计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(4):1−9 [doi: 10.15888/j.cnki.csa.007327]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

利用外部知识辅助和多步推理的选择题型机器阅读理解模型①

盛艺暄, 兰　曼

(华东师范大学 计算机科学与技术学院, 上海 200062)
通讯作者: 兰　曼, E-mail: mlan@cs.ecnu.edu.cn

摘　要: 选择题型机器阅读理解的答案候选项往往不是直接从文章中抽取的文本片段, 而是对文章内容中相关片段的归纳总结、文本改写或知识推理, 因此选择题型机器阅读理解的问题通常需要从给定的文本中甚至需要利用外部知识辅助进行答案推理. 目前选择题型机器阅读理解模型大多数方法是采用深度学习方法, 利用注意力机制对文章、问题和候选项这三者的信息进行细致交互, 从而得到融合三者信息的表示进而用于答案的预测. 这种方式只能利用给定的文本进行回答, 缺乏融入外部知识辅助, 因而无法处理需外部知识辅助推理的问题. 为了解决需外部知识辅助推理的问题, 本文提出了一个采用外部知识辅助多步推理的选择题型机器阅读理解模型, 该模型首先利用注意力机制对文章、问题和候选项及与这三者相关的外部知识进行信息交互建模, 然后采用多步推理机制对信息交互建模结果进行多步推理并预测答案. 本文在 2018 年国际语义测评竞赛 (SemEval) 中任务 11 的数据集 MCScript 上进行对比实验, 实验结果表明本文提出的方法有助于提高需要外部知识辅助的选择题型问题的准确率.

关键词: 机器阅读理解; 选择题型问答; 注意力机制; 多步推理机制; 外部知识辅助

引用格式: 盛艺暄,兰曼.利用外部知识辅助和多步推理的选择题型机器阅读理解模型.计算机系统应用,2020,29(4):1–9. http://www.c-s-a.org.cn/1003-3254/7327.html

Leveraging Commonsense Knowledge to Assist Multi-Step Reasoning for Multiple Choice Machine Reading Comprehension

SHENG Yi-Xuan, LAN Man

(School of Computer Science and Technology, East China Normal University, Shanghai 200062, China)

Abstract: Typically, the options of multiple choice Machine Reading Comprehension (MRC) are not directly extracted from the given document. Thus the answers need to be summarized or rewritten or even inferred from document or from the world's knowledge. Most existing models adopt attention mechanism to generate an interactive representation of document, question, and option. However, these models are limited by only using the given document rather than common knowledge, leading to poor result when dealing with questions requiring external knowledge assistance reasoning. To address questions requiring external knowledge assistance reasoning, we propose a novel neural model by integrating external commonsense knowledge to assist multi-step reasoning. Our model first interacts information among document, question, options, and related external knowledge by attention mechanism and then predicts answer by multi-step reasoning through the interaction results. The experimental results on the SemEval-2018 MCScript corpus show that the proposed model improves the accuracy of question answering requiring common knowledge reasoning.

Key words: machine reading comprehension; multiple-choice question answering; attention mechanism; multi-step reasoning mechanism; commonsense knowledge

① 收稿时间: 2019-08-14; 修改时间: 2019-09-06; 采用时间: 2019-09-19; csa 在线出版时间: 2020-04-05

机器阅读理解是自然语言处理的核心任务之一, 目标是利用机器自动理解自然语言文本, 从给定文本中获取回答问题的答案信息. 高质量的机器阅读理解对于搜索、问答、智能对话等任务都发挥重要作用. 根据答案是否是从文章中直接抽取的文本片段, 机器阅读理解任务可以分为两类, 抽取式和非抽取式. 典型的抽取式任务有 CNN/DailyMai[1], CBTest[2]等这类完形填空任务, 以及 SQuAD[3], TriviaQA[4]等这类片段抽取型任务. 典型的非抽取式任务有 MCTest[5], RACE[6], MCScript[7]等选择题型任务.

选择题型任务与上述直接抽取式任务最大不同在于答案候选项往往是与文章内容相关的归纳总结或文本改写, 而非完全截取自文章中某文本片段. 因此, 需要机器通过从文章或借助外部知识进行推理, 获得正确答案. 表 1 是来自 MCScript 数据集中两种类型问题举例, 问题 1 的回答可由文章中下划线句子推理得出去蒸桑拿是为了解压: 而问题 2 则需要借助外部知识辅助推理, 因为全文并没有提及蒸桑拿所穿服装的信息, 因此结合文章中粗斜体句子信息 (蒸桑拿很热), 并利用 (steam bath, RelatedTo, sauna), (bath suit, RelatedTo, swimsuit), (nude, DerivedFrom, swim), (jeans, UsedFor, clothing), (clothing, UsedFor, warm) 等从外部知识库得到的相关外部知识, 帮助机器正确推理并回答问题 2. 相比问题 1 这类仅需推理文章可得到答案的问题, 问题 2 这类需外部知识辅助的问题属于选择题问题中难度大的问题, 因为它不仅需要文章信息, 还需要外部知识辅助才能得到答案, 对模型推理要求更高.

表 1 选择题中两种类型问题举例

文章	
After an extremely hard week at work, I decided that I wanted to treat myself to something nice. For this something nice, I decided to go to a sauna. I had never been to a sauna before but I had heard a lot of my friends talk about how much they loved it and how relaxing it was for them, so I figured I would give it a try. I got to the sauna and was greeted by a very calm and soothing employee who first directed me to a massage room. I waited and got a great massage which made me very relaxed and ready to go into the sauna. I entered the sauna and the first thing I noticed was that it smelled very woody, almost like being deep in a forest. ***It was also very hot inside*** but it wasn't unpleasant. I sat down in the corner and closed my eyes, trying to relax. I spent about thirty minutes in the sauna enjoying the heat and took a shower when I was finally ready to leave. Going to the sauna was the perfect relaxation I needed after my stressful week!	
没有用到外部知识辅助回答的问题	用到外部知识辅助回答的问题
问题 1: Why did they go to the sauna? 选项: (A). They had a free coupon for a sauna session. (B). They were stressed out. 正确答案: (B)	问题 2: What should they wear? 选项: (A). Swimsuit or nude. (B). Jeans and a T-shirt. 正确答案: (A)

解决阅读理解的问题时, 可能会同时遇到问题 1 这类仅需推理文章就可获得答案的问题, 以及问题 2 这类需外部知识辅助推理的问题, 需要模型对两类问题都能较好地处理才能取得总体性能的提高. 然而, 目前选择题型机器阅读理解模型的研究大多是解决仅需推理文章可获得答案的问题. 这些模型精心设计文章、问题和候选项的语义表示以及这三者之间信息交互方式. 例如, Parikh 等[8]利用正交表示排除不相关的选项, 并根据排除操作的结果增强文章的语义表示. Zhu 等人[9]利用选项之间的相关性建模选项语义表示. Wang 等人[10]将问题和选项分别与文章同时进行交互得到既包含问题信息又包含选项信息的文章表示用于答案预测. 此外, 也有一些研究工作意识到推理在选择题型阅读理解中的重要性, 设计了多种多步推理方式. 例如, Lai 等人[6]将 GA Reader[11]应用于选择题, 利用多步推理机制与文章问题之间的注意力机制结合的方式, 进行固定次数问题和文章之间的信息交互用于答案预测. Xu 等人[12]在计算文章、问题和侯选项三者交互的语义表示后, 采用强化学习进行动态多步推理得出正确答案. 然而, 上述模型都是针对仅需推理文章可获得答案的问题, 没有外部知识的辅助, 模型处理类似问题 2 需外部知识辅助推理的问题时能力受限, 无法正确推理回答问题, 因此需外部知识辅助推理的问题是这类模型提高总体性能的一个突破点.

为帮助回答问题 2 这类需外部知识辅助推理的问题, Wang 等人[13]和 Chen 等人[14]尝试把从外部知识库中得到的词之间关系的表示向量加入词的嵌入表示中, 通过隐式方式引入相关外部知识信息, 但这些模型仅

仅只利用外部知识信息以及文章、问题和候选项这三者的信息进行浅层的语义匹配. 然而, 利用外部知识辅助的问题往往对模型推理要求更高, 上述模型并没有进行多步推理, 导致该类模型不能深入地利用外部知识信息以及文章、问题和候选项这三者的信息共同进行融合推理.

为了解决需外部知识辅助推理的问题, 从而帮助提升选择题机器阅读理解模型的总体性能. 本文提出一个利用外部知识辅助和多步推理的选择题型机器阅读理解模型, 该模型不仅引入与回答问题相关的外部知识, 进行外部知识与文章、问题和候选项的信息交互用于建模语义匹配表示, 还利用包含外知识信息的语义匹配表示进行多步答案推理, 从而更好地利用外部知识信息帮助模型推理出答案. 本文在国际语义测评竞赛 2018 年任务 11 的数据集 MCScript[7]上进行了对比实验, 结果表明本文提出的方法有助于提高需外部知识辅助回答的问题的准确率.

本文的组织结构如下: 首先在上述内容中介绍了本文的研究动机及相关工作, 然后将在第 1 节介绍本文提出的模型; 在第 2 节中说明实验设置, 包含数据准备, 模型参数设置和评估标准; 在第 3 节中展示实验结果并分析; 最后在第 4 节总结本文工作.

1 利用外部知识辅助和多步推理的选择题型模型

本章将介绍本文所提出的利用外部知识辅助和多步推理的选择题型机器阅读理解模型. 本文目标是提高需外部知识辅助推理的问题性能来帮助提高总体性能, 所以需要引入一定量的外部知识, 然而仅需推理文章得到答案的问题原本只需要利用文章信息就可推理出答案, 额外引入的外部知识对其而言可能会是噪声信息, 容易影响仅需推理文章得到答案的问题性能. 因此, 在实现本文提高需要外部知识辅助的问题的目标的同时, 应至少保持仅需推文章得到答案的问题性能不降低. 因此本文基于 Co-Matching 模型[9](一个针对解决仅需推理文章得到答案的问题的鲁棒性强的选择题型机器阅读理解模型), 增加对外部知识的显示利用和多步知识推理功能, 通过进一步提升模型解决需外部知识辅助推理问题的能力, 从而提高模型总体性能. 下面首先介绍本文提出的模型, 然后阐述本模型与 Co-Matching 模型的不同.

与解决仅需推理文章信息得到答案的模型不同, 本文模型除了输入文章、问题和选项, 还需要输入与文章、问题和选项这三者有关的外部知识. 通常外部知识是采用图结构存储的知识库形式, 每条外部知识采用三元组的形式 (外部知识选取的相关内容将在第 2 节实验配置介绍). 图 1 为本文提出的模型的概括图, 展示了本文模型预测一个选项作为答案的概率的过程. 将所有数据输入以后, 该模型通过嵌入层、编码层、外部知识融入层、文章问题及选项信息交互层、答案层进行逐层处理, 并在答案层输出预测的答案. 下面从模型的输入开始, 按数据处理的顺序逐步介绍上述 5 层.

1.1 模型的输入

为便于模型描述, 本文将采用下列符号化的表示定义选择题型机器阅读理解任务. 该任务输入一个以四元组表示的选择题型阅读理解的全部数据$<D,Q,C,K>$, 模型从答案候选项集合C中选择概率最高的选项作为预测答案输出. 四元组中给定的文章 D 包含s个句子$\{DS_1,DS_2,\cdots,DS_s\}$, 其中第$j$句$DS_j$的长度为$l_{DS_j}$个词, 问题 Q 的长度为l_Q个词, 答案候选项集合C是包含y个选项的集合$\{c_1,c_2,\cdots,c_y\}$, 其中第i个选项c_i的长度为l_{c_i}个词, 相关知识集 K 是从外部知识库中筛选出的与 D, Q, C 有关的集合, 是包含p个外部知识的三元组的集合, 其中每个知识三元组k表示为$k=(subj,rel,obj)$. $subj$和obj是至少包含一个词的三元组头、尾节点, 长度分别为l_{subj}和l_{obj}. rel是$subj$和obj之间的关系, 长度为 1.

1.2 嵌入层

嵌入层将输入的四元组中的每个词投射到一个语义表示的向量空间. 以问题 Q 为例, 其中每个词表示为词嵌入后得到$E^Q=\{e_1^Q,\cdots,e_{l_Q}^Q\}$, 其中$e_n^Q$表示问题中第$n$个词的嵌入表示, 本文采用预训练的 Glove 词向量[15]并拼接人工特征作为每个词的嵌入表示. 人工特征为多个二值特征, 表示该单词是否出现在文章 D 中、问题 Q 中、候选项$c_i(i\in[1,y])$中、同时出现在文章D和候选项$c_i(i\in[1,y])$中, 同时出现在问题 Q 和候选项$c_i(i\in[1,y])$. 同样, 对文章的句子$DS_j(j\in[1,s])$和选项$c_i(i\in[1,y])$中每个词进行向量映射后, 分别得到E^{DS_j}和E^{c_i}. 此外, 外部知识集合中每个三元组的词进行向量映射后, 可表示为$k=(E^{subj},E^{rel},E^{obj})$.

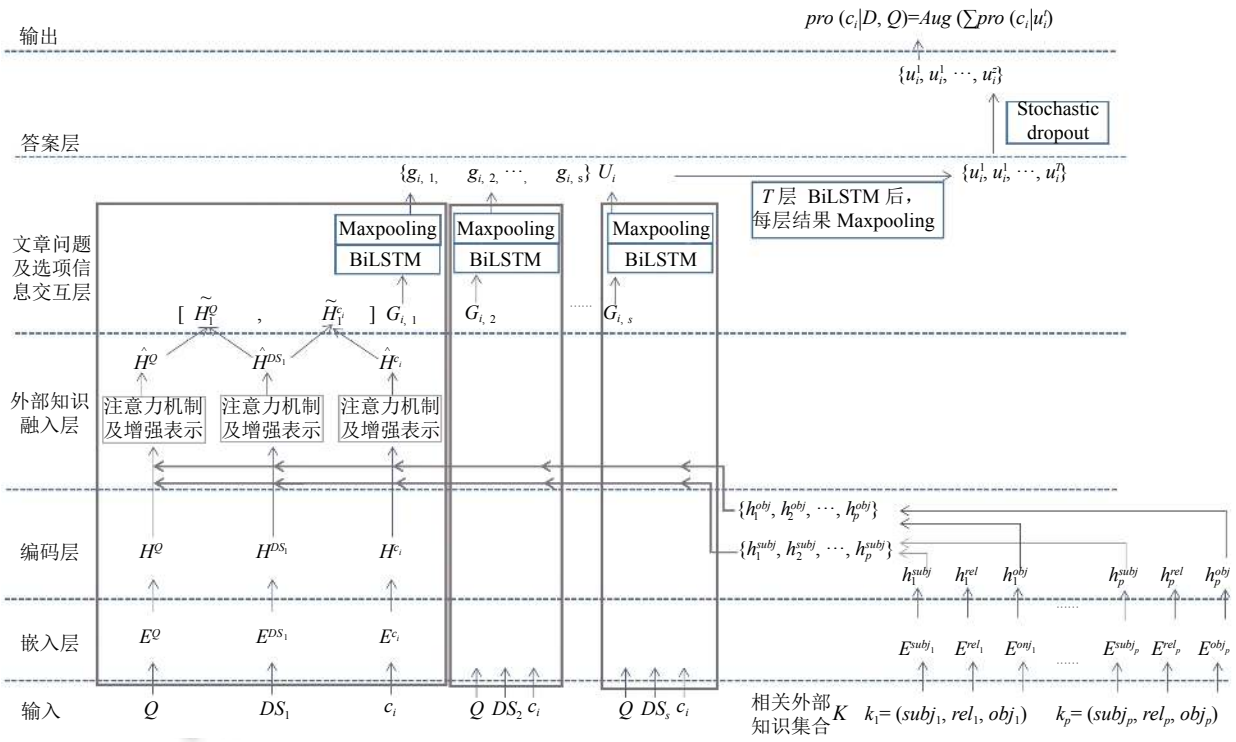

图 1 利用外部知识辅助和多步推理的选择题阅读理解模型

1.3 编码层

编码层是为每个词的嵌入表示生成上下文依赖的编码表示. 本文使用双向长短期记忆网络 (BiLSTM) 对文本进行语义编码. 以问题Q为例, 采用如式 (1) 经过 BiLSTM 得到其语义表示$H^Q \in R^{l_Q \times 2h}$:

$$H^Q = BiLSTM(E^Q) = \{h_1^Q, \cdots, h_{l_Q}^Q\} \tag{1}$$

其中, $h_n^Q \in R^{1\times 2h}$表示问题中第n个词在 BiLSTM 对应各时刻的隐层拼接, h 为隐层大小. 对文章中的句子和选项也进行同样的处理得到各自的语义表示H^{DS_j}和H^{c_i}. 对每一条外部知识三元组, 本文参考 Mihaylov 等人[16]提出的方法, 把外部知识三元组映射到与问题、文章、选项三者相同的向量空间. 同样采用 BiLSTM 对知识三元组中的E^{subj}, E^{rel}和E^{obj}分别进行编码, 可得到编码后的外部知识三元组表示$k = (h^{subj}, h^{rel}, h^{obj})$, 其中$h^{subj}$, h^{rel}和$h^{obj} \in R^{1\times 2h}$, 是各自双向编码的最后时刻的隐层拼接.

1.4 外部知识融入层

外部知识融入层是为了把外部知识引入模型, 受 Mihaylov 等人[16]解决完形填空中针对名词类型问题的启发, 本文在进行文章问题及选项的信息交互之前, 通过外部知识和文章、问题及选项中每个词的语义匹配关系得到每条外部知识的权重, 并按各条外部知识的权重表示文章、问题和选项的每个词, 再将每个词分别按外部知识的权重和上下文编码的权重的结果进行按权重的线性加和 (见式 (4)), 这样得到既包含外部知识信息又包含上下文依赖信息的词表示. 以问题Q中第n个词为例, 如式 (2), 本文首先采用注意力机制求得外部知识集合 K 中每条外部知识三元组和该词之间的关系权重, 并按权重线性加和得到$v_n^Q \in R^{1\times 2h}$:

$$v_n^Q = softmax(h_n^Q\{h_1^{subj}, \cdots, h_p^{subj}\}^{\mathrm{T}})\{h_1^{obj}, \cdots, h_p^{obj}\} \tag{2}$$

其中, $\{h_1^{subj}, \cdots, h_p^{subj}\}$和$\{h_1^{obj}, \cdots, h_p^{obj}\} \in \mathrm{R}^{p\times 2h}$. 以前的研究[17,18]经验地发现对$v_n^Q$用式 (3) 进行 $ReLU$ 的非线性处理后, 通常语义表示效果更好, 本文借鉴了他们的处理, 如下式 (3) 计算后可得问题Q中第n个词优化后的外部知识按权重的表示$\bar{h}_n^Q \in R^{1\times 2h}$:

$$\bar{h}_n^Q = ReLU(W_1[v_n^Q - h_n^Q, v_n^Q * h_n^Q] + b_1) \tag{3}$$

其中, $ReLU$是一种非线性激活函数, [,]是拼接操作, −和∗是元素减和元素乘操作, W_1和b_1是训练参数. 随后将Q中第n个词优化后的外部知识按权重的表示向量$\bar{h}_n^Q$和上下文编码向量h_n^Q进行有权线性加和, 如式 (4)

得到$\hat{h}_n^Q$:

$$\hat{h}_n^Q = \alpha * \bar{h}_n^Q + (1-\alpha) * h_n^Q \tag{4}$$

其中, $\alpha \in R^{1\times 2h}$是一个随机初始化可训练的向量. 经该层处理后, 问题Q表示为$\hat{H}^Q = \{\hat{h}_1^Q, \cdots, \hat{h}_{l_Q}^Q\}$, 其中每个词表示都包含外部知识信息和上下文信息. 类似地, 对文章中的句子和选项采用同样操作可分别得到$\hat{H}^{DS_j}$和$\hat{H}^{c_i}$.

1.5 文章问题及选项信息交互层

为了生成文章、问题和候选项这三者之间的交互语义匹配表示, 使模型关注到这三者匹配度高的部分, 信息交互层在 Wang 等人[9]提出的 Co-Matching 匹配方式的基础上进行了改进. Co-Matching 模型仅仅利用文章、问题和候选项三者的上下文编码直接进行语义匹配, 缺乏外部知识的辅助, 因而回答需外部知识辅助推理的问题时能力受限, 而为了使外部知识的信息能够参与语义匹配中, 本文使用外部知识融入层的输出替代只有上下文编码信息的表示, 并通过注意力机制同时计算问题和文章中每个句子的语义匹配关系以及各个选项和文章中每个句子的语义匹配关系. 以问题Q和文章中第j句DS_j的信息交互为例, 用如式 (5) 和式 (6) 计算得到包含外部知识信息和问题信息的文章中第j句的表示$\tilde{H}_j^Q$:

$$\hat{V}_j^Q = softmax(\hat{H}^{DS_j}(W_2\hat{H}^Q + b_2)^T)\hat{H}^Q \tag{5}$$

$$\tilde{H}_j^Q = ReLU(W_3[\hat{V}^Q - H^{DS_j}, \hat{V}^Q \times H^{DS_j}] + b_3) \tag{6}$$

同时对选项c_i进行类似处理可得$\tilde{H}_j^{c_i}$. 然后将$\tilde{H}_j^Q$和$\tilde{H}_j^{c_i}$进行拼接可得既包含外部知识信息又包含问题Q和选项c_i信息的文章中第j句的语义匹配表示$G_{i,j} = [\tilde{H}_j^Q, \tilde{H}_j^{c_i}]$. 对每个问题、文章中某句子和某一个候选项构成的任一组合$\{Q, DS_j(j \in [1,s]), c_i(i \in [1,y])\}$都可采用上述四层处理, 得到一个语义匹配的表示$G_{i,j}$, i和j分别表示第i个选项和文章第j句. 以第i个候选项c_i为例, 在得到融入了外部知识、问题和候选项信息的文章中所有句子的表示$\{G_{i,1}, G_{i,2}, \cdots, G_{i,s}\}$后, 采用如式 (7) 将文章句子的矩阵表示$G_{i,j}$变成向量表示$g_{i,j}$:

$$g_{i,j} = maxpooling(BiLSTM(G_{i,j})) \tag{7}$$

由此可以得到由文章中所有句子语义匹配表示向量组成的语义匹配表示$U_i = \{g_{i,1}, g_{i,2}, \cdots, g_{i,s}\}$, U_i中含有经过语义匹配的外部知识辅助信息、以及问题文章和选项的信息.

1.6 答案层

答案层采用多步推理机制实现答案预测. 用多层 BiLSTM 对信息交互层得到的语义匹配表示U_i进行推理, 根据每层 BiLSTM 的输出, 得到一系列输出结果$\{U_i^1, U_i^2, \cdots, U_i^T\}$, 其中 T 为总推理步数, U_i^t为第 t 层 BiLSTM 的输出. 然后, 对每层 BiLSTM 的输出$\{U_i^1, U_i^2, \cdots, U_i^T\}$使用最大池化操作得到$\{u_i^1, u_i^2, \cdots, u_i^T\}$. 在多步推理的过程中, 为避免出现某步推理偏执的问题 (“step bias problem”), 本文对多步推理的结果$\{u_i^1, u_i^2, \cdots, u_i^T\}$使用 Liu 等人[19]提出的随机丢失策略 (Stochastic dropout), 即按一定的随机丢失概率随机丢弃$\{u_i^1, u_i^2, \cdots, u_i^T\}$中的一些结果, 随机留下$z(z \le T)$步推理结果. 然后本文根据留下的推理结果计算选项$c_i$为预测答案的概率, 并计算余下步骤的概率均值, 得到综合考量多个推理步骤后选项c_i作为预测答案的平均概率:

$$pro(c_i|u_i^t) = \frac{\exp(wu_i^t + b)}{\sum_{j=1}^{y} exp(wu_j^t + b)} \tag{8}$$

$$pro(c_i|D,Q) = Avg(\sum^{z} pro(c_i|u_i^t)) \tag{9}$$

其中, w和b是训练参数, $pro(c_i|u_i^t)$是选项c_i在第t步推理时作为答案输出的概率. $pro(c_i|D,Q)$是给定模型文章和问题后, 最终得到的选项c_i作为答案输出的概率. 模型输出概率最高的选项为预测答案.

1.7 本文模型与 Co-Matching 模型的异同点比较

本文提出的模型是基于 Co-Matching 模型, 但与 Co-Matching 模型在各层中都存在不同. (1) 在嵌入层, 本文加入了人工特征表示词匹配关系, 对回答一些候选项与文章有词共现的问题有帮助. (2) 在编码层, 本文增加对外部知识的编码. (3) 额外增加外部知识融入层丰富词表示, 使模型具有灵活丰富的外部知识的信息辅助. (4) 在文章问题及选项信息交互层, Co-Matching 模型仅利用文章, 问题和选项的信息做语义匹配, 本文模型使用外部知识融入层的输出, 在进行三者信息交互时有外部知识的信息参与, 增强了机器阅读理解模型回答需要外部知识辅助类问题时的语义匹配能力. (5) 在答案层, 采用了多步推理机制而并非单层答案预测层, 可更进一步利用外部知识信息帮助推理, 解决了仅利用外部知识信息进行浅层语义匹配的缺陷, 增强了模型的推理能力, 更有利于推理需外部知

识辅助的问题.

2 实验设置

本章介绍实验的设置, 包括数据准备, 模型参数设置和评估指标.

2.1 数据准备

数据集: 本文使用国际语义测评竞赛 2018-SemEval 的任务 11 的数据集 MCScript 进行实验. MCScript 是旨在于让大家探索使用外部知识进行机器阅读理解的数据集, 其中每篇文章对应若干问题, 每个问题有 2 个选项, 需从中选择一个正确答案. 数据集中同时包含需要外部知识辅助推理的问题 (表中“cs”类型) 和仅需文章信息可推理答案的问题 (表中“text”类型), “cs”类问题和“text”类问题比例约 3:7, “cs”型的问题数量比“text”类型的问题的数量少. 因此, “cs”类型问题应该有较大变化才能影响数据集的总体效果. 数据集的具体分布如表 2 所示, “all”是所有问题, “#”表示数量.

表 2 数据集 MCScript 的数据分布

数据集	文章	问题		
		#text	#cs	#all
train	1470	7032	2699	9731
dev	219	1006	405	1411
test	430	2074	723	2797

数据预处理: 首先采用 Stanford CoreNLP 对文章、问题、候选项和外部知识等文本进行分词、小写、词形还原等预处理.

外部知识集合 K 的构建: 由于要给模型输入相关的外部知识, 本文采用如下 3 个步骤挑选得到与 D, Q 和 C 有关的外部知识集合 K. 第 1 步, 从外部知识库中搜索包含 D, Q 和 C 中词语 (去停词) 的知识三元组. 本文使用图结构存储的 ConceptNet[20] 作为外部知识库, ConceptNet 做外部知识库是因为其采用多种资源构建, 包含 OMCS[21], Open Multilingual Word Net[22], OpenCyc[23], DBPedia[24], JMDict[25] 和“Games with a purpose”[26–28]. 知识图中每两个节点和其连接边可组成一个外部知识三元组, 两节点分别是 $subj$ 和 obj, 边为两节点的关系 rel. 第 2 步, 根据以下两点对外部知识进行打分: (1) 外部知识中的词在给定的不同文本中出现的重要性不同, 在候选项中出现最重要, 问题其次, 文章更次; (2) 外部知识三元组里出现在给定文本中的词数在外部知识总词数的占比, 并排序选出得分最高的前 50 条知识. 其中一条外部知识三元组 k 的相关性得分的计算如式 (10) 和式 (11):

$$score_k = (score_{subj} + score_{obj}) * weight \tag{10}$$

$$weight = \frac{cnt((subj \cup obj) \cap (D \cup Q \cup C))}{cnt(subj) + cnt(obj)} \tag{11}$$

式中, $score_{subj}$ 为 $subj$ 分数, 有 4 个值, 取值方式为: $subj$ 中的词如果出现在选项中给 4 分, 如果出现在问题中为 3 分, 如果在文章中给 2 分, 不出现给 0 分. $score_{obj}$ 类似. 式 (11) 中 $cnt()$ 为词语数量, $\cup$ 表示两者中出现过的所有词, $\cap$ 表示两者中共同出现的词. 第 3 步, 为防止外部知识出现频率对模型选答案造成影响, 使两条选项所对应的外部知识数量一致, 本文首先把选项对应的外部知识进行排序, 保留所有选项前 x 条外部知识 (x 为对应外部知识最少的选项的外部知识数量). 所有只与文章或与问题有关的外部知识则都保留.

2.2 模型参数配置

本文提出的模型采用 Pytorch 框架构建实现. 词嵌入为 300 维预训练 Glove 词向量[15] 和 8 维人工特征拼接的 308 维向量. MCScript 数据集中每道题对应 2 个选项, 人工特征为 8 维, 即是否出现在文章 D, 问题 Q, 选项 c_1, 选项 c_2, 同时出现在文章 D 和候选项 c_1, 同时出现在文章 D 和候选项 c_2, 问题 Q 和候选项中 c_1, 问题 Q 和候选项 c_2 中. 由于外部知识三元组中 rel 是外部知识库自己定义的字符表示并且种类固定 (ConceptNet 中为 34 种), 本文采用 34 个 308 维随机初始化向量分别表示这些关系. BiLSTM 的隐层是 150 维, 损失函数是 NLLLoss. 采用 Adamax 算法[29] 以 0.002 为初始学习率开始训练, 并在每 10 个批大小 (batch size) 进行更新. 为防止过拟合, 嵌入层和编码层都采用概率 0.4 的丢失策略 (dropout). 答案层采用 8 层的 BiLSTM 进行 8 步推理, 答案层随机丢失的概率也为 0.4. 模型共训练 50 轮 (epoch).

2.3 模型评估

本文采用准确率来评估模型, 如式 (12) 计算.

$$Accuracy = \frac{\text{某类问题答对的数目}}{\text{某类问题总数目}} \tag{12}$$

3 实验分析

3.1 与其他模型的比较

表 3 展示了本文模型与其他模型的实验对比结果.

其中“—”表示数据未提供, “*”表示是使用原作者提供的代码复现的结果, 除了 MITRE 均为单模型结果.

表 3 中 Random, Sliding Window, Bilinear Model 和 Attentive Reader 是 MCScript 上的实验基线系统. Co-Matching[10]是本文模型的基线系统. HMA[30], TriAN[12]是竞赛前两名的单模型结果. MITRE[31]是竞赛第 3 名的集成模型的结果, 因为其使用不同结构的模型集成, 并非同一结构不同随机种子训练的模型集成, 因此将其 3 个集成模型 MITRE(NN-T), MITRE (NN-GN) 和 MITRE(LR) 的结果也列在表中. 其中 GCN[13]和 TriAN[12]是隐式引入外部知识到模型词嵌入层中的模型, 而本文模型是在编码之后显示引入外部知识的模型, 其余模型都未引入外部知识.

表 3 不同模型的准确率对比 (单位: %)

模型	all	text	cs
人类水平[7]	98.20	—	
Random[7]	50.00	50.00	50.00
Sliding Window[7]	55.00	55.70	53.10
Bilinear Model[7]	70.20	69.80	71.40
Attentive Reader[7]	72.00	70.90	75.20
GCN[14]	78.97	—	—
Co-Matching*[10]	80.01	81.21	77.03
HMA[30]	80.94	—	—
TriAN(扩展训练数据)[13]	**81.94**	—	—
TriAN(未扩展训练数据)[13]	80.51	—	—
MITRE(集成模型)[31]	**82.27**	**83.00**	79.00
MITRE(NN-T)[31]	80.23	—	—
MITRE(NN-GN)[31]	80.12	—	—
MITRE(LR)[31]	79.66	—	—
本文模型	**81.01**	**81.53**	**79.95**

本文模型在 MCScrip 中需要外部知识辅助的问题 (“cs”类型的问题) 上取得了 79.95% 的准确率, 超过 Co-Matching 的准确率 2.92%, 甚至超过了 MITRE 集成模型在“cs”类型的问题上 0.95% 的准确率. 而仅由文章推理可得答案的问题 (“text”类型的问题) 取得了 81.53% 的准确率, 相比于 Co-Matching 的准确率提升 0.32%, 说明本文模型在引入外部知识时保持“text”类型的问题性能不降低. 从总体性能和问题占比情况看, 本文模型取得了 81.01% 的准确率, 相比 Co-Matching 提升了 1%, 这主要是靠在数据集中相对数量少的、占比约 30% 的“cs”类型问题近 3% 的提升所贡献的. 该实验结果与本文目标通过提升需外部知识辅助问题的性能, 从而帮助选择题型阅读理解是一致的. 另外, 与表 3 中总体性能 (“all”问题的准确率) 超过本文模型的集成模型 MITRE 和单模型 TriAN(扩展训练数据) 比较, MITRE 效果好于本文模型是因为其使用 3 个子模型集成, 然而其 3 个子模型各自的在“all”问题的准确率均低于本文单模型效果; 与单模型 TriAN(扩展训练数据) 比较, 本文模型准确率比 TriAN 低 0.93%, 这是因为 TriAN 使用了额外的大规模单选题数据集 RACE 做训练, 而本文未使用额外训练数据. 当 TriAN 只用 MCScript 进行训练时结果为 80.51%, 本文模型的结果高于其结果 0.50%.

3.2 不同外部知识数量的影响

为了深入分析不同外部知识数量对模型的影响, 表 4 展示了不同外部知识的结果. 当知识数量为 100 条时, “cs”类问题的效果最好, 但总体性能相比 50 条时略有下降. 整体观察表格中数据情况, 可以发现当外部知识数量越高, 模型的效果整体呈降低趋势, 这可能是由于随着外部知识的引入同时会带来噪声, 说明构建外部知识集合 K 的方式、外部知识被引进模型及与文章问题选项三者交互的方式仍可以被改进.

表 4 不同外部知识数量情况下 test 上的准确率 (单位: %)

#外部知识	all	text	cs
50	81.01	81.53	79.95
100	80.91	81.19	80.07
150	80.86	81.61	79.01
200	80.77	81.38	78.97

3.3 模型模块消减实验

本文还进行了消减实验用以分析外部知识的引入以及利用外部知识和文章信息一起进行推理的作用. 首先本文模型分别与如下 3 种模型对比: (1)“-外部知识引入”是指仅采用上下文编码信息进行匹配和多步推理, 仅在本文模型上删去外部知识引入层; (2)“-答案多步推理”是在本文模型上删去多步推理答案层, 仅使用 Co-Matching 模型原本的答案层, 但保留了外部知识的引入; (3)“-外部知识引入&答案多步推理”是在本文模型中删除外部知识引入和多步推理两种模块. 表 5 给出了消减实验的结果, “-”表示删减某个模块.

对比表 5 中“本文模型”和“-外部知识引入”发现删除外部知识的辅助在“cs”类型问题上有 1.94% 的下降, 并且表中包含外部知识引入的模型 (“本文模型”和“-答案多步推理”) 在“cs”类型问题的结果都高于其余两种

无外部知识辅助模型的效果, 证明加入外部知识信息是有助于回答需要外部知识辅助类型问题的.

对比表 5 中“本文模型”和“-答案多步推理”, 发现在“cs”类型问题的表现上下降 0.94%, 说明在语义匹配的基础上进一步利用外部信息推理是对答题有帮助的. “-答案多步推理”和“-外部知识引入”对比, 发现删去多步推理机制比删去外部知识引入对“text”类型的问题影响更大. 对比“-答案多步推理”和“-外部知识引入&答案多步推理”, 发现在“text”类型的问题上“-答案多步推理”要略逊 0.10%, 说明引入外部知识信息后若仅采用语义匹配, 不增加推理步骤可能会对“text”类型的回答略有影响, 也证明了利用外部知识进行推理 (而不仅仅是语义匹配) 在使模型尽可能不降低“text”类型的效果上略有帮助.

表 5 test 上的消减实验准确率 (单位: %)

模型	all	text	cs
本文模型	81.01	81.53	79.95
-外部知识引入	80.41	81.24	78.01
-答案多步推理	80.44	81.11	79.01
-外部知识引入&答案多步推理	80.01	81.21	77.03

3.4 模型运行时间分析

表 6 中为本文模型和基线模型 Co-Matching 每一轮 (epoch) 平均训练时间、训练至目标函数的损失值波动平稳的收敛时间和全部测试数据的测试时间. 本文模型对基线模型 Co-Matching 的各层都增加了功能, 时间的主要增长来源应是额外增添的外部知识融入层和多步推理答案层, 因此训练时间、收敛时间和测试时间都有所增加, 但增加的时间在合理可控的范围内.

表 6 每轮的运行时间分析 (单位: s)

模型	训练时间	收敛时间	测试时间
本文模型	196	9800	20
Co-Matching	158	7900	13

4 结论与展望

针对需要外部知识辅助的单项选择题阅读理解问题, 本文提出了一个利用外部知识辅助和多步推理的选择题型机器阅读理解模型, 这个模型在保证引入外部知识的情况下, 不降低回答仅需文章推理的问题性能的同时, 提高需要外部知识辅助问题的性能, 从而提高模型的整体性能. 本文的实验结果证明提出的模型的有效性, 然而对比人类的结果还有巨大的差距, 因此接下来将更深入分析人类回答两类问题的特点, 模拟人类的回答方式探索更细致有效的外部知识引入模型的方式, 研究不同外部知识引入方式对模型性能的影响.

参考文献

1 Hermann KM, Kočiský T, Grefenstette E, *et al*. Teaching machines to read and comprehend. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, QC, Canada. 2015. 1693–1701.

2 Hill F, Bordes A, Chopra S, *et al*. The goldilocks principle: Reading children's books with explicit memory representations. arXiv preprint arXiv:1511.02301, 2015.

3 Rajpurkar P, Zhang J, Lopyrev K, *et al*. SQuAD: 100,000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 2383–2392.

4 Joshi M, Choi E, Weld D, *et al*. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada. 2017. 1601–1611.

5 Richardson M, Burges CJC, Renshaw E. MCTest: A challenge dataset for the open-domain machine comprehension of text. Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing. Seattle, WC, USA. 2013. 193–203.

6 Lai GK, Xie QZ, Liu HX, *et al*. RACE: Large-scale ReAding comprehension dataset from examinations. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark. 2017. 785–794.

7 Ostermann S, Modi A, Roth M, *et al*. MCScript: A novel dataset for assessing machine comprehension using script knowledge. Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Miyazaki, Japan. 2018. 3567–3574.

8 Parikh S, Sai AB, Nema P, *et al*. ElimiNet: A model for eliminating options for reading comprehension with multiple choice questions. Proceedings of the 27th International Joint Conference on Artificial Intelligence. Stockholm, Sweden. 2018. 4272–4278.

9 Zhu HC, Wei FR, Qin B, *et al*. Hierarchical attention flow for multiple-choice reading comprehension. Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. New Orleans, LA, USA. 2018. 6077–6084.

10 Wang SH, Yu M, Jiang J, *et al*. A co-matching model for multi-choice reading comprehension. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia. 2018. 746–751.

11 Dhingra B, Liu HX, Yang ZL, *et al*. Gated-attention readers for text comprehension. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, BC, Canada. 2017. 1832–1846.

12 Xu YC, Liu JJ, Gao JF, *et al*. Dynamic fusion networks for machine reading comprehension. arXiv preprint arXiv:1711.04964, 2017.

13 Wang L, Sun M, Zhao W, *et al*. Yuanfudao at SemEval-2018 Task 11: Three-way attention and relational knowledge for commonsense machine comprehension. Proceedings of the 12th International Workshop on Semantic Evaluation. New Orleans, LA, USA. 2018. 758–762.

14 Chen WY, Quan XJ, Chen CB. Gated convolutional networks for commonsense machine comprehension. Proceedings of the 25th International Conference on Neural Information Processing. Siem Reap, Cambodia. 2018. 297–306.

15 Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1532–1543.

16 Mihaylov T, Frank A. Knowledgeable reader: Enhancing cloze-style reading comprehension with external commonsense knowledge. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia. 2018. 821–832.

17 Tai KS, Socher R, Manning CD. Improved semantic representations from tree-structured long short-term memory networks. Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Beijing, China. 2015. 1556–1566.

18 Wang SH, Jiang J. A compare-aggregate model for matching text sequences. arXiv preprint arXiv:1611.01747, 2016.

19 Liu XD, Shen YL, Duh K, *et al*. Stochastic answer networks for machine reading comprehension. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia. 2018. 1694–1704.

20 Speer R, Chin J, Havasi C. ConceptNet 5.5: An open multilingual graph of general knowledge. Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. San Francisco, CA, USA. 2017. 4444–4451.

21 Singh P, Lin T, Mueller ET, *et al*. Open mind common sense: Knowledge acquisition from the general public. Proceedings of OTM Confederated International Conferences on the Move to Meaningful Internet Systems. Irvine, CA, USA. 2002. 1223–1237.

22 Bond F, Foster R. Linking and extending an open multilingual WordNet. Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia, Bulgaria. 2013. 1352–1362.

23 Elkan C, Greiner R. Building large knowledge-based systems: Representation and inference in the CYC project: D.B. Lenat and R.V. Guha. Artificial Intelligence, 1993, 61(1): 41–52. [doi: 10.1016/0004-3702(93)90092-P]

24 Auer S, Bizer C, Kobilarov G, *et al*. Dbpedia: A nucleus for a Web of open data. In: Aberer K, Choi KS, Noy N, *et al*., eds. The Semantic Web. Berlin, Heidelberg: Springer, 2007. 722–735.

25 Breen J. JMDict: A Japanese-multilingual dictionary. Proceedings of the Workshop on Multilingual Linguistic Resources. Geneva, Switzerland. 2004. 65–72.

26 Von Ahn L, Kedia M, Blum M. Verbosity: A game for collecting common-sense facts. Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Montréal, Québec, Canada. 2006. 75–78.

27 Kuo YL, Lee JC, Chiang KY, *et al*. Community-based game design: Experiments on social games for commonsense data collection. Proceedings of the ACM SIGKDD Workshop on Human Computation. Paris, France. 2009. 15–22.

28 Nakahara K, Yamada S. Development and evaluation of a web-based game for common-sense knowledge acquisition in Japan. Unisys Technology Review, 2011, 30(4): 295–305.

29 Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980, 2014.

30 Chen ZP, Cui YM, Ma WT, *et al*. HFL-RC system at SemEval-2018 Task 11: Hybrid multi-aspects model for commonsense reading comprehension. arXiv preprint arXiv:1803.05655, 2018.

31 Merkhofer E, Henderson J, Bloom D, *et al*. MITRE at SemEval-2018 Task 11: Commonsense reasoning without commonsense knowledge. Proceedings of the 12th International Workshop on Semantic Evaluation. New Orleans, LA, USA. 2018. 1078–1082.