CCL23-Eval 任务9总结报告：汉语高考阅读理解对抗鲁棒评测

郭亚鑫[1]，闫国航[1]，谭红叶[1,2,⋆]，李茹[1,2]
[1]山西大学 计算机与信息技术学院，山西 太原 030006
[2]山西大学 计算智能与中文信息处理教育部重点实验室，山西 太原 030006
{202112407002, 202222407055}@email.sxu.edu.cn
{tanhongye,liru}@sxu.edu.cn

摘要

汉语高考阅读理解对抗鲁棒评测任务致力于提升机器阅读理解模型在复杂、真实对抗环境下的鲁棒性。本次任务设计了四种对抗攻击策略（关键词扰动、推理逻辑扰动、时空属性扰动、因果关系扰动），构建了对抗鲁棒子集GCRC_advRobust。任务需要根据给定的文章和问题从4个选项中选择正确的答案。本次评测受到工业界和学术界的广泛关注，共有29支队伍报名参赛，但由于难度较大，仅有8支队伍提交了结果。有关该任务的所有技术信息，包括系统提交、官方结果以及支持资源和软件的链接，可从任务网站获取[1]。

关键词： 机器阅读理解；鲁棒性；对抗攻击

Overview of CCL23-Eval Task 9: Adversarial Robustness Evaluation for Chinese Gaokao Reading Comprehension

Yaxin Guo[1], Guohang Yan[1], Hongye Tan[1,2,⋆], Ru Li[1,2]
[1]School of Computer and Information Technology, Shanxi University, Taiyuan, Shanxi 030006, China
[2]Key Laboratory of Computational Intelligence and Chinese Information Processing of Ministry of Education, Shanxi University, Taiyuan, Shanxi 030006,China
{202112407002, 202222407055}@email.sxu.edu.cn
{tanhongye,liru}@sxu.edu.cn

Abstract

The Adversarial Robustness Evaluation for Chinese Gaokao Reading Comprehension Task aims to improve the robustness of machine reading comprehension models in complex and realistic adversarial environments. This task includes four types of adversarial attack strategies: Keyword perturbation, Reasoning logic perturbation, Temporal/spatial perturbation, and Cause-effect perturbation. The task constructs an adversarial robust subset called GCRC_advRobust. Participants are required to select the correct answer from four options based on the given passage and questions. A total of 29 teams registered for the competition, but due to the high difficulty, only 8 teams submitted their results. All technical information related to this task, including system submissions, official results, and links to supporting resources and software, can be found on the task website[1].

Keywords: Machine Reading Comprehension , Robustness , Adversarial attack

⋆ 通讯作者 Corresponding Author
[1]http://cuge.baai.ac.cn/#/ccl/2023/gcrc

项目资助： 国家重点研发计划项目(2020AAA0106100)、国家自然科学基金面上项目(62076155)

1 背景和动机

机器阅读理解(Machine Reading Comprehension，MRC) 是指让机器阅读文本，然后回答与文本内容相关的问题。它是自然语言处理和人工智能领域的重要前沿课题，对于提升机器的智能水平、使机器具有持续知识获取的能力等具有重要价值，近年来受到学术界和工业界的广泛关注。

MRC模型的鲁棒性是衡量该技术能否在实际应用中大规模落地的关键(Jia and Liang, 2017)。随着技术的进步，现有模型已经能够在封闭测试集上取得较好的性能，但在面向开放、动态、真实环境下的推理与决策时，其鲁棒性仍表现不佳 (Wu and Xu, 2020; Zhou et al., 2020; Ren et al., 2023; Gan and Ng, 2019)。为了评估模型的鲁棒性，Tang等人(2021)从现有数据库和人类的写作中进行检索并构建分散注意力的问题，创建了中文MRC鲁棒性基准数据集。Wu等人(2020)为了分析影响模型鲁棒性的因素，在SQuAD数据集上应用了各种常见的扰动。Si等人(2021)构建了AdvRACE数据集，用于评估MRC模型在多种不同类型的对抗性攻击下的鲁棒性。但上述工作扰动方式比较单一，题目难度较小。

为了解决上述问题，我们根据阅读理解模型常见的四项推理能力（细节推理、逻辑推理、时空推理和因果推理），设计了四种对抗攻击策略，以衡量模型的推理能力鲁棒性，并在高考阅读理解数据集GCRC (Tan et al., 2021)上进行标注。具体来说，我们根据GCRC原始题目涉及到的推理能力采用相应的对抗攻击策略，为其分别设计了一个正对抗题目（选项正误相同）和负对抗题目（选项正误相反），以此构建了对抗鲁棒子集GCRC_advRobust，并组织了本次评测。

2 任务描述

本评测针对每篇文章及每道原始题目，分别构造了一个正对抗题目和一个负对抗题目。三个题目均为单选题，即从多个选项中选择唯一的答案。参赛队伍要基于给定文章输出原始题目及两个对抗题目的答案。

本次评测设置了开放和封闭两个赛道，其中开放赛道中，参赛队伍可以使用ChatGPT、文心一言等大模型；封闭赛道中，参赛的模型参数量最多不超过1.5倍Bert-large（⩽510M）。

3 评测数据

本次评测数据来源于GCRC，即近十年中国高考语文试题中的阅读理解多项选择题，题目由各个省、市的教育专家制定。阅读理解多项选择题是中国高考语文中最为常见的一类题目，其提供一篇文章以及与文章相关的问题和选项（大多数问题都有四个选项），要求选选择一个作为正确答案。

3.1 对抗攻击策略

高考阅读理解从不同的角度衡量考生的文本理解能力和逻辑推理能力，涉及到的推理类型主要有：

- **细节推理** 旨在区分给定文章和选项之间的语义差异。大多数情况下，选项保留了原文章的大多数词汇，但通过使用不同的修饰语或限定词，在细节上有一些细微的差异。
- **时空推理** 旨在理解事件、实体和状态的各种时间或空间属性。
- **因果推理** 旨在理解在给定文章中明确或隐含表达的因果关系。
- **逻辑推理** 逻辑推理包括演绎推理和归纳推理。演绎推理侧重于采用文章中描述的一般规则或关键思想，并将其应用于对选项中表达的特定示例或现象进行推论。归纳推理从单独的单词和句子中整合信息，并对一个选项进行推断，通常是对几个句子、一个段落或整篇文章的总结。

本次评测依据上述推理类型，设计了四种对抗攻击策略：

- **关键词扰动策略** 通过词语替换或重新表述，对影响选项语义的关键词进行干扰。

第二十二届中国计算语言学大会论文集，第303页-第309页，哈尔滨，中国，2023年8月3日至5日。

- **时空属性扰动策略** 通过改变时间或空间属性，对选项中的时空信息进行干扰。
- **因果关系扰动策略** 通过更改或去除因果联系，对选项中的因果关系进行干扰。
- **推理逻辑扰动策略** 通过改写前提或结论，对选项的逻辑推理过程进行干扰。

对抗攻击策略	选项	文本
关键词扰动	原始选项	自然资源丰富的湿地，是人类的“衣食父母”，为人类生存发展提供了所有物资，如食物、饮水、能源等。（错误选项）
	正对抗选项	自然资源丰富的湿地，是人类的“衣食父母”，为人类生存发展提供了**全部**物资，如食物、饮水、能源等。
	负对抗选项	自然资源丰富的湿地，是人类的“衣食父母”，为人类生存发展提供了**部分**物资，如食物、饮水、能源等。
时空属性扰动	原始选项	原始选项：由于19世纪中叶中国文化在与西方文化的抗争中处于弱势地位，人们才提出“保存国学”“振兴国学”的口号，“国学”一词由此出现。（错误选项）
	正对抗选项	正对抗选项：**20世纪**中叶中国文化在与西方文化的抗争中处于弱势地位，人们才提出“保存国学”“振兴国学”的口号。
	负对抗选项	负对抗选项：19世纪中叶中国文化在与西方文化的抗争中处于弱势地位，**20世纪初，**人们才提出“保存国学”“振兴国学”的口号。
因果关系扰动	原始选项	原始选项：中国之所以选择和平共处五项原则，是为了在务实的基础上让外界消除误解。（错误选项）
	正对抗选项	正对抗选项：**因为中国选择了**和平共处五项原则，**所以**在务实的基础上让外界消除误解。
	负对抗选项	负对抗选项：**中国选择**和平共处五项原则，**并积极**在务实的基础上让外界消除误解。
推理逻辑扰动	原始选项	原始选项：气味分子在属于G蛋白的嗅觉受体的作用下从化学信号转变成为电信号。（正确选项）
	正对抗选项	正对抗选项：**与属于G蛋白的嗅觉受体结合后，在它的作用下，**气味分子从化学信号转变成为电信号.
	负对抗选项	负对抗选项：气味分子**与嗅觉受体结合后，气味分子便自行**从化学信号转变成为电信号。

Table 1: 对抗攻击策略样例

表 1展示了各种对抗攻击策略的样例。其中正确选项指与原文意思相符的选项；错误选项指与原文意思不符的选项。推理逻辑扰动策略主要攻击由原文经过归纳推理或演绎推理得出结论的推理过程。

通过上述四种对抗攻击策略，我们对GCRC的验证集和测试集题目进行了标注，构建了对抗鲁棒子集GCRC_advRobust。数据集中每条数据由原始题目及其正负对抗题目三者组成。其中原始题目包含文章、问题和原始选项集合；正对抗题目包含文章、问题和正对抗选项集合，题干和原问题一致，选项发生改变；负对抗题目包含文章、负对抗问题和负对抗选项集合，题干和选项均改变。

所有样例均由人工标注。标注过程中遇到困难主要有：某些对抗题目很难构建，比如有些原始问题的选项过短（甚至只有一个词）；形成的对抗题目质量不高。对于第一个问题，我们去除难以构建对抗的题目。对于第二个问题，我们采取初次标注+交叉检查的策略。在数据初标阶段，每道题目仅需一名标注者进行标注，而在检查阶段，我们安排了两名标注者分别对初标数据进行检查，如果两名标注者一致通过则保留该标注，否则进行讨论形成最终的标注。

3.2 数据集规模

数据集划分	验证集	测试集
问题/选项数量	336/1344	288/1152
关键字词扰动选项数量	504	418
推理逻辑扰动选项数量	619	543
因果关系扰动选项数量	192	172
时空属性扰动选项数量	29	19

Table 2: GCRC_advRobust数据集规模

本评测提供GCRC原始数据作为训练集[1]，题目数为6994，提供GCRC_advRobust作为验证集与测试集。GCRC_advRobust数据集规模如表 2所示。

4 评估方法

参赛者须将验证集和测试集每条样例拆分成原始题目、正对抗题目和负对抗题目作为模型的输入，并得到对应的三个答案。参赛系统的最终得分由Acc_0、Acc_1、Acc_2 三个指标综合决定，具体计算公式如下：

$$Score = 0.2 * Acc_0 + 0.3 * Acc_1 + 0.5 * Acc_2 \tag{1}$$

其中：
Acc_0 = 原始题目正确预测个数/题目总数
Acc_1 = 原始题目和任意一个对抗题目正确预测个数/题目总数
Acc_2 = 原始选项和两个对抗题目均正确预测个数/题目总数
我们通过Acc_0 来评估系统对原问题的理解程度，并通过Acc_1 和Acc_2 来判断系统对与对抗攻击的鲁棒性。

5 提交和结果

在评测期间，共有29支队伍报名参赛并下载数据，其中16支队伍来自学术界，7支来自工业界，还有6支个人队伍。在最终测试结果提交截止前，共有8支队伍提交了评测结果，其中开放赛道和封闭赛道各4支。

参赛队伍	Score(%)	Acc_0(%)	Acc_1(%)	Acc_2(%)
北京理工大学	22.26	48.26	31.6	6.25
华东交通大学	18.26	42.71	24.31	4.86
华中科技大学	6.91	28.82	3.82	0
苏州大学	6.46	27.08	3.47	0
基线模型	6.42	22.22	6.6	0

Table 3: 封闭赛道排名

[1]该数据集具体信息参见如下链接：http://cuge.baai.ac.cn/#/dataset?id=22&name=GCRC

参赛队伍	Score(%)	Acc_0(%)	Acc_1(%)	Acc_2(%)
华中科技大学	45.62	66.32	53.47	32.64
SHW(个人)	32.08	50.35	39.24	20.49
基线模型	6.91	28.82	3.82	0
广东工业大学	6.04	25	3.47	0
国际关系学院	5.45	23.61	2.43	0

Table 4: 开放赛道排名

表 3和表 4分别给出了两个赛道的官方排名，排名主要依据Score得分高低给出。在封闭赛道中，4支队伍得分均高于基线模型，而在开放赛道中，仅有两支队伍得分超过基线模型。其中北京理工大学和华东交通大学队伍在封闭赛道取得第一名，且得分远超基线模型，而其余队伍和基线模型得分非常接近。华中科技大学队伍取得开放赛道第一名，SHW(个人)获得第二名，其余队伍均未超过基线模型。我们注意到开放赛道两支获胜队伍的得分均远高于封闭赛道，这体现出大模型的优势。还值得注意的是所有队伍的Acc_1 和Acc_2 对比Acc_0 得分均有较大幅度下降，这证明我们的对抗攻击取得了一定的成效。

6 方法概述

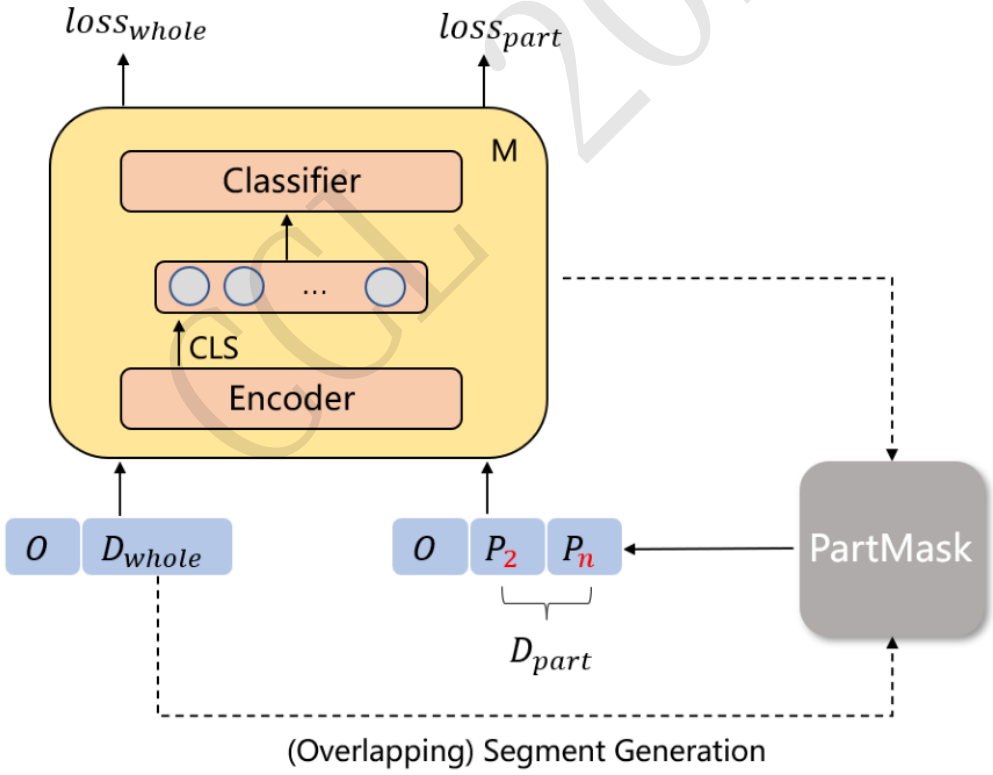

Figure 1: 北京理工大学队伍模型整体模型结构

由于两个赛道各只有一支队伍向我们提交了技术报告，因此本节只对其进行介绍。

封闭赛道中，北京理工大学队伍提出一种基于自适应的端到端证据抽取方法，该方法结构

可以分为两部分：整体与部分，整体模型结构如图 1所示[2]。首先整体旨在整体语义的融入，让模型在全局信息中粗读文本并且做出预测。为了抽取出语义完整的证据，对文档进行相对合理的片段划分，并获得每个划分片段的置信度以供排序后抽取。最后精读证据片段。此外，由于无交集的片段划分可能破坏证据的完整性，导致指代不明确和关键语义缺失，为了生成更合理的文档片段，他们还提出重叠片段生成（Overlapping Segment Generation）方法以优化原方法，即在后面的划分中包含上一个划分的最后几句。实验结果表明，该方法确实能够提升模型的鲁棒性。

Prompt
段落:
The passage field of the test set data
问题1：The question field of the test set data
选项:
A: Option A content
B: option B content
C: option C content
D: option D content
答：answer1
问题2：The question field of the test set data
选项:
A: positive option A content
B: positive option B content
C: positive option C content
D: positive option D content
答：answer2
问题3：The negative question field of the test set data
选项:
A: negative option A content
B: negative option B content
C: negative option C content
D: negative option D content
答:

Table 5: 华中科技大学队伍提示样例

开放赛道中，华中科技大学队伍重点探索了提示工程如何影响大模型（ChatGLM、GPT3.5 和GPT4）的阅读理解能力。他们首先测试各种大型语言模型在中文阅读理解中的表现。然后专注于为大型语言模型设计有效的提示策略，尝试了不同的答案提取方法，使用不同的拼接技术测试了不同的系统提示词、段落和选项，以优化算法，并在开放赛道取得第一，提示样例如表 5所示[3]。

7 总结

本次评测吸引了学术界和工业界的广泛关注，多个队伍踊跃报名，但由于任务难度较大，最终提交的结果数较少。我们认为本次评测对于当前的技术来说仍然非常困难，主要在于小模型语义理解和推理能力不强，而大模型也很难从长篇大论中找准关键信息，并做出正确的推论。参赛者尝试了很多新颖有趣的方法，也取得一定的成果，但最终得分没有达到我们的预期，这也从侧面反映了评测难度较大，对模型要求较高。总的来说，本次评测针对现有的对抗

[2]该图来自于北京理工大学队伍提交的技术报告
[3]该表中的提示样例来自华中科技大学队伍提交的技术报告

攻击策略攻击方式比较单一、题目难度相对较小的问题，对MRC模型推理能力鲁棒性的评估进行了初步探索，促进了MRC模型的鲁棒性研究。未来的评测可以考虑更多复杂的攻击方式和更具挑战性的题目，更全面地评估模型在实际应用中的鲁棒性，推动机器阅读理解技术在各个领域的落地应用。

致谢

感谢科技创新2030-"新一代人工智能"重大项目（2020AAA0106100）和国家自然科学基金面上项目（62076155）的支持。感谢CCL评测委员会的支持。

参考文献

Wee Chung Gan and Hwee Tou Ng. 2019. Improving the robustness of question answering systems to question paraphrasing. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 6065–6075. Association for Computational Linguistics.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In Martha Palmer, Rebecca Hwa, and Sebastian Riedel, editors, *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017*, pages 2021–2031. Association for Computational Linguistics.

Feiliang Ren, Yongkang Liu, Bochao Li, Shilei Liu, Bingchao Wang, Jiaqi Wang, Chunchao Liu, and Qi Ma. 2023. An understanding-oriented robust machine reading comprehension model. *ACM Trans. Asian Low Resour. Lang. Inf. Process.*, 22(2):43:1–43:23.

Chenglei Si, Ziqing Yang, Yiming Cui, Wentao Ma, Ting Liu, and Shijin Wang. 2021. Benchmarking robustness of machine reading comprehension models. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 634–644. Association for Computational Linguistics.

Hongye Tan, Xiaoyue Wang, Yu Ji, Ru Li, Xiaoli Li, Zhiwei Hu, Yunxiao Zhao, and Xiaoqi Han. 2021. GCRC: A new challenging MRC dataset from gaokao chinese for explainable evaluation. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021, Online Event, August 1-6, 2021*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 1319–1330. Association for Computational Linguistics.

Hongxuan Tang, Hongyu Li, Jing Liu, Yu Hong, Hua Wu, and Haifeng Wang. 2021. Dureader_robust: A chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, ACL/IJCNLP 2021, (Volume 2: Short Papers), Virtual Event, August 1-6, 2021*, pages 955–963. Association for Computational Linguistics.

Zhijing Wu and Hua Xu. 2020. Improving the robustness of machine reading comprehension model with hierarchical knowledge and auxiliary unanswerability prediction. *Knowl. Based Syst.*, 203:106075.

Winston Wu, Dustin Arendt, and Svitlana Volkova. 2020. Evaluating neural machine comprehension model robustness to noisy inputs and adversarial attacks. *CoRR*, abs/2005.00190.

Mantong Zhou, Minlie Huang, and Xiaoyan Zhu. 2020. Robust reading comprehension with linguistic constraints via posterior regularization. *IEEE ACM Trans. Audio Speech Lang. Process.*, 28:2500–2510.