 使用 2 个不同的双向 LSTM 编码器对文档 d 和问题 q 进行语义编码。但只将问题 q 最后时刻的前向与反向隐藏状态进行拼接, 形成问题表示 u 。对于文档 d 中的每个词, 计算其与 u 的注意力权重, 并进行加权和形成文档表示 r , 通过非线性变换 g , 融合问题 u 与文档 r 的语义特征, 最后使用 softmax 进行答案预测。问题与文档的注意力权重可理解为模型从问题出发, 对文档中词的关注程度。由于注意力权重是一维向量, 所以称之为一维匹配模型。

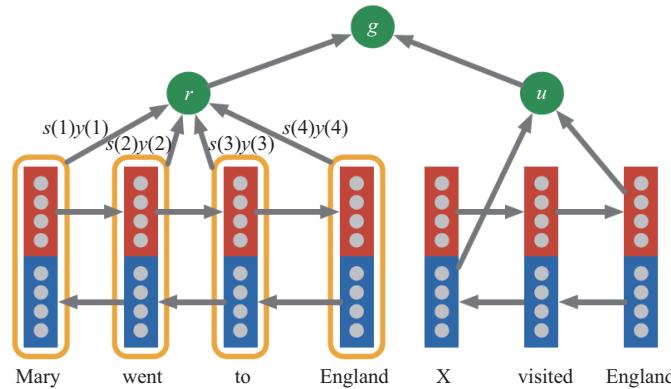


图 1 Attentive Reader 架构图
Fig. 1 Architecture of Attentive Reader

后续的许多工作都在 Attentive Reader 基础上进行改进。Chen 等人^[25]在 Attentive Reader 的注意力模块上进行两处优化, 提出 Standford Attentive Reader。在注意力计算上,

模型采用双线性函数而非 Attentive Reader 使用的 Tanh 函数计算注意力权重, 双线性函数的有效性已在机器翻译等领域得到验证, Chen 等人^[25]对 CNN/Daily Mail 数据集进行抽样分析, 认为该数据集不需要复杂的推理能力, 在得到文档的注意力加权和表示之后, 没有进行非线性变换, 直接使用 softmax 进行答案预测, 简化模型。最终 Standford Attentive Reader 的性能超过 Attentive Reader, 间接证明 Chen 等人^[25]抽样分析的结论。

Kadlec^[26]提出 Attention Sum Reader (AS Reader) 模型, 在文献 [25] 的基础上精简答案预测层。对问题 q 与文档 d 编码, 计算问题与文档中每一个词的注意力权重。受 Pointer Network^[27]启发, 直接使用归一化后的注意力权重作为答案的概率, 对相同词的概率进行累加。虽然 AS Reader 答案预测模块太过简单, 但只要对问题、文档进行充分的特征抽取, 也能得到不错的效果。这种精简的答案预测策略对后续二维匹配模型有着深刻的影响。

一维注意力模型主要针对填空式机器阅读任务。模型采用一层循环神经网络以及一维注意力机制就能很好地对问题、文档进行表示。但一维的问题表示向量会损失部分语义信息, 增加模型的推理难度, 降低模型答案抽取准确率。与一维匹配模型相比, 二维匹配模型使用二维注意力机制充分抽取问题、文档的文本特征以及捕捉它们之间的语义交互关系。

2.2.2 二维匹配模型

一维匹配模型将问题压缩成一个稠密向量, 可能会导致问题语义特征的损失。这种损失对于简单的机器阅读任务来说, 影响不大。对于需要复杂推理的机器阅读任务, 问题语义特征的损失会增加模型推理压力, 进而影响模型性能。本节介绍二维匹配模型, 它能充分挖掘问题语义信息, 增强文档、问题间的语义交互。

文献 [24] 中的 Impatient Reader 是首个二维匹配模型, Attentive Reader 采用一维匹配模式, 即问题 q 与文档 d 之间的注意力权重是一维向量, 相当于从问题 q 出发, 进行一次文档阅读。但人在做阅读任务时, 往往从问题出发, 阅读文档, 回到问题, 再次阅读文档等。受此过程的启发, Impatient Reader 阅读问题 q 中每个词时, 计算该词与文档中所有词的注意力, 并做加权和来聚合文档信息。图 2 显示 Impatient Reader 的架构图, 使用不同时刻问题表示 u 来不断提取文档特征 r 。这种注意力机制使得模型编码问题时, 都能从问题 q 的每一个词出发, 不断聚合文档信息。由于计算问题中每个词与文档中所有词的注意力权重, 最终将得到二维注意力权重向量, 因此称此类模型为二维匹配模型。

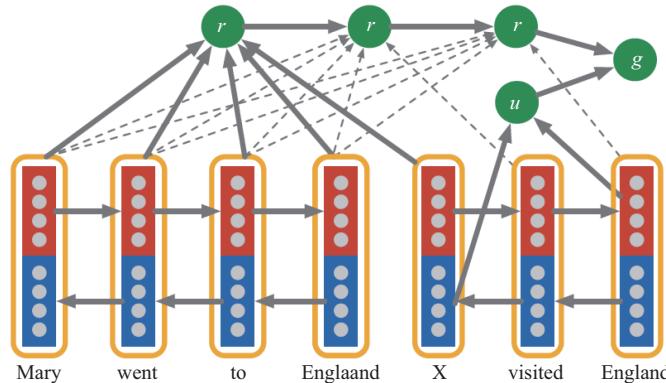


图 2 Impatient Reader 架构图
Fig. 2 Architecture of Impatient Reader

受 Impatient Reader 的启发, 文献 [28-31] 分别从不同方面挖掘文本特征, 进行答案预测。

Cui 等人^[28]提出 Consensus Attention Sum Reader (CSA Reader), 该模型在答案预测阶段

与 AS Reader 都使用 Pointer Network 网络。CAS Reader 针对问题中的每一个词, 计算与文档中所有词的相似度(又称文档端注意力), 分别使用 Sum, Average, Max 三种启发模式将问题中每一个词对应的文档端注意力融合成一维注意力, 最后采用 Pointer 预测最终答案。同时文献[28]又提出中文填空式机器阅读数据集 Children's Fairy Tale(CFT), 促进中文机器阅读的发展。

文献[29]提出 Attention of Attention Reader (AoA Reader)。AoA Reader 进一步改进 CAS Reader 的注意力融合模块。AoA Reader 基于文档中的每个词获得问题端的注意力, 再根据问题中每个词获取文档端注意力。对文档端的注意力权重进行融合时, 文献[29]认为 CAS Reader 启发式的融合方案不具备解释性, 且未充分利用问题端的注意力信息。因此 AoA Reader 对所有问题端注意力取平均, 得到和问题长度相同的一维向量, 向量的每一个元素对应问题中的一个词, 该元素值可理解为所对应的词对于整个文档的匹配程度, 最终将一维向量对文档端的注意力进行加权和, 采用 Pointer Network 挑选出多个候选答案。AoA Reader 没有将概率最高的作为预测答案, 而是挑出概率最高的前 k 个作为候选答案并对其进行排序。模型将候选答案填入问题的空缺处, 形成候选句子。为验证每个候选句子的合理性, AoA Reader 分别采用 Global N-gram LM, Local N-gram LM, Word-class LM 三个不同的验证器来对候选句子进行打分排序, 挑出最优候选句子所对应的答案作为最终预测答案。

上述的二维注意力模型主要应用于填空式任务, 下文所介绍的机器阅读模型既可应用于填空式任务, 也可应用于段落抽取式任务。

Wang 等人^[30]提出 Match-LSTM 模型, 使用 2 个不同的 RNN 对文档和问题进行编码。对于文档中的每个词, 计算问题端注意力权重。Match-LSTM 模型没有采用双线性函数或者 Tanh 函数来计算相似度, 正如论文标题中的 Match-LSTM, 将 LSTM 模块与 Tanh 函数结合来计算问题端注意力权重, 再通过加权和对问题信息进行聚合。最后在答案预测模块, 引入 Boundary Pointer Network^[27]网络进行答案预测。虽然引入 LSTM 参与注意力计算能够更好地建模问题与文档之间的关系, 但 LSTM 串行的特点使得 Match-LSTM 模型的训练与推断时间都比较长, 不适用于低延时的问答场景。

Seo 等人^[31]提出 Bi-Directional Attention Flow (BiDAF) 模型。模型在词的表示上使用 glove 词向量, 同时引入字符嵌入^[32], 有效缓解未登录词的表征问题。文献[28-30]提出的模型, 或计算问题端注意力权重, 聚合问题信息; 或计算文档端注意力权重, 聚合文档信息。而 BiDAF 使用双端注意力权重, 同时聚合文档和问题信息, 模拟文档和问题交互, 极大地降低文档和问题的信息损失。BiDAF 没有采用 Pointer Network 来预测答案, 而采用全连接层分别预测答案的起始位置与结束为止。在训练时间与测试时间上, BiDAF 明显优于 Match-LSTM。

文献[33]提出 R-Net 模型, R-Net 使用问题端注意力来聚合问题信息, 并将门控机制和自注意力机制引入机器阅读中。这种门控机制的设计思路来源于人类阅读过程。人类做阅读理解时, 过滤掉与问题无关的信息, 记忆与问题最相关的文档信息。当获得融合问题信息的文档表示时, 文献[33]认为获取正确答案不仅需要文档的局部上下文信息, 同时也应关注其他部分的文档信息, 因此 R-Net 模型引入自注意力机制, 将每个词的感受域扩展到全文档, 增强模型对文档的整体理解能力。

上述的二维匹配模型核心模块是模拟问题和文档的语义交互, 但这些模型重点关注深层语义特征的学习, 而忽略低级语法、句法特征的重要性。Huang 等人^[34]提出 FusionNet 模型, 充分融合问题和文档不同层次的信息, 如低层次的命名实体识别、词性信息和高层次的语义信息, 来对问题、文档语义关系建模, 提升模型的文本特征抽取能力。文献[35]提出 Stochastic Answer Network (SAN) 模型, 该模型的创新点在答案预测层。模型在推断的过程中会产生多个候选答

案, SAN 会随机丢弃某些答案, 对剩下的候选答案取平均来进行答案预测。实验表明, SAN 随机丢弃候选答案的特性没有降低模型性能, 显著增强了模型的鲁棒性。

由于填空式机器阅读任务不需要复杂的推理能力, 所以一维匹配模型只使用单层的循环神经网络以及一维匹配机制。简单的一维模型会损失问题的部分语义信息, 稍稍降低模型答案预测的准确率。文献[30-35]中的二维匹配模型主要应用于需要复杂推理的抽取式机器阅读任务, 模型多次使用循环神经网络以及复杂的注意力机制来模拟人类的推理过程, 充分抽取问题、文档特征, 进而提升模型的答案抽取能力。

2.3 推理模型

在机器阅读领域, 具有推理过程的模型称为推理模型。模型中的推理过程由堆叠同一模块或相似模块(大多由循环神经网络组成), 增加网络深度来实现。在推理过程中, 问题与文档信息不断的更新、交互, 进而更好地对问题、文档之间的语义关联以及表示进行建模。文献[36]提出记忆网络, 文献[37]在记忆网络的基础上引入 multiple hop(堆叠)策略来实现多轮推理。文献[38-40]将 multiple hop 策略引入各自的模型, 增强模型推理能力, 提升答案抽取准确率。

大部分基于注意力机制的模型需要 RNN 对问题和文档进行编码。RNN 通过稠密的隐藏状态来保存文本序列的历史语义信息, 但稠密向量会损失部分语义。针对此问题, 文献[36]提出记忆网络架构, 通过外部记忆模块来缓解 RNN 隐藏状态信息丢失的问题, 但 Weston 等人^[36]只提供一种模型框架, 没有明确指明各个模块的具体实现且此时的记忆网络不能实现端到端的训练, 应用场景较少。文献[37]提出端到端的记忆网络, 并在此基础上引入 multiple hop 策略来实现多轮推理。

Sukhbaatar 等人^[37]提出单层和多层的端到端记忆网络, 多层模型是单层模型的堆叠(multiple hop)。图 3 显示网络架构。图 3(a) 的单层记忆网络共有 3 个模块。分别是输入模块, 输出模块和答案预测模块。输入模块基于多个句子组成的文档 d 以及问题 q , 将文档中的每个句子经过输入嵌入矩阵 A , 变换到特征空间中, 并存储在记忆槽 m_i 中(图中蓝色线条)。同时将问题 q 经过嵌入矩阵 B , 变换到同维度的特征空间, 形成特征向量 u , 并将 u 与每个记忆 m_i 进行相似度匹配, 得出相似性权重。输出模块, 将文档中的每个句子经过输出嵌入矩阵 C , 映射到输出特征空间, 并使用输入模块中的相似性权重做加权和, 得到输出特征表示。答案预测模块对输入特征向量 u 与样本输出特征表示进行线性变换, 对候选答案进行打分。

限于单层记忆网络有限的推理能力, 文献[37]对单层记忆网络进行堆叠(multiple hop), 如图 3(b) 所示。在推理过程中, 文档的输入嵌入矩阵 A 和 C 在推理过程中保持不变, 但问题语义信息在与文档的多次交互过程中不断更新。这一过程相当于带着问题多次阅读文档, 不断理解问题, 排除问题中的无效信息, 根据文档抽取出问题答案。

文献[38]提出 Gated Attention Reader(GA Reader), 模型在引入 multiple hop 策略的同时又增加门控注意力机制。在推理过程中, 原问题语义信息保持不变, 多次使用循环神经网络 LSTM 和门控注意力机制精炼文档语义表示。同时使用问题语义信息对文档中词的不同语义维度进行过滤, 门控注意力机制对文档进行提纯, 使得模型具有较强的文档理解能力。

Sordoni 等人^[39]提出迭代交替注意力模型 Iterative Attention Reader(IA Reader), 这种迭代类似于 GA Reader 的多轮推理。IA Reader 采用 GRU 模块进行多轮推理, 模型的交替性体现在使用 GRU 模块的隐藏状态对问题进行一维注意力匹配, 理解问题语义。然后利用问题的语义表示以及 GRU 的隐藏状态对文档进行一维注意力匹配, 提炼文档信息, 整个过程不断地交替进

行, 实现多轮推理.

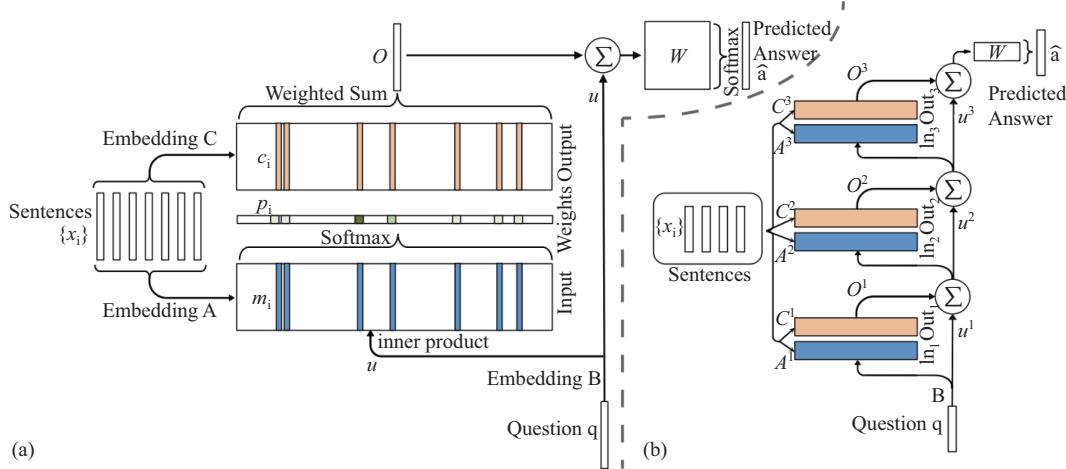


图3 端到端的记忆网络架构图

Fig. 3 Architecture of End-To-End Memory Networks

文献[37-40]中的模型采用堆叠机制来模拟人类的多轮推理过程,但推理的作用点不同。文献[37]更加注重问题的理解,在多次推理过程中不断地使用注意力机制提炼问题语义特征,过滤掉问题的无效信息,根据文档不断地加深模型对问题的理解。文献[38]的GA Reader模型保持问题语义不变。在推理过程中,带着问题反复阅读文档、理解文档、提炼文档,加深模型对文档的理解。文献[39]取上述模型的优点,交替性地提取问题及文档中有效特征,显著提升模型的表示与推理能力。

文献[37-39]预定义推理次数,但不同的问题需要不同的推理能力。简单问题,过度的推理可能陷入过拟合。Shen等人^[40]提出动态推理次数的ReasoNet模型。ReasoNet将编码后的文档d与问题q作为外部记忆模块,将问题最后时刻的隐藏状态作为内部控制器的初始状态,并使用二维注意力计算当前时刻下的问题-文档语义信息,然后同当前控制器的内部状态一起,作为GRU模型输入,更新控制器内部状态。终止门以控制器内部状态作为输入,动态地决定是否有必要继续阅读文档。终止门产生的是二元值:True,结束推理并进行答案预测;False,继续阅读文档。该二元输出并不可导,不能使用梯度算法训练模型。故Shen等人将强化学习^[41]引入ReasoNet进行模型训练。

堆叠推理策略显著提升模型的答案抽取能力,促进机器阅读领域的发展。但是目前推理模型的网络结构还太过单一(大多数采用循环神经网络结构),推理策略还不够灵活。故开发灵活的推理网络结构,更加高效的推理策略已成为该方向急需解决的问题。

2.4 其他模型

机器阅读模型不仅限于注意力模型、推理模型,同时还有基于卷积神经网络(Convolutional Neural Networks, CNN)和预训练的阅读理解模型。由于预训练语言模型具有十分强大的文本表征能力,它还适用分词、命名实体识别、文本分类等自然语言处理任务。

2.4.1 基于CNN的机器阅读模型

大多数基于注意力的模型都使用RNN进行文本编码。RNN循环特性能很好处理文本序列内部的依赖关系,但随之而来的问题是RNN只能串行计算,处理长文本时不能满足低延时的问答场景。因此文献[42-43]探索利用CNN并行计算以及局部特征抽取能力来处理序列数据。

递归网络能够对序列内部的长期依赖进行建模,但串行运算成为低延时问答场景的主要瓶

颈. Wu 等人^[42]提出门控线性空洞残差网络 (Gated Linear Dilated Residual Network, GLDR) 来代替 RNN, 在不降低性能的前提下, 有效降低模型推断时间. GLDR 模块的组成如图 4 所示.

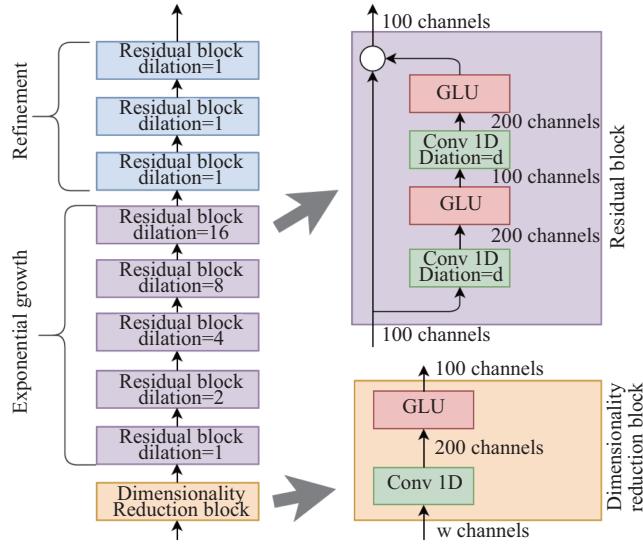


图 4 门控线性空洞残差网络架构图

Fig. 4 Architecture of Gated Linear Dilated Residual Network

模型分为 3 部分: 降维模块, 感知模块, 聚合模块. 降维模块将词向量维度降低至 100 维, 先使用普通一维卷积变换到 200 维. 鉴于文献 [44] 提出的线性门控单元在语言建模中的良好性能, 将前 100 维加上 sigmoid 变换, 作为控制单元, 对后 100 维的词向量进行语义过滤. 感知模块由空洞卷积以及门控线性单元所构成的残差块组成, 残差网络可以有效防止梯度消失等问题, 空洞卷积使得模型看到更大范围的上下文, 模型的长期依赖建模能力得以增强. 聚合模块由普通一维卷积构成, 再次扩大模型的感受域, 提高模型文本表示能力. 在其他同等条件下, 文献 [42] 将 BiDAF 模型的 LSTM 模块替换成 GLDR 模块, 模型性能稍微提升, 训练与推断时间显著减少. 相比 LSTM, GLDR 的训练时间减少到原来的 $1/6$, 单个样本的推断时间减少到原来的 $1/12$. 虽然 GLDR 能够通过使用空洞卷积来增大感受域, 但空洞卷积并不能为文本序列内部单词的位置关系进行建模. Yu 等人^[43]也使用 CNN 解决机器阅读理解任务, 同时还引入位置嵌入来强化单词序列之间的位置信息.

Yu 等人^[43]提出只使用卷积、自注意力的机器阅读模型 QANet. 与 GLDR 不同, 它简化文献 [45] 中复杂的位置嵌入, 仅使用正弦函数、余弦函数模拟位置嵌入增强 QANet 对序列单词的位置感知能力. 同时采用文献 [46] 中的深度可分离卷积 (depthwise separable convolutions) 来进行序列建模. 无论是空洞卷积还是深度可分离卷积, 单层卷积网络的感受域十分有限, 文献 [47] 中的多头自注意力机制直接地处理文本序列内部各个单词之间的依赖关系, 使得模型能够从多个方面来理解文本语义. 鉴于 QANet 训练时间短且数据量相对较少等原因, Yu 等人^[43]使用回译数据增强技术来进行数据增强, 显著提升模型性能. 图 5 显示 QANet 网络架构.

文献 [42-43] 虽然都采用不同的卷积网络对文本进行建模. 但 Wu 等人^[42]的对比实验 (Ablation study) 表明, 去掉空洞卷积对实验结果的影响远小于门控线性单元. Yu 等人^[43]的对比实验 (Ablation study) 表明, 使用常规卷积代替深度可分离卷积, 模型性能会有稍微下降,

但数据增强却能显著提升模型性能.

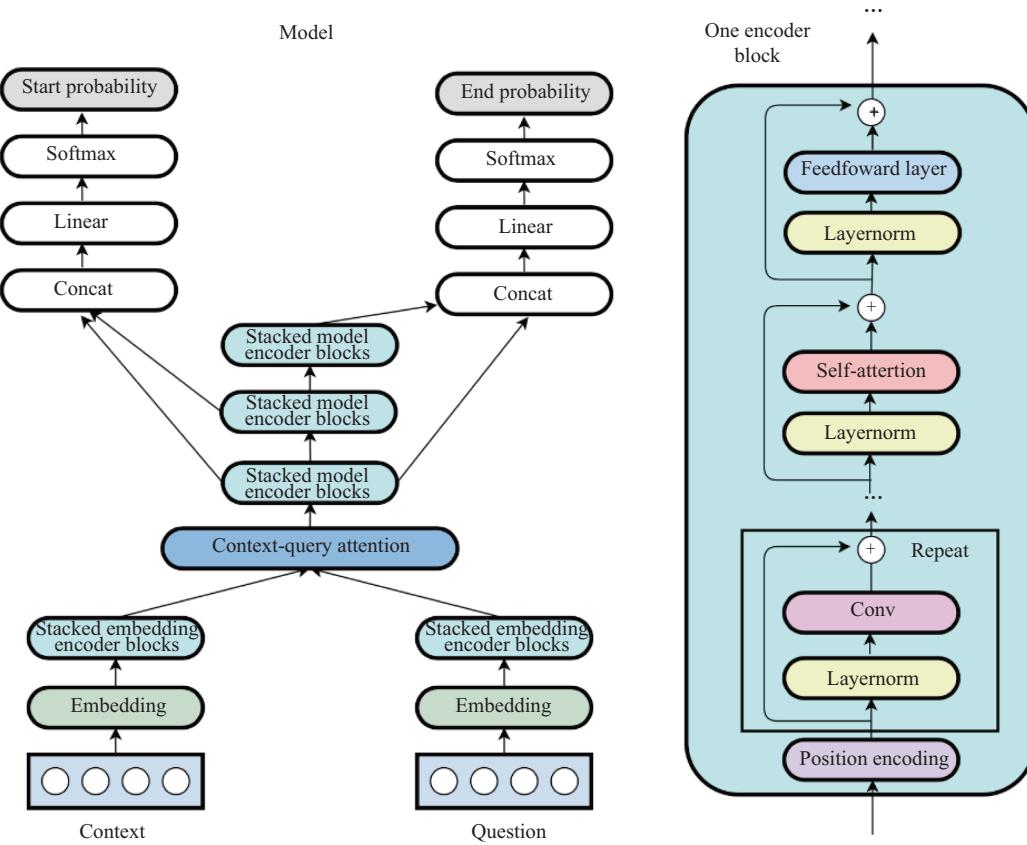


图 5 QANet 网络架构图

Fig. 5 Architecture of QANet

2.4.2 基于预训练语言模型

在大规模语料中预训练语言模型, 将语言模型所包含的知识迁移到其他任务并做微调已成为 NLP 领域的新范式. 使用机器阅读理解模型的答案预测模块替换预训练语言模型的输出层, 再通过微调就能获得相当优异的答案抽取效果. 本小节介绍文献 [48-49] 中的预训练语言模型, 并比较各个预训练模型的优缺点.

2.4.1 节中的对比实验表明, 卷积神经网络的各个变种在文本建模方面并不强大. 当文献 [47] 提出只基于自注意力和全连接的 Transformer 模型取得最佳翻译效果时, 验证了自注意力以及全连接模块的特征抽取能力强于 RNN. 鉴于 Transformer 强大的文本表示能力, 文献 [48] 提出仅由 Transformer 的解码器堆叠而成的单向预训练语言模型 GPT. 由于训练好的语言模型包含语料中大量知识, 将训练好的语言模型迁移到机器阅读任务, 预训练语言模型仅需少量的监督数据和微调时间就能达到优异的性能.

文献 [49] 提出双向预训练语言模型 BERT. 该模型提出 Mask Language Model 和 Next Sentence Prediction 两种预训练子任务. MLM 预训练任务会随机遮掉一句话中部分词, 充分利用被遮词的上文、下文来预测被遮词, 这使得语言模型中的词表示包含上下文信息. Next Sentence Prediction (NSP) 预训练任务用来判断语料中的两句话是否具有前后关系, 该子任务对句子关系进行建模, 有利于机器阅读这种需要模拟文档、问题语义交互任务.

相比于文献[48]中单向预训练语言模型 GPT, BERT 这种双向语言模型包含更丰富的语义与语言结构信息。同时, GPT 仅仅使用单向语言模型, 而 BERT 中的 NSP 训练子任务使得模型能够更充分地捕获句子对之间的语义关系。在多项自然语言推断任务上(特别是句子对关系建模的任务, 如句子对匹配任务、机器阅读任务)表明, BERT 的模型性能普遍优于 GPT。

2.5 模型比较与最新进展

2.2 节至 2.4 节, 介绍注意力模型, 推理模型以及卷积、预训练模型。它们之间的优点以及不足如下:

多数一维匹配模型用来处理填空式任务, 模型将问题的表示压缩成固定的稠密向量, 导致问题信息丢失。但相比段落抽取式任务, 填空式阅读任务复杂程度低, 一维匹配模型更适合此类任务。二维匹配模型针对填空式任务以及段落抽取式任务, 这类模型, 对问题、文档语义关系进行建模。相比一维匹配模型, 它能充分提取文本特征, 提升模型的文本表示能力。

注意力模型伴随着 RNN 的使用, RNN 将历史语义压缩成稠密向量, 丢失部分语义特征。推理模型引入外部记忆模块来更好地存储单词序列历史信息, 同时使用 multi-hop 策略来模拟人类的多轮推理过程。相比注意力模型, 推理模型能够更好地存储历史语义信息, 增强模型的复杂推理能力。

基于注意力模型、推理模型都会使用 RNN 模块来处理序列信息。RNN 循环的特性能够很好地处理单词序列依赖关系, 但 RNN 不能并行计算, 难以满足低延时的问答场景。因此 2.4 节探索使用卷积网络来处理序列数据, 相比于注意力、推理模型, 基于卷积网络模型在性能不降低的前提下, 又能大幅度减少模型训练时间与推断时间, 非常适用于低延时的问答场景。

目前使用预训练语言模型与微调相结合的方法在机器阅读以及多数自然语言处理任务中取得最优结果, 并成为当前火热的研究方向。这种预训练语言模型采用自注意力机制, 序列内部的各个单词直接相互作用, 在文本特征抽取能力上已超越传统的 RNN。由于预训练语言模型参数众多且训练语料充足, 它能够从语料中学习大量的知识, 显著减少 NLP 任务的监督数据和训练时间。

3 数据集和评估标准

大规模公开数据集是训练神经机器阅读模型的基石, 可靠的评估标准是衡量模型语言理解能力的重要指标。3.1 节先介绍近年来所提出的数据集, 概述不同类型数据集之间的区别。3.2 节介绍数据集的评估标准。3.3 节对代表性模型进行比较。

3.1 机器阅读数据集

不同的数据集要求不同的机器阅读任务。本文将已有的机器阅读数据集分为两类: 填空式数据集, 段落抽取式数据集。表 4 显示近年来被提出的机器阅读数据集。

不同类别机器阅读数据集要求模型对文本理解的程度、推理能力各不相同。Chen 等人^[25]经过采样分析表明, 填空式机器阅读任务不需要复杂的网络模型, 它对于模型的推理能力要求不高。单段落以及多段落抽取式机器阅读数据集需要模型具有较强的文本特征抽取能力和推理能力。填空式机器阅读的答案是一个实体或者动词。单段落抽取式机器阅读数据集需从文档中抽取部分内容作为答案, 多段落抽取式数据集要求分析各个段落之间的关系, 从多个段落中抽取答案。所以段落抽取式任务对模型提出更高的要求, 多段落抽取任务的难度也比单段落抽取的难度要高。

3.2 数据集的评估标准

评估标准是衡量模型是否具有自然语言理解能力的重要指标, 本文阐述的两类数据集有两种不同的评估标准。

表 4 机器阅读数据集总结

Tab. 4 Summary of Machine reading datasets

任务	数据集	语言	规模	问题来源	文档来源	答案
填空式	MCTest [5]	EN	2K/500	Crowdsourced	Fictional stories	Multi. choices
	CNN/DM [24]	EN	1.4M/300K	Synthetic cloze	News	Fill in entity
	RACE [19]	ZH	870K/50K	English exam	English exam	Multi. choices
	HLF_RC [28]	ZH	100K/28K	Synthetic cloze	Fairy/News	Fill in word
	CBT [50]	EN	688K/108	Synthetic cloze	Project Gutenberg	Multi. choices
段落抽取式	SQuAD [15]	EN	100K/536	Crowdsourced	WiKi	Span of words
	TriviaQA [16]	EN	40K/660k	Trivia websites	WiKi/Web doc	Span of words
	NewsQA [17]	EN	100K/10K	Crowdsourced	CNN	Span of words
	SearchQA [20]	EN	140K/6.9M	QA site	Web doc	Span of words
	NarrativeQA [18]	EN	46K/1.5K	Crowdsourced	Book&Movie	Manual summary
	MS MARCO [15]	EN	100K/200K	User logs	Web doc	Manual summary
	DuReader [21]	ZH	200K/1M	Web doc	Web doc/CQA	Manual summary

填空式数据集的答案仅仅是一个实体词或动词, 所以填空式阅读任务的评估指标是准确率。对于段落抽取式数据集, 它的答案是文档中的一个片段, 它的评估指标有精确匹配(Exact Match, EM)和模糊匹配(F_1 -score)。EM 要求模型预测的答案与真实答案完全一样, F_1 计算模型对于单个样本的召回率(recall)、准确率(precision)的调和平均, 最后对所有样本的调和平均取均值。

$$F_1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}. \quad (2)$$

3.3 模型性能对比

表5显示具有代表性的一维匹配、二维匹配以及推理模型在 CNN/Daily Mail 数据集上的性能比较。

表 5 模型在 CNN/Daily Mail 上的性能比较

Tab. 5 Performance comparison of models on CNN/Daily Mail

模型	CNN		Daily Mail	
	Valid	Test	Valid	test
Sukhbaatar 等人 (End to End Memory network) ^[37]	63.4	66.8	NA	NA
Hermann 等人 (Attentive Reader) ^[24]	61.6	63.0	70.5	69.0
Hermann 等人 (Impatient Reader) ^[24]	61.8	63.8	69.0	68.0
Chen 等人 (Standford Attentive Reader) ^[25]	72.4	72.4	76.9	75.8
Kadlec 等人 (AS Reader) ^[26]	68.6	69.6	75.0	73.9
Cui 等人 (CAS Reader) ^[28]	68.2	70.0	NA	NA
Cui 等人 (AoA Reader) ^[29]	73.1	74.4	NA	NA
Sordoni 等人 (Iterative Attention) ^[39]	72.6	73.3	NA	NA
Seo 等人 (BiDAF) ^[31]	76.3	76.9	NA	NA
Shen 等人 (ReasoNet) ^[40]	72.9	72.4	NA	NA

实验表明, 文献[29,31]的二维匹配模型以及文献[39-40]的推理模型在 CNN 数据集上的准确率明显高于一维匹配模型。由此说明, 相对于一维匹配模型, 二维匹配模型与推理模型具有更强的语义关系建模能力, 能更好地对问题、文档进行特征抽取。

表6显示具有代表性的二维匹配模型、基于卷积的机器阅读模型以及预训练语言模型在SQuAD数据集上的性能比较。

表 6 模型在 SQuAD 数据集上的性能比较

Tab. 6 Performances comparison of models on SQuAD dataset

模型	EM	F1
Wang 等人 Match LSTM [30]	60.474	70.695
Seo 等人 (BiDAF) ^[31]	67.974	77.323
Shen 等人 (ReasoNet) [40]	70.555	79.364
Liu 等人 (SAN) [35]	76.828	84.396
Huang 等人 (FusionNet) ^[34]	75.968	83.900
Wu 等人 (GLDR) [42]	69.325	77.886
Wang 等人 (R-Net) [33]	81.391	88.170
Yu 等人 (QANet) ^[43]	82.471	89.306
Devlin 等人 (BERT) [49]	85.083	91.835

实验表明, 基于自注意力机制的二维匹配模型优于传统二维匹配模型, 这表明自注意力机制具有更强的信息提纯能力。同时预训练语言模型在机器阅读领域达到最优性能, 这表明预训练语言模型能够从无监督语料中学习有用知识与文本表示, 有效缓解模型的推理压力, 大幅提高模型性能。

4 总结与展望

本文对神经机器阅读模型进行系统的研究与分析, 概述机器阅读理解的定义, 对现有的神经机器阅读模型进行分类, 总结各自的优点及不足。同时总结现有不同类型的数据集, 概述不同类型数据集的特点, 最后对神经机器阅读理解模型在公开数据集上的性能进行比较。

虽然基于神经网络的机器阅读理解模型具有强大的表示能力, 并在公开数据集上取得较好的效果。但神经网络阅读器在诸如教育自动问答中的应用依然存在以下问题, 值得进行下一步研究, 具体如下。

- 多模态输入数据。虽然基于机器阅读的自动问答系统能够很好表征文本、理解文本。但实际应用中, 学生抛给自动问答系统的输入数据可能是多模态的, 如几何题中文本与图片的混合数据, 如何表征多模态数据是自动问答面临的首要问题。目前Transformer已成为自然语言领域最强的特征抽取器, 如何应用Transformer于多模态数据特征抽取将是下一步研究的热点。
- 推理能力不强。本文虽然介绍推理模型, 但这些模型的推理能力还非常弱, 这种基于multi-hop的推理机制还太过单一, 计算效率低下。最近新兴的图卷积网络不仅计算效率高且具备较强的推理能力, 如何将图卷积的推理能力应用在自动问答领域也值得研究。
- 模型的鲁棒性低。在实际中自动问答系统所接受的用户输入是复杂多变的。对于加入噪声的输入极有可能成为机器阅读模型的对抗样本。机器阅读模型面对对抗样本时, 就会错误地理解文本, 影响模型泛化性能。对抗训练可以缓解上述问题, 但如何将对抗训练引入机器阅读模型的训练过程也值得下一步探究。

[参 考 文 献]

- [1] CHEN D Q. Neural reading comprehension and beyond [D]. CA: Standford University, 2018.
[2] LEHNERT W G. The process of question answering [R]. Yale Univ New Haven Conn, 1977.

- [3] HIRSCHMAN L, LIGHT M, BRECK E, et al. Deep read: A reading comprehension system [C]// Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics. Association for Computational Linguistics, 1999: 325-332.
- [4] RILOFF E, THELEN M. A rule-based question answering system for reading comprehension tests [C]// Proceedings of the 2000 ANLP/NAACL Workshop on Reading comprehension tests as evaluation for computer-based language understanding systems-Volume 6. Association for Computational Linguistics, 2000: 13-19.
- [5] RICHARDSON M, BURGES C J C, RENSHAW E. Mctest: A challenge dataset for the open-domain machine comprehension of text [C]//Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, 2013: 193-203.
- [6] SACHAN M, DUBEY K, XING E, et al. Learning answer-entailing structures for machine comprehension [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 239-249.
- [7] NARASIMHAN K, BARZILAY R. Machine comprehension with discourse relations [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), 2015: 1253-1262.
- [8] WANG H, BANSAL M, GIMPEL K, et al. Machine comprehension with syntax, frames, and semantics [C]// Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers), 2015: 700-706.
- [9] BENGIO Y, DUCHARME R, VINCENT P, et al. A neural probabilistic language model [J]. *Journal of Machine Learning Research*, 2003(3): 1137-1155.
- [10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]// Advances in neural information processing systems, 2013: 3111-3119.
- [11] PENNINGTON J, SOCHER R, MANNING C. Glove: Global vectors for word representation [C]// Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014: 1532-1543.
- [12] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. *Neural computation*, 1997, 9(8): 1735-1780.
- [13] CHUNG J, GULCEHRE C, CHO K H, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling [J]. arXiv preprint, arXiv: 1412.3555, 2014.
- [14] RAJPURKAR P, ZHANG J, LOPYREV K, et al. Squad: 100,000+ questions for machine comprehension of text [J]. arXiv preprint, arXiv: 1606.05250, 2016.
- [15] NGUYEN T, ROSENBERG M, SONG X, et al. MS MARCO: A Human-Generated MAchine Reading COmprehension Dataset [J]. *Neural Information Processing Systems*, 2016.
- [16] JOSHI M, CHOI E, WELD D S, et al. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension [J]. arXiv preprint, arXiv: 1705.03551, 2017.
- [17] TRISCHLER A, WANG T, YUAN X, et al. Newsqa: A machine comprehension dataset [J]. arXiv preprint, arXiv: 1611.09830, 2016.
- [18] KOČSKÝ T, SCHWARZ J, BLUNSMON P, et al. The narrativeqa reading comprehension challenge [J]. *Transactions of the Association for Computational Linguistics*, 2018(6): 317-328.
- [19] LAI G, XIE Q, LIU H, et al. Race: Large-scale reading comprehension dataset from examinations [J]. arXiv preprint, arXiv: 1704.04683, 2017.
- [20] DUNN M, SAGUN L, HIGGINS M, et al. Searchqa: A new q&a dataset augmented with context from a search engine [J]. arXiv preprint, arXiv: 1704.05179, 2017.
- [21] HE W, LIU K, LIU J, et al. Dureader: a chinese machine reading comprehension dataset from real-world applications [J]. arXiv preprint, arXiv: 1711.05073, 2017.
- [22] TAYLOR W L. "Cloze procedure": A new tool for measuring readability [J]. *Journalism Bulletin*, 1953, 30(4): 415-433.
- [23] BAHdanau D, CHO K, BENGIO Y. Neural machine translation by jointly learning to align and translate [J]. arXiv preprint, arXiv: 1409.0473, 2014.
- [24] HERMANN K M, KOCISKÝ HERMANN K M, KOCISKÝ T, et al. Teaching machines to read and comprehend [C]// Advances in neural information processing systems. 2015: 1693-1701.
- [25] CHEN D, BOLTON J, MANNING C D. A thorough examination of the cnn/daily mail reading comprehension task [J]. arXiv preprint, arXiv: 1606.02858, 2016.
- [26] KADLEC R, SCHMID M, BAJGAR O, et al. Text understanding with the attention sum reader network [J]. arXiv preprint, arXiv: 1603.01547, 2016.
- [27] VINYALS O, FORTUNATO M, JAITLEY N. Pointer networks [C]// Advances in Neural Information Processing Systems, 2015: 2692-2700.
- [28] CUI Y, LIU T, CHEN Z, et al. Consensus attention-based neural networks for chinese reading comprehension [J]. arXiv preprint, arXiv: 1607.02250, 2016.
- [29] CUI Y, CHEN Z, WEI S, et al. Attention-over-attention neural networks for reading comprehension [J]. arXiv preprint, arXiv: 1607.04423, 2016.

- [30] WANG S, JIANG J. Machine comprehension using match-lstm and answer pointer [J]. arXiv preprint, arXiv: 1608.07905, 2016.
- [31] SEO M, KEMBHAVI A, FARHADI A, et al. Bidirectional attention flow for machine comprehension [J]. arXiv preprint, arXiv: 1611.01603, 2016.
- [32] KIM Y. Convolutional neural networks for sentence classification [J]. arXiv preprint, arXiv: 1408.5882, 2014.
- [33] WANG W, YANG N, WEI F, et al. Gated self-matching networks for reading comprehension and question answering [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), 2017: 189-198.
- [34] HUANG H Y, ZHU C, SHEN Y, et al. Fusionnet: Fusing via fully-aware attention with application to machine comprehension [J]. arXiv preprint, arXiv: 1711.07341, 2017.
- [35] LIU X, SHEN Y, DUH K, et al. Stochastic answer networks for machine reading comprehension [J]. arXiv preprint, arXiv: 1712.03556, 2017.
- [36] WESTON J, CHOPRA S, BORDES A. Memory networks [J]. arXiv preprint, arXiv: 1410.3916, 2014.
- [37] SUKHBAATAR S, WESTON J, FERGUS R. End-to-end memory networks [C]// Advances in neural information processing systems, 2015: 2440-2448.
- [38] DHINGRA B, LIU H, YANG Z, et al. Gated-attention readers for text comprehension [J]. arXiv preprint, arXiv: 1606.01549, 2016.
- [39] SORDONI A, BACHMAN P, TRISCHLER A, et al. Iterative alternating neural attention for machine reading [J]. arXiv preprint, arXiv: 1606.02245, 2016.
- [40] SHEN Y, HUANG P S, GAO J, et al. Reasonet: Learning to stop reading in machine comprehension [C]// Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, 2017: 1047-1055.
- [41] WILLIAMS R J. Simple statistical gradient-following algorithms for connectionist reinforcement learning [J]. Machine learning, 1992, 8(3/4): 229-256.
- [42] WU F, LAO N, BLITZER J, et al. Fast reading comprehension with convnets [J]. arXiv preprint, arXiv: 1711.04352, 2017.
- [43] YU A W, DOHAN D, LUONG M T, et al. Qanet: Combining local convolution with global self-attention for reading comprehension [J]. arXiv preprint, arXiv: 1804.09541, 2018.
- [44] DAUPHIN Y N, FAN A, AULI M, et al. Language modeling with gated convolutional networks [C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 933-941.
- [45] GEHRING J, AULI M, GRANGIER D, et al. Convolutional sequence to sequence learning [C]// Proceedings of the 34th International Conference on Machine Learning-Volume 70. JMLR. org, 2017: 1243-1252.
- [46] CHOLLET F. Xception: Deep learning with depthwise separable convolutions [C]// Proceedings of the IEEE conference on computer vision and pattern recognition, 2017: 1251-1258.
- [47] VASWANI A, SHAZER N, PARMAR N, et al. Attention is all you need [C]// Advances in neural information processing systems, 2017: 5998-6008.
- [48] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding with unsupervised learning [R/OL]. Technical report, OpenAI, 2018. [2019.08.01]. https://s3-us-west-2.amazonaws.com/openai-assets/research-covers/language-unsupervised/language_understanding_paper.pdf.
- [49] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding [J]. arXiv preprint, arXiv:1810.04805, 2018.
- [50] HILL F, BORDES A, CHOPRA S, et al. The goldilocks principle: Reading children's books with explicit memory representations [J]. arXiv preprint, arXiv:1511.02301, 2015.

(责任编辑: 李万会)