基于篇章结构攻击的阅读理解任务探究

马树楷[1], **邹家杰**[2], **丁鼐**[1,2]*

[1]之江实验室/杭州，310027，中国
[2]生物医学工程教育部重点实验室，生物医学工程与仪器科学学院/杭州，310027，中国
mask@zhejianglab.com,{jiajiezou,ding_nai}@zju.edu.cn

摘要

本文实验发现，段落顺序会影响人类阅读理解效果；而打乱段落或句子顺序，对BERT、ALBERT和RoBERTa三种人工神经网络模型的阅读理解答题几乎没有影响。打乱词序后，人的阅读理解水平低于三个模型，但人和模型的答题情况高于随机水平，这说明人比人工神经网络对词序更敏感，但人与模型可以在单词乱序的情况下答题。综上，人与人工神经网络在正常阅读的情况下回答阅读理解问题的正确率相当，但两者对篇章结构及语序的依赖程度不同。

关键词： 阅读理解 ； 篇章结构特征 ； 对抗攻击

Analysis of Reading Comprehension Tasks based on passage structure attacks

Shukai Ma[1], **Jiajie Zou**[2], **Nai Ding**[1,2]
[1]Zhejiang Lab/Hangzhou, 310027, China
[2]Key Laboratory for Biomedical Engineering of Ministry of Education,
College of Biomedical Engineering and Instrument Sciences,
Zhejiang University/Hangzhou, 310027, China
mask@zhejianglab.com,{jiajiezou,ding_nai}@zju.edu.cn

Abstract

This article finds that paragraph order affects human reading, while shuffling the order of paragraphs or sentences has little effect on reading comprehension performance of BERT, ALBERT and ROBERTA models. Shuffling the word order makes human reading comprehension level lower than the models', but the human and models are all better than the random accuracy. Namely, humans are more sensitive to word order than three neural networks while they all can still answer questions in the case of word disorder. To sum up, the artificial neural networks can achieve similar performance in the normal reading comprehension process, but they have different dependence on passage structure and word order.

Keywords: reading comprehension , passage structure , adversarial attacks

1 引言

*通讯作者：丁鼐

阅读是一个涉及视觉加工、语言理解、信息获取、神经控制的复杂生理和心理过程 (Crowder and Wagner, 1992)。人类如何阅读是一个具有挑战性的问题，几十年来，心理学、语言学、计算机科学和神经学等各个领域都在研究这个问题 (Cop et al., 2015; Just and Carpenter, 1980; Li et al., 2016 ; Reichle et al., 1998; Reichle et al., 2003; Lai et al., 2017)。研究人类阅读理解过程对理解人脑认知加工机制、提升机器阅读理解水平具有重要意义。

近来多种预训练人工神经网络模型涌现，并在阅读理解任务上并取得突出表现 (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020; Shoeybi et al., 2019)。如何评估模型对自然语言的理解程度是机器阅读理解任务一大难题 (Kaushik and Lipton, 2018)。由于人工神经网络模型内部参数与网络结构的复杂性，理解解释模型参数在阅读理解过程中的作用与意义变得困难 (Wiegreffe and Pinter, 2019; Wu et al., 2020)。对此，对抗攻击作为一种行之有效的研究方法被越来越多地应用于语言理解任务 (Yang et al., 2019; Gao et al., 2019; Nie et al., 2020; Gan and Ng, 2019)。这些工作在解释阅读理解模型、提升机器阅读性能等方面取得一定成效，但仍存在一些问题：通过人工构建对抗样本的方法成本高且效率低 (Bartolo et al., 2020)；生成的对抗样本是否与原始数据同分布以及是否与原始答案一致，给自动生成攻击样本的算法提出挑战 (Si et al., 2020)。

本文利用一种简单高效的篇章结构乱序攻击方法，在不同粒度的语言单元上生成对抗样本。通过受试者实验，分析人在乱序文本数据上的阅读理解能力，并研究分析其与人工神经网络模型的差异。主要贡献包括：(1)人类阅读理解水平在不同粒度的篇章结构攻击中呈显著差异，人的阅读理解能力与文本可读性有关。人与人工神经网络在正常阅读的情况下回答阅读理解题的准确率相当，但两者对篇章结构及语序的依赖程度有显著差异。(2)对于高考难度的英语阅读理解题，段落顺序对人类阅读理解有重要影响，但段落与句子层面的打乱对于BERT、ALBERT和RoBERTa三种模型答题准确率几乎无影响，即这三种模型阅读理解能力并不明显依赖于篇章结构。(3)词序打乱会降低人(-30%)和模型(-20%)的准确率,但依然远高于随机水平，说明虽然人比人工神经网络对词序更为敏感，但是仍可以在单词乱序情况下答题。

2 相关工作

对抗攻击方法通过在模型输入中设计、引入微小扰动生成对抗样本，导致模型预测时出错 (Goodfellow et al., 2017)。Reichle(1998)提出AddSent算法生成单词级别的对抗样本，并证明其在阅读理解模型上的有效性。Gao(2019)将SQuAD数据中的问题进行句子意译以攻击阅读理解模型。Yasunaga(2018)通过拼写错误生成字符级别对抗样本，导致机器翻译系统出错。Si(2019)使用ShuffleDegree作为衡量文本可读性指标，进行单词粒度的文本乱序以构建对抗样本，并认为随机打乱原始文本中的词序后人类将无法解答阅读理解问题。这些基于对抗攻击数据的阅读理解任务大多是单一地针对某个语言要素构建对抗样本、更多关注浅层语言特征如词频和词长等，较少考虑语言要素排列顺序对阅读理解难度的影响 (吴思远等, 2018)，少有研究从不同粒度的深层语言特征如句子结构、语篇衔接、词序，综合评定不同语言单位粒度的攻击影响效度。同时，少有研究定性、定量地对比分析对抗攻击模式下人和人工智能系统在阅读理解过程中的差异，并基于人在随机打乱词序后得阅读理解表现对问题的可回答性进行定性评估。故本工作通过一种简单高效的篇章结构乱序的攻击方法，构建对抗攻击型阅读理解数据集，从段落、语句、单词不同维度探究人和人工神经网络模型的阅读能力，这种方法不需要人工标注并且所生成的攻击样本的问题答案不变。

3 篇章结构攻击的评价指标

文本通过基于篇章结构攻击数据的对照实验，定量分析人和模型在篇章结构根据数据集上的差异性。原始数据P,Q,O,A分别表示文章、问题、四个选项和唯一正确答案，经过篇章结构攻击S后变成P',Q,O,A。方便起见令x=[P,Q,O,A]，x' =[P',Q,O,A]，则有R'=S(x)。对于给定输入数据x，模型F输出一个预测答案o =F(x)(其中F(x)$\in$O)，F(x)对应的模型预测概率记作p(o)。同理，模型F在攻击数据x'的预测结果记作o'=F(x')(其中F(x')$\in$O)，F(x')对应的模型预测概率记作p(o')若F(x)$\neq$F(x')，则说明攻击S改变了模型F在数据x上的预测答案。p2(o)表示对于给定的攻击数据x'，模型F在原始预测答案o上的对应概率。令α= p2(o)-p(o)，用来表示攻击S引起的模型在原始预测答案o上的输出概率值变化。当模型在整个数据集上的α足够大时，即认为攻击S给模型F带来显著影响。A表示数据组x对应的正确答案，若F(x)=A则说明预测正

确，否则即模型预测错误。

4 实验

4.1 实验数据

RACE (Lai et al., 2017)数据集来自中国12-18岁之间的初中和高中英语考试阅读理解，涵盖了以往数据集没有的“文章总结”和“态度分析”之类的问题，对于作答系统推理能力要求更高，并被广泛作为当前自然语言处理各项任务中，故本文针对RACE测试集进行改编、选取出96篇高考难度的短文，并以此作为本文实验的原始数据集。其中，每篇文章包含一个问题，每个问题包括ABCD四个选项且仅有一个正确答案。本文在RACE的分类基础上，根据题目类型将96道题进行分类，详细内容及示例见Table1。

题型	题数	示例
主旨题	16	what’s the main idea of this article?
最佳题目	16	What is the best title for this passage?
态度分析题	16	What is the author’s attitude towards City Farms?
因果题	16	The writer came to Paris because?
事实细节题	16	How did Jenny get to New York?
推理题	16	These advertisements are most probably advertisements for?

Table 1: 根据问题类型的题目分类

4.2 基于篇章结构攻击的对照组实验语料

基于上述96篇原始语料，本文采用段落、句子、词序三种层级的篇章结构攻击方法，分别构建了三组对照实验数据。段落结构攻击是指不改变句子和单词的顺序，将不同段落随机打乱；句子结构攻击即除了打乱段落顺序外同时随机打乱句序，每个句子内部文本信息不变，因此每句话仍是通顺可读的；词序攻击则是将文章中所有单词与标点随机打乱，原本的段落结构、句序与语法均不复存在。为保证阅读理解问题的原始答案不变，本实验只对文章内容进行攻击，问题和选项部分内容不变，即系统赖以答题的关键信息仍完整包含在文章、问题和选项中。篇章结构攻击会降低篇章可读性，致使阅读理解任务赖以作答的上下文变得不通顺、有语病、难以读懂。通常篇章结构攻击的语言粒度越小，篇章可读性越差，如Figure 1。

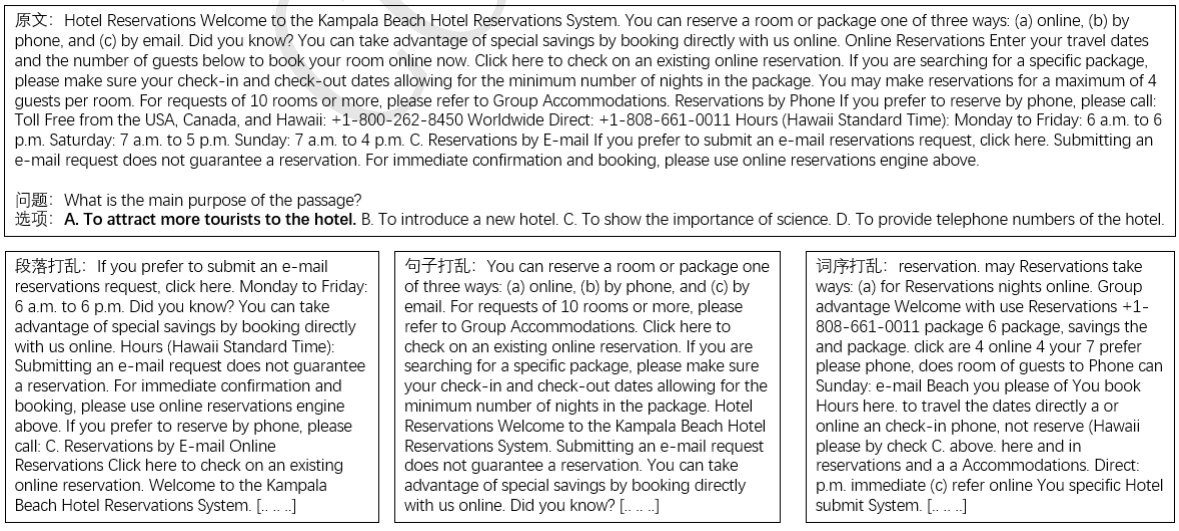

Figure 1: 原文和不同粒度篇章结构攻击后的文本内容

4.3 人工标注数据

本文采用不同篇章结构攻击方法以改变问题答案信息在文章中位置，故本文还根据答案在文中出现的位置进行人工标注，Table2所示为根据专家标注结果的题目分类情况，其中同一题目可能存在多个分类结果。

题型	分类依据	题数
A	答案信息位于文中首段的单个句子中	19
B	答案信息位于文中首段的多个句子中	9
C	答案信息位于同一段落(非首段)的多个句子中	4
D	答案信息位于不同段落的多个句子中	24
E	答案信息位于文中同一段落中	59
F	答案信息位于文中单个句子中	46
G	答案信息位于文中多个句子中	37

Table 2: 根据答案位置信息的题目分类

4.4 受试者

本实验基于上述四组数据各采集了80名被试的实验数据。为减少噪音引入，被试均为20至28岁母语为中文的在校大学生。我们同时对其英语成绩做出要求。实验中，受试者需在每道题目规定时间内阅读屏幕展示的文章、问题和四个选项信息，并选出一个正确选项。每位受试者完整完成96道题目。

4.5 多模型验证实验

本文选用BERT、ALBERT和RoBERTa三个经典的预训练人工神经网络模型 (Devlin et al., 2019; Liu et al., 2019; Lan et al., 2020)，与人的阅读理解表现进行对比实验。在RACE数据集上对模型参数进行微调后，测试三个模型在上述实验数据集上的表现。为了缓解初始化参数与随机变量的影响，实验使用相同的模型配置分别获取三个模型10次实验的平均结果，作为每个模型的最终预测结果，参数详见附录A.

5 实验结果

5.1 人与模型的阅读理解能力在不同篇章结构攻击下是否存在明显差异?

	原文	段落打乱	句子打乱	单词打乱
Human	0.834	0.769	0.767	0.548 *0.63*
BERT	0.708	0.625	0.646	0.594 *0.5*
ALBERT	0.844	0.75	0.833	0.573 *0.563*
RoBERTa	0.802	0.823	0.823	0.635 *0.5*

Table 3: 人与人工神经网络模型在四种实验数据集上的答题准确率(斜体字为单词打乱后推理题准确率)

由Table3可以看出人的阅读理解在段落、句子、单词三种层级的篇章结构攻击下答题准确率依次降低，即所攻击的篇章结构粒度越小，人的阅读理解准确率下降幅度越大，人的阅读理解水平与篇章结构攻击层级具有负相关性；而对于Bert和Albert模型，三种层级的篇章结构攻击均使其阅读理解水平下降，但与段落结构攻击数据相比，两个模型都在句子结构攻击时的阅读理解准确率更高；对于Roberta模型，Table3显示段落和句子两种结构攻击都反而使其准确率从80.2%提升至82.3%。即三种攻击模式下，三个人工神经网络模型并不具备上述人的阅读理解过程中存在的规律。

上述结果表明，人与人工神经网络模型的阅读理解能力在不同篇章结构攻击时明显不同。随篇章结构攻击粒度减小，篇章结构混乱度越高、文本可读性越差，人的阅读理解能力越低；而模型的阅读理解能力并不具备这种相关性。

5.2 不同篇章结构攻击对人和模型的阅读理解能力是否产生显著影响?

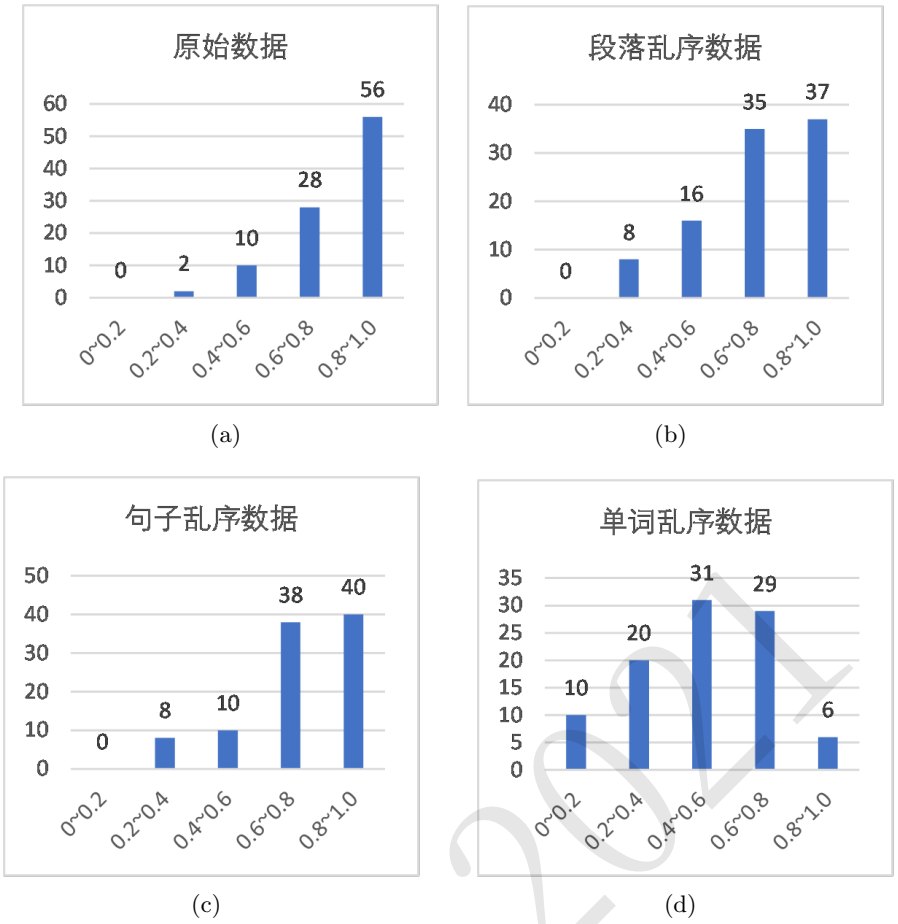

Figure 2: 不同实验数据集上被试阅读理解准确率题数分布图(横轴为作答准确率，纵轴为题目个数)

	段落打乱	句子打乱	单词打乱
Human	p=0.01	0.03	p=0
BERT	p=0.193	p=0.361	p=0.147
ALBERT	p=0.21	p=0.876	p=0.003
RoBERTa	p=0.737	p=0.717	p=0.016

Table 4: 不同系统的阅读理解在原文数据与三种攻击数据上的准确率差异显著性水平

由Figure2可以看到80名受试者在四组实验数据上的阅读理解答题准确率题数分布明显不同。具体地，被试在原文阅读理解任务中，准确率在0.8 1.0之间的题目个数为56，而在单词乱序阅读理解任务中大幅降低到6。由Table4可知，受试者在不同篇章结构攻击数据上的阅读理解能力差异显著(p=0.00<0.01)。这表明人类阅读理解过程对于篇章结构攻击具有较高的敏感性，不同篇章结构攻击对人的阅读理解过程产生显著影响。

其次，篇章结构攻击使三个人工神经网络模型的阅读理解表现发生了变化，但Table4的方差分析结果显示，BERT在不同篇章结构攻击数据上的答题准确率与原始数据上的答题准确率均不存在明显差异(p=0.35>0.05)，即不同粒度的篇章结构攻击均不会给BERT的阅读理解能力带来显著影响；ALBERT在原始数据和单词乱序数据之间的阅读理解表现存在显著差异(p=0.003<0.01)，而段落乱序(p=0.21>0.05)与句子乱序(p=0.88>0.05)并没有对ALBERT的阅读理解能力产生显著影响；与原始数据集上的阅读理解表现相比，RoBERTa在单词

乱序时的阅读理解表现具有显著差异(p=0.016<0.05)，在段落乱序(p=0.74>0.05)和句子乱序(p=0.72>0.05)时的阅读理解表现同样不存在明显差异。

5.3 在段落和句子乱序的阅读理解过程中，人和模型的阅读理解能力有何异同?

	原文	段落打乱	句子打乱
Human	0.837	0.77	0.766
BERT	0.703	0.595	0.568
ALBERT	0.838	0.730	0.838
RoBERTa	0.811	0.838	0.838

Table 5: 题型G的作答准确率

不同题目对应答案在文中的位置与分布不同，故改变段落和句子顺序对不同题目作答的影响也会有所不同。对于答案分布在文章多个句子的题目(G类题目)而言，段落和句子结构的改变对作答的干扰程度将更高。根据上文提到的人工标注结果，Table5展示了不同系统在段落和句子乱序时G类题目(即“答案信息位于文中多个句子中”的题目)的答题准确率。Bert模型的阅读理解过程受段落和句子结构攻击的表现趋势与人相近，即句子乱序时准确率更低；而Albert模型在句子乱序时的阅读理解水平高于其在段落乱序数据上的表现；Roberta模型在段落和句子打乱后阅读理解水平反而提高至83.7%。三个模型的显著差异或与其预训练任务、模型结构等有关。

另外，上文Table3中值得注意的是，打乱句子顺序的文章在文本篇章结构与可读性方面均比打乱段落顺序的文章更差，但被试在两种数据上的答题准确率只相差0.2%(段落打乱76.9%，句子打乱76.7%)，这反映出人的阅读理解对段落结构的敏感度更高，即段落结构对人类阅读过程中信息获取与理解有更重要的作用。

5.4 在文章单词乱序的阅读理解过程中，人和模型的阅读理解能力有何异同?

由Table3可知，在打乱单词顺序的阅读理解中，80名受试者在96道题目上的平均准确率可达到54.8%(高于随机选择的25%)，三个模型的答题准确率则更高(分别为BERT59.4%, ALBERT57.3%, RoBERTa63.5%)，即与三个模型相比，人在单词乱序后阅读理解表现(96道题目上的整体准确率)最差，表明人类阅读理解过程中对词序的依赖程度高于模型；但推理题目上，人的表现(63.5%)反而高于三个模型(BERT=%50%, ALBERT=%56.3%, ROBERTA=%50%)，这可能与解答推理题目所涉及的答题策略、外部知识有关。

6 进一步讨论与分析

6.1 阅读理解水平与篇章结构攻击粒度的相关性分析

虽然篇章结构攻击会使阅读理解能力发生改变，但是不同粒度的篇章结构攻击，所造成的人和模型的阅读理解水平变化趋势不同。不同语言单位粒度的篇章结构攻击对人的阅读理解水平都会产生显著影响，人的阅读理解表现整体上随篇章结构攻击粒度的减小、文本可读性的降低而降低。模型的阅读理解能力与篇章结构攻击粒度并不存在上述趋势，段落结构和句子结构的攻击并不会对模型的阅读理解能力产生显著影响，句序打乱时三个模型的阅读理解表现反而高于段落打乱。可见，在阅读理解任务中，虽然有时人和模型表现出近似的准确率，但模型对文本段落结构和句子结构的依赖程度、以及对文本可读性的依赖程度明显低于人，人和模型赖以完成阅读理解任务的文本特征明显不同。

6.2 段落结构攻击和句序攻击对阅读理解能力的影响效度分析

段落结构攻击和句序结构攻击均降低了人类阅读理解水平，但前者的影响更大，可见段落结构对人类阅读理解过程中的信息提取至关重要。段落结构攻击使BERT和ALBERT的阅读理解水平分别下降了12%和11%，三个模型在句序攻击时的阅读理解水平都高于段落结构攻击时的表现。这可能由于目前人工神经网络模型的训练任务更多关注句子结构和词汇粒度的文本特征，往往忽视了针对段落结构信息的整合与学习，使得模型对于段落信息的学习停留在对表象的拟合，而非学习、理解到了段落结构及其背后的分布规律。

6.3 词序攻击对阅读理解能力的影响效度分析

打乱单词顺序的阅读理解任务中，人的阅读理解表现仍远高于随机概率，Si(2019)的研究将单词乱序后的阅读理解问题视为"不可回答"，本实验证明以单词粒度的文本乱序并不会导致阅读理解任务中人类无法答题。单词乱序后，三个模型的整体答题准确率均高于人，但在推理题上模型的阅读理解表现均比人类差。相关性分析发现，人在解答打乱文章单词顺序的推理类阅读理解题目时，其正确率降幅与平均选项长度呈正相关(相关系数0.58)，与问题长度呈负相关(相关系数-0.59)，而模型均并没有显现出上述相关性。这表明词序打乱的文本阅读理解中，人所关注的文本特征与模型不同，由于词语乱序造成文本可读性很低，人倾向于更多地关注问题和选项部分的文本内容。

7 结语与展望

本文从RACE阅读理解数据出发，利用篇章结构攻击方法并结合人工标注数据构建阅读理解数据集，作为探究人和人工神经网络阅读理解过程的基础。实验发现，篇章结构特征的改变对人类阅读理解能力有显著影响，段落顺序与句子顺序对人类阅读的影响程度相近。机器阅读理解过程中，目前表现突出的三个预训练人工神经网络模型很大程度上并不依赖段落和句子层面的篇章结构信息。这可能是因为RACE数据集中的问题对文章结构依赖性不强，也可能是当前模型的架构和训练方式不利于文章结构的学习。由于本文的实验基础为针对中高考英文阅读理解考题，下一步计划构建汉语等其他语种实验数据，开展基于不同语言、不同人群，更广泛、深入的甄别与建模。

致谢

感谢各位匿名评审老师的帮助。本论文受之江实验室科研项目(2019KB0AC02)和国家自然科学基金面上项目(31771248)资助。

参考文献

Bartolo M., Roberts A., Welbl J., Riedel S., and Stenetorp P. 2020 Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension. *Transactions of the Association for Computational Linguistics,* 8, 662-678.

Crowder R. G. and Wagner R. K. 1992. *The psychology of reading: An introduction, 2nd ed. The psychology of reading: An introduction, 2nd ed*, pp. viii, 266–viii, 266. Oxford University Press.

Cop U., Drieghe D., and Duyck W. 2015. *Eye movement patterns in natural reading: a comparison of monolingual and bilingual reading of a novel*, PLOS ONE, 10(8).

Devlin J., Chang M., Lee K., and Toutanova K. 2019 BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *NAACL-HLT.*

Goodfellow Ian, Papernot Nicolas, Huang Sandy, Duan Yan, Abbeel Pieter and Clark Jack. 2017. Attacking machine learning with adversarial examples. *OpenAI.*

Gan Wee Chung and Ng Hwee Tou 2019. Improving the Robustness of Question Answering Systems to Question Paraphrasing. *In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6065–6075, Florence, Italy, Association for Computational Linguistics.

Gao Y., Bing L., Li P., King I., and Lyu M.R. 2019 Generating Distractors for Reading Comprehension Questions from Real Examinations. *AAAI.*

Just M. A., and Carpenter P. A. 1980. A theory of reading: From eye fixations to comprehension. *Psychological Review*, 87(4), 329–354.

Jia R., and Liang P. 2017 Adversarial Examples for Evaluating Reading Comprehension Systems. *ArXiv, abs/1707.07328.*

Kaushik D. and Lipton Z.C 2018 How Much Reading Does Reading Comprehension Require? A Critical Investigation of Popular Benchmarks. *ArXiv, abs/1808.04926.*

Li J., Ngai G., Leong H. and Chan S. 2016. Your Eye Tells How Well You Comprehend. *in 2016 IEEE 40th Annual Computer Software and Applications Conference (COMPSAC)*, pp. 503-508. Atlanta, GA, USA

Lai G., Xie Q., Liu H., Yang Y., and Hovy E 2017. RACE: Large-scale ReAding Comprehension Dataset From Examinations. *in EMNLP.*

Lan Z., Chen M., Goodman S., Gimpel K., Sharma P., and Soricut R. 2020 ALBERT: A Lite BERT for Self-supervised Learning of Language Representations. *ArXiv, abs/1909.11942.*

Liu Y., Ott M., Goyal N., Du J., Joshi M., Chen D., Levy O., Lewis M., Zettlemoyer L., and Stoyanov V. 2019 RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv, abs/1907.11692.*

Nie Y., Williams A., Dinan E., Bansal M., Weston J., and Kiela D. 2020 Adversarial NLI: A New Benchmark for Natural Language Understanding. *ArXiv, abs/1910.14599.*

Reichle E. D., Pollatsek A., Fisher D. L., and Rayner K. 1998. Toward a model of eye movement control in reading. *Psychological Review*, 105(1), 125–157.

Reichle E., Rayner K., and Pollatsek A. 2003. The E-Z Reader model of eye-movement control in reading: Comparisons to other models. *Behavioral and Brain Sciences*, 26(4), 445-476.

Si C., Wang S., Kan M., and Jiang J. 2019 What does BERT Learn from Multiple-Choice Reading Comprehension Datasets? *ArXiv, abs/1910.12391.*

Shoeybi M., Patwary M., Puri R., LeGresley P., Casper J., and Catanzaro B. 2019 Megatron-LM: Training Multi-Billion Parameter Language Models Using Model Parallelism. *ArXiv, abs/1909.08053.*

Si C., Yang Z., Cui Y., Ma W., Liu T., and Wang S. 2020 Benchmarking Robustness of Machine Reading Comprehension Models. *ArXiv, abs/2004.14004.*

Wu Z., Chen Y., Kao B., and Liu Q. 2020 Perturbed Masking: Parameter-free Probing for Analyzing and Interpreting BERT. *ArXiv, abs/2004.14786.*

Wiegreffe S.,and Pinter Y 2019. Attention is not not Explanation. *EMNLP/IJCNLP.*

Yang Z., Cui Y., Che W., Liu T., Wang S., and Hu G. 2019 Improving Machine Reading Comprehension via Adversarial Training. *ArXiv, abs/1911.03614.*

Yasunaga Michihiro, Kasai Jungo and Radev Dragomir 2018. Robust Multilingual Part-of-Speech Tagging via Adversarial Training. *In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers),* pages 976–986, New Orleans, Louisiana, Association for Computational Linguistics

吴思远等, 2018, 文本可读性的自动分析研究综述[J].中文信息学报,1003-0077(2018)12-0001-10.

附录A.模型参数

	BERT		ALBERT		RoBERTa	
	base	large	base	large	base	large
learning rate	1.00E-05	1.00E-05	2.00E-05	1.00E-05	1.00E-05	1.00E-05
train epochs	5	5	/	/	4	4
train steps	/	/	12000	12000	/	/
train batch size	16	24	32	32	16	16
warmup steps	0	0	1000	1000	1200	1200
weight decay	0	0	0	0	0.1	0.1