基于机器阅读理解的论文辅助阅读系统构建

秘蓉新[1]，姚文文[1]，阮宏坤[2]
1. 国家计算机网络应急技术处理协调中心，北京 100029；
2. 北京邮电大学计算机学院，北京 100876

摘要
在信息化和数字化时代，科技论文数量的迅速增加带来了一系列问题，如论文冗长、信息提取困难、阅读时间成本居高不下等，研究者面临着更加烦琐、耗时的文献阅读挑战。通过语言模型落地创新，设计了科技论文辅助阅读系统来应对这些挑战。以机器阅读理解技术为核心，通过解析论文文本和预先设定问题，达到自动回答的效果。充分利用预训练语言模型PERT，提升系统对语义的理解和信息的提取能力，解决科技论文阅读过程中存在的各种问题，从而帮助读者提高科技文献阅读效率。

关键词
自然语言处理；机器阅读理解；预训练语言模型

中图分类号：TP391.1 文献标志码：A doi: 10.11959/j.issn.2096-0271.2024039

Construction of a paper-assistant reading system based on machine reading comprehension

MI Rongxin[1], YAO Wenwen[1], RUAN Hongkun[2]
1. National Computer Network Emergency Response Technical Team/Coordination Center of China, Beijing 100029, China
2. Faculty of Computer Science, Beijing University of Posts and Telecommunications, Beijing 100876, China

Abstract
In the era of informatization and digitization, the rapid increase in the number of scientific papers has given rise to various challenges, such as lengthy articles, difficulty in information extraction and high time costs associated with reading. Literature reading challenges for researchers are increasingly tedious and time-consuming. By utilizing the language models, the assited reading system of scientific papers has been designed to address these challenges. By adopting machine reading comprehension technology as the core, the system parses scientific texts and offers some common questions to achieve automated response capabilities. By fully utilizing the pre-trained language model PERT, the system enhances

its capabilities in semantic understanding and information extraction, effectively resolving various challenges in reading scientific papers and helping readers improve the efficiency of scientific literature review.

Key words
natural language processing, machine reading comprehension, pre-trained language model

0 引言

近年来，科技的飞速发展使科技论文的数量急速增长。当全世界的科研工作者将自己的研究成果发布在各种各样的期刊和论文库中时，构筑的论文体系将更加庞大。大量的论文公开散布在网络上，为科研人员提供了相对充足的参考资料，同时也暴露了以下问题。

- 部分科技论文的组织较为混乱且分散，导致许多科研人员在查找论文时存在一定的困难。一些论文内容冗余，篇幅较长，研究人员要从论文中精准快速地抽取所需信息的难度和时间成本越来越高。
- 科研人员对自己要查询的内容模糊不清，在阅读和筛选论文时浪费了较多的时间和精力。

上述问题产生了许多需求场景，比如在论文检索时，不单单考虑论文的客观信息，也可以尝试基于文章在一些常见问题的回答方面返回检索结果，并且在阅读文章时，读者往往会因为内容较为晦涩难懂而产生想要快速定位到核心内容、解决内心困惑的想法。目前已知的一些论文检索和阅读工具更注重论文的收藏量、阅读标注等与阅读习惯相关的功能和文章翻译等需求，而忽略了基于文章内容进行理解的应用场景，往往这部分需求更能有效提高使用者的工作效率。

如今，自然语言处理领域的机器阅读理解（machine reading comprehension，MRC）技术不断发展，各种大规模基准数据集的出现也使深度神经网络能够实现机器阅读理解，预训练语言模型（pre-trained language model，PLM）在阅读理解任务上表现出色。论文辅助阅读系统的目标是帮助研究人员快速获取所需信息，而MRC任务的引入使系统能够直接回答用户提出的问题，实现更高效的信息检索和阅读，提高了系统信息检索的精度和准确性，节省了用户的时间和精力。此外，基于MRC的系统可以根据用户提出的具体问题提供回答，为用户提供更好的阅读体验。MRC任务为论文辅助阅读系统的发展提供了重要支持，使系统能够更好地满足用户的信息需求，提高了系统的效率和用户体验。因此，构建能够理解论文内容并回答使用者问题的科技论文辅助阅读系统具有一定的研究意义。本文的主要成果如下。

- 归纳目前机器阅读理解领域常用的预训练模型，选择贴合应用场景的模型算法。将深度神经网络与实际应用相结合，为机器阅读理解理论研究的应用实践方式提供参考。
- 设计、开发具有论文库管理、论文检索和论文机器预阅读与问答等功能的应用系统。该系统能够对网络中繁杂的资料进行筛选，其问答功能有助于科研人员快速抽取出问题的答案。

1 相关技术

1.1 机器阅读理解技术

使用计算机理解自然语言文本并回答

相关问题，是自然语言处理领域具有较强挑战性和较大难度的经典任务之一，得到完全准确的结果也是相关领域不断追求的目标[1]。机器阅读理解有很悠久的研究历史，但是由于数据集规模以及计算机硬件设备的制约，直到最近几年才重新受到国内外科研人员的广泛关注。MRC在近期快速发展最重要的两个原因为：①百科类知识库以及众包群体智慧服务模式的发展，极大地推动了大规模的阅读理解监督数据集的创建，这类数据集以（段落，问题，答案）三元组的形式存储，为MRC模型的性能测评以及准确率的提高提供了可能性[2]；②计算机硬件设备性能的显著提升推动了以神经网络为代表的深度学习技术的全面发展。

机器学习任务从形式和数据集等方面可分为四大类，即完形填空式（cloze tests）、多项选择式（multiple choice）、抽取式（也称跨距预测类型，span extraction）和自由回答式（free answering）[3]。

- 完形填空式：通过从文章中删除一些单词或实体产生问题。完形填空式给阅读增加了障碍，需要理解上下文和词汇的使用。
- 多项选择式：要求根据提供的上下文选择正确的答案。多项选择题的答案并不仅仅局限于上下文中的单词或实体，具有更加灵活多样的答案形式。
- 抽取式：给定上下文和问题，要求机器从相应的上下文中提取一段文本作为答案。
- 自由回答式：需要机器对文本的多个片段进行推理，并总结证据。

在上述4个任务中，自由回答式是最复杂的，因为它的答题形式没有限制，对于辅助阅读需求来说，过多的生成内容不能保证科技论文阅读的准确性和严谨性。

按照人类阅读习惯，多数MRC模型架构可分为4个模块，如**图1**所示，以给定的原文内容和问题为输入，输出预测答案。MRC模型通过嵌入编码将文本转换为向量表示，通过特征提取来提取有用的特征信息，通过信息交互进行推理和理解，最后通过答案预测生成最终的答案。

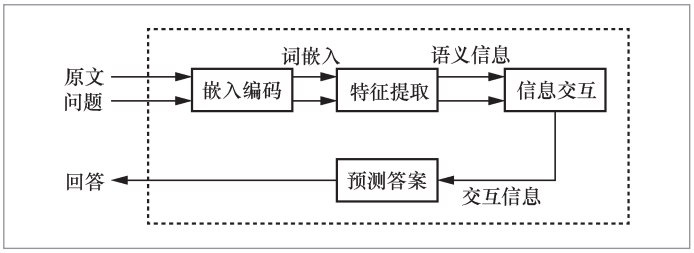

图1 MRC模型基本架构[4]

1.2 经典预训练模型

机器阅读理解任务在以往研究中主要使用统计学的方法来完成，在SQuAD（Stanford question answering dataset）发布之后，出现了一些基于注意力机制的匹配模型。此后，BERT（bidirectional encoder representation transformers）于2018年由Google AI团队发布[5]，在各主流任务上的表现取得了质的飞跃，极大地推动了自然语言处理领域的发展。此后几年陆续出现很多围绕着BERT进行优化的模型，例如RoBERTa[6]、ALBERT[7]以及MacBERT[8]等，这类预训练模型通过微调解决基础模型使用的问题。此外，GPT（generative pre-trained transformer）[9-11]作为当下火热的大语言预训练模型，在各项自然语言处理任务上表现优异，获得各界的广泛关注。随着当前技术的发展，规模更大、效果更好的大语言模型也不断涌现。

BERT的核心部分采用了Transformer架构，是一个多层Transformer的编码器（encoder），输入的词通过嵌入逐层的编

码器进行编码转换，再连接到不同的下游任务。BERT采用了双向并行输入的方式，即将整个句子输入模型中，而不是逐个输入单词，大大提升了模型的运行效率。BERT通过两个无监督任务进行参数预训练，一个是遮蔽语言模型，这种训练流程可以让模型学习到单词在上下文中的分布表示，另一个是连续句子预测任务，让模型理解两个句子之间的关系，从而应用于机器阅读理解以及自然语言推理等下游任务中。

ALBERT是一个轻量级的BERT模型，其最大的特点是参数少、速度快。ALBERT结合了两种参数的约简技术：一是对嵌入参数进行因式分解，通过将大的词汇表嵌入矩阵分解为两个小的矩阵，将隐藏层的大小与词汇表嵌入的大小分离开来；二是跨层参数共享，这种技术可以防止参数数量随着网络深度的增加而增加。

RoBERTa在更大的数据集上对模型进行了更多的迭代训练。RoBERTa去掉了BERT训练中的下一序列预测（next sentence prediction，NSP），通过更大的数据集和更长时间的迭代来提高模型的表现，从而在多个自然语言处理任务上表现出更好的性能。

GPT是OpenAI团队开发的一种大型自回归语言模型，利用连续的数据预测来进行模型参数的优化。在GPT-1的实现中，首先在无标注的数据上进行无监督预训练，然后针对具体的任务，比如文本分类和问答，进行有监督的微调。GPT-2基于GPT-1的架构，扩大了模型规模和训练数据集，增强了模型的泛化能力。GPT-3在模型容量上做了进一步的提升，并通过上下文学习，在许多零样本或小样本的任务上表现出先进的性能，甚至在某些任务上超过了经过微调的模型。ChatGPT基于GPT-3.5的架构，通过结合人类反馈的强化学习技术，在对话问答方面表现出了高性能[12]。

1.3 PERT模型选择

目前，预训练语言模型在各种自然语言处理任务中均表现出了优异的性能，主要分为自编码和自回归两种。自编码PLM不但能使用掩码语言模型任务作为预训练任务，而且在一段文本中随机打乱几个字不会影响人们对这一段文本的理解。因此本文基于乱序语言模型的预训练模型PERT[13]构建论文辅助阅读系统。

PERT模型的基本输入、输出格式如**图2**所示。PERT遵循与BERT相同的范式，因此对各种下游任务进行微调时，PERT可以直接用于BERT的任何调优脚本。要注意的是，PERT模型在微调阶段使用自然输入序列，而不是改变单词顺序的序列。

中文语料数据来自维基百科等。对比发现，PERT模型在MRC任务上表现较好。PERT模型相比MacBERT[8]模型在部分精度上略有提高，在执行中文MRC任务时能很好地体现该模型的优点。各常见模型在中文机器阅读理任务上的结果见**表1**，选用的数据集为CMRC 2018（简体中文）[14]和DRCD（繁体中文）[15]。

英文语料使用英语维基百科等数据作为预训练数据，使用与BERT中相同的WordPiece[16]分词器和词汇量为30 522的英文BERT-Base-Uncased词汇表。与中文自然语言理解任务的结果类似，$PERT_{base}$在MRC任务上表现出很好的性能。然而，$PERT_{large}$的效果并未达到最好，但是与其他模型的差距并不显著。

目前，能够进行机器阅读理解任务的模型在英文预测中的表现明显好于在中文预测上的表现。然而，在前科技论文中，中文文献的比例也在不断增加[17]。经过综合考虑，由于PERT模型能够较好地适配

所需的功能，本文最终选择使用该模型执行系统的相关任务。与更大规模的模型相比，该模型具有更轻量、更高效、更容易部署等特点，适用于需要实时推理的应用场景。此外，PERT模型具有更好的泛化能力，更容易解释和理解，并且在成本和资源效率方面具有优势。

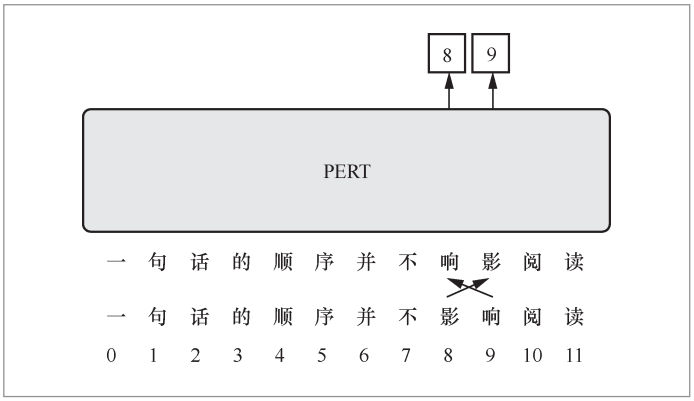

图 2 PERT 模型的基本输入、输出格式 [13]

2 论文辅助阅读系统构建实践

2.1 系统需求

从实际场景出发，论文辅助阅读系统的主要功能需求分析如下。

● 支持多种检索关键字，如论文的基本信息与基于文章自动回答的问题结果。支持在线阅读和阅读时的问题解答，将机器阅读理解模型部署在应用程序中。

● 支持文章收藏，对于用户感兴趣的文章，该系统支持重复阅读和查看，支持用户构建自己的收藏夹。支持论文文件自主上传，由于论文初始数量有限，该系统需要有“边使用边扩充”的能力，在使用过程中，对数据库的内容存量进行提升。

● 问答环节支持自主选择“段落”和“问题”，除了允许用户自由手动输入内容外，还可以根据数据库中事先存储的对论文解析得到的“摘要”“相关工作”“结论”等章节进行提问，加快用户的段落选择速度。同时也可以准备多个在阅读文献过程中较常出现的问题作为备选，从而加快问答环节互动节奏。

2.2 系统架构设计

科技论文辅助阅读系统总体架构设计如**图3**所示，设计分为3层，每层的功能描述如下。

● 交互表示层：负责交互任务，由其他层次相应模块提供支持。

● 业务处理层：负责全部运算过程，主

表 1 各模型在中文阅读理解任务上的实验结果 [8]

系统	CMRC 2018				DRCD			
	Dev set		Test set		Dev set		Test set	
	EM	F1	EM	F1	EM	F1	EM	F1
$BERT_{base}$	65.6%	85.0%	70.0%	87.0%	84.5%	90.9%	83.0%	89.9%
$RoBERT_{base}$	66.5%	86.5%	71.4%	88.8%	85.9%	92.2%	85.2%	91.7%
$ELECTRA_{base}$	68.0%	84.6%	72.7%	86.9%	87.0%	92.3%	86.6%	91.7%
$MacBERT_{base}$	67.3%	87.1%	72.4%	89.2%	89.2%	94.1%	88.7%	93.5%
$PERT_{base}$	68.1%	87.1%	72.5%	89.0%	88.9%	93.6%	88.5%	93.2%
$RoBERT_{large}$	67.6%	87.9%	72.4%	90.0%	89.1%	94.4%	88.9%	94.1%
$ELECTRA_{large}$	68.2%	84.5%	72.8%	86.6%	88.7%	93.2%	88.2%	93.2%
$MacBERT_{large}$	68.6%	88.2%	73.2%	90.1%	90.8%	95.3%	90.9%	95.3%
$PERT_{large}$	71.0%	88.8%	75.5%	90.4%	90.8%	95.2%	90.7%	95.1%

要包含消息的格式控制、PDF论文结构拆解、预先自动阅读、论文管理、检索计算、问答环节。这些运行模块分别与相应的页面对应，数据操作由数据库提供，其中问答环节和论文自动阅读部分通过部署的MRC模型辅助实现。

● 数据存储层：负责存储系统所需的全部数据，且有专用的论文存储位置用于论文库的构建。

该系统的功能模块如下，主要功能结构如**图4**所示。

（1）用户管理模块

用户管理模块负责用户的注册与登录。

（2）论文管理模块

● 内容解析与拆分：将可用程序拆分的论文进行内容解析。

● 自动阅读：对解析后的论文内容进行自动阅读，回答预先设定的问题。

● 信息添加：将论文的信息、拆分内容、问题答案等全部信息存储到数据库。

● 原文存储：将原文以文件格式存储至论文库中。

● 论文移除：后台管理人员通过操作数据库对要移除的论文进行删除操作。

（3）收藏管理模块

收藏管理模块包含用户收藏感兴趣的论文和从收藏夹中移除论文。

（4）检索模块

检索模块包含关键字检索、作者检索、主题检索、研究方法检索、解决问题检索等，可将检索结果呈现给用户。

（5）阅读模块

阅读模块包含文章预览和自动问答，自动问答环节为自动阅读过程的拓展，可以根据用户实时操作进行自动阅读。

2.3 系统功能实现

论文添加需要进行鉴权处理。为防止论文库被垃圾文件污染，选择PDF格式文

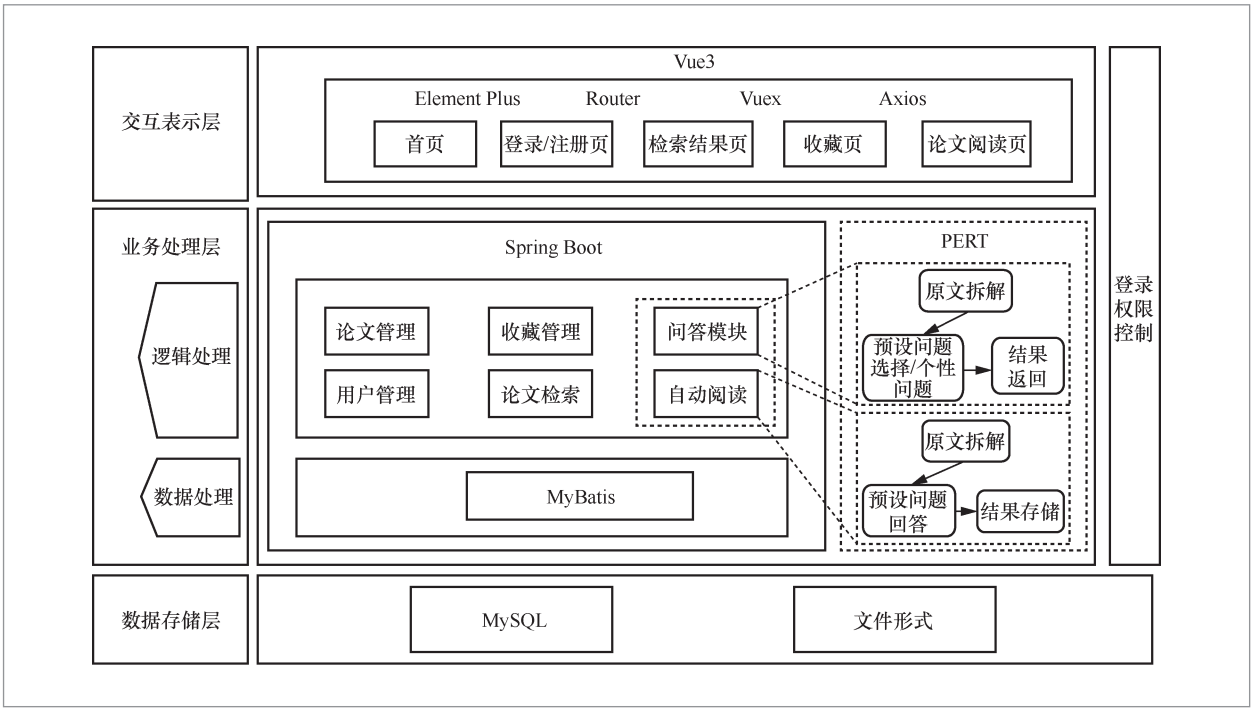

图 3 总体架构设计

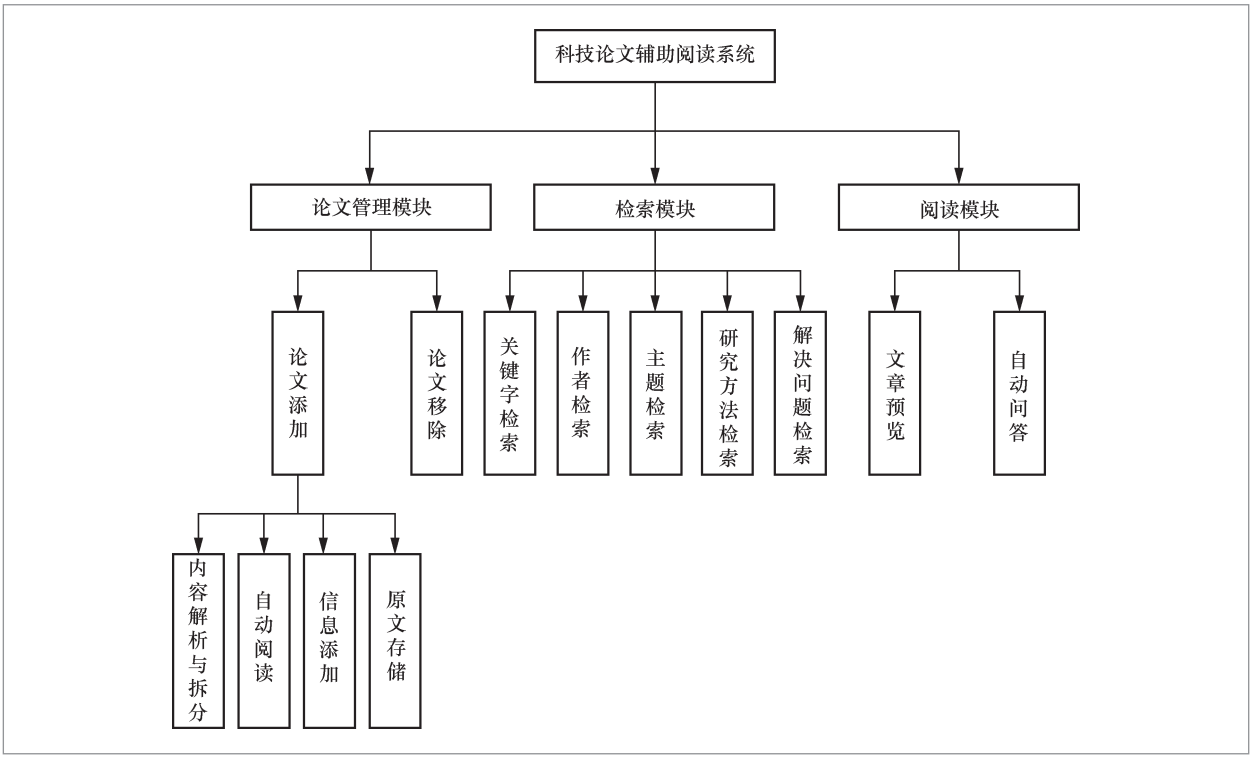

图 4 系统主要功能结构

件进行上传扩充论文库。在上传文件的同时，可以输入文章标题和作者信息。文件上传至后台后，需要对论文进行处理，包括结构拆解处理和自动阅读处理，经过处理后，将所有得到的内容存储至数据库的论文相关表，将论文源文件统一存储。

处理论文文件，使用已有代码库解析PDF文件，通过该方式对文件进行逐步解析，然后将文本内容合并。将解析的内容存储到数据库中以供检索。检索模块通过待检索内容和不同检索模式需求，检索得到符合要求的论文。

选定基础机器阅读理解模型PERT，并基于该系统特定任务进行微调。针对不同语言类型，分别调用使用中文语料和英文语料预训练的两个结构相同但训练数据不同的模型。PERT模型的开源版本包含Transformer部分的权重，对各个下游任务进行微调，与BERT共享相同的主要神经架构。下载多篇会议、期刊的科技文章作为语料，并对其进行提问与答案标注，构建小型专业知识类数据集，对其中英文模型分别进行微调以贴合应用场景。

在对论文各模块进行划分后，对这些模块进行预先自动阅读处理。按章节模块处理的同时，考虑到过长文本直接作为输入的不合理性，采取“滑动窗口”的方式，对长文本内容进行截断，改变输入的结构从而优化答案。基于论文拆分后的结果，对论文正文部分进行滑动窗口式截断处理，对每次窗口划分的内容进行机器阅读理解。最后，需要综合各个章节的答案，对所有的回答进行综合考虑，选择更具有代表性的回答作为最终答案。该模型还用于问答环节对阅读过程进行辅助，可通过提问来快速了解圈定内容，关于提出问题和划定内容，有预先准备的范式以供选择，也可进行个性化提问。

3 结束语

本文基于机器阅读理解构建的论文辅助阅读系统基本满足阅读和检索需求，未来研究可从以下方面进行优化。

● 提高系统鲁棒性，扩充系统功能，收集存储更多的论文数据，为专业人员提供支持。

● 目前模型在MRC任务中的表现还有待提高，不断提升答案的准确度才能满足更严谨和更专业的数据理解任务的需求，且在论文解析阶段还需要改进方法，增强系统的自动化程度，降低人工检查环节比重。

● 考虑科技论文的严谨性，当前模型主要实现了原文内容的抽取式回答，随着机器阅读理解在生成式任务和跨文本阅读任务上表现的提升，该系统将在未来更好地支持复杂推理和多步骤问题解答，进而提高答案的深度和准确性。

● 随着大语言模型的不断发展，该系统也可以紧跟技术更新，融合不同的模型，提升用户体验。

参考文献:

[1] HERMANN K M, KOČISKÝ T, GREFENSTETTE E, et al. Teaching machines to read and comprehend[J]// Advances in Neural Information Processing Systems. 2015(1): 1693–1701.

[2] 顾迎捷, 桂小林, 李德福, 等. 基于神经网络的机器阅读理解综述[J]. 软件学报, 2020, 31(7): 2095–2126.
GU Y J, GUI X L, LI D F, et al. Survey of machine reading comprehension based on neural network[J]. Journal of Software, 2020, 31(7): 2095–2126.

[3] LIU S S, ZHANG X, ZHANG S, et al. Neural machine reading comprehension: methods and trends[J]. Applied Sciences, 2019, 9(18): 3698.

[4] 张少华. 面向复杂文本的抽取式机器阅读理解研究[D]. 荆州: 长江大学, 2023.
ZHANG S H. Research on extractive machine reading comprehension for complex textual corpus[D]. Jingzhou: Yangtze University, 2023.

[5] DEVLIN J, CHANG M W, LEE K, et al. Bert: pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the 2019 Conference of the North, Minneapolisy. Minnesotd: Association for Computational Linguistics, 2019.

[6] LIU Y, OTT M, GOYAL N, et al. Roberta: a robustly optimized Bert pretraining approach[EB]. arXiv preprint, 2019, arXiv: 1907.11692.

[7] LAN Z, CHEN M, GOODMAN S, et al. Albert: a lite Bert for self-supervised learning of language representations[EB]. arXiv preprint, 2019, arXiv: 1909.11942.

[8] CUI Y, CHE W, LIU T, et al. Revisiting pre-trained models for Chinese natural language processing[C]//Proceedings of Findings of the Association for Computational Linguistics: EMNLP 2020. Stroudsburg: ACL, 2020: 657–668.

[9] RADFORD A, NARASIMHAN K, SALIMANS T, et al. Improving language understanding by generative pre-training[EB]. arXiv preprint, 2018, arXiv: 49313245.

[10] RADFORD A, WU J, CHILD R, et al. Language models are unsupervised multitask learners[J]. OpenAI Blog, 2019, 1(8): 9.

[11] BROWN T B, MANN B, RYDER N, et al. Language models are few-shot learners[EB]. arXiv preprint, 2020, arXiv: 2005.14165.

[12] 卢经纬, 郭超, 戴星原, 等. 问答ChatGPT之

后: 超大预训练模型的机遇和挑战[J]. 自动化学报, 2023, 49(4): 705-717.
LU J W, GUO C, DAI X Y, et al. The ChatGPT after: opportunities and challenges of very large scale pre-trained models[J]. Acta Automatica Sinica, 2023, 49(4): 705-717.

[13] CUI Y, YANG Z, LIU T, Pert: pre-training Bert with permuted language model[EB]. arXiv preprint, 2022, arXiv: 2203.06906.

[14] CUI Y, LIU T, CHE W, et al. A span-extraction dataset for Chinese machine reading comprehension[C]//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Stroudsburg: ACL, 2019: 5883-5889.

[15] SHAO C C, LIU T, LAl Y T, et al. Drcd: a Chinese machine reading comprehension dataset[EB]. arXiv preprint, 2018, arXiv: 1806.00920.

[16] WU Y, SCHUSTER M, CHEN Z, et al. Google's neural machine translation system: bridging the gap between human and machine translation[EB]. arXiv preprint, 2016, arXiv: 1609.08144.

[17] 万小军. 智能文本生成: 进展与挑战[J]. 大数据, 2023, 9(2): 99-109.
WAN X J. Intelligent text generation: progress and challenges[J]. Big Data Research, 2023, 9(2): 99-109.

作者简介

秘蓉新(1984-), 女, 国家计算机网络应急技术处理协调中心助理研究员, 主要研究方向为人工智能、网络安全、信息内容安全。

姚文文(1984-), 女, 国家计算机网络应急技术处理协调中心助理研究员, 主要研究方向为网络安全战略、信息科技发展态势。

阮宏坤(2001-), 男, 北京邮电大学计算机学院硕士生, 主要研究方向为数据预处理、人工智能。

收稿日期: 2024-04-23
通信作者: 姚文文, yww@cert.org.cn