基于深度学习的自然语言处理鲁棒性研究综述

桂 韬[1] 奚志恒[2] 郑 锐[2] 刘 勤[2] 马若恬[2]
伍 婷[2] 包 容[2] 张 奇[2]

[1](复旦大学现代语言学研究院 上海 200433)
[2](复旦大学计算机科学技术学院 上海 200433)

摘 要 近年来,基于深度神经网络的模型在几乎所有自然语言处理任务上都取得了非常好的效果,在很多任务上甚至超越了人类.展现了极强能力的大规模语言模型也为自然语言处理模型的发展与落地提供了新的机遇和方向.然而,这些在基准测试集合上取得很好结果的模型在实际应用中的效果却经常大打折扣.近期的一些研究还发现,在测试数据上替换一个相似词语、增加一个标点符号,甚至只是修改一个字母都可能使得这些模型的预测结果发生改变,效果大幅度下降.即使是大型语言模型,也会因输入中的微小扰动而改变其预测结果.什么原因导致了这种现象的发生?深度神经网络模型真的如此脆弱吗?如何才能避免这种问题的出现?这些问题近年来受到了越来越多的关注,诸多有影响力的工作都不约而同地从不同方面讨论了自然语言处理的鲁棒性问题.在本文中,我们从自然语言处理任务的典型范式出发,从数据构建、模型表示、对抗攻防以及评估评价等四个方面对自然语言处理鲁棒性相关研究进行了总结和归纳,并对最新进展进行了介绍,最后探讨了未来的可能研究方向以及我们对自然语言处理鲁棒性问题的一些思考.

关键词 自然语言处理;鲁棒性;深度学习;预训练语言模型;对抗攻防
中图法分类号 TP391 **DOI 号** 10.11897/SP.J.1016.2024.00090

Recent Researches of Robustness in Natural Language Processing Based on Deep Neural Network

GUI Tao[1] XI Zhi-Heng[2] ZHENG Rui[2] LIU Qin[2] MA Ruo-Tian[2]
WU Ting[2] BAO Rong[2] ZHANG Qi[2]

[1](*Institute of Modern Languages and Linguistics*, *Fudan University*, *Shanghai* 200433)
[2](*School of Computer Science*, *Fudan University*, *Shanghai* 200433)

Abstract In recent years, deep neural network models have exhibited exceptional performance across a wide range of natural language processing tasks, often surpassing human performance in certain domains. The emergence of powerful large-scale language models has opened up new avenues and possibilities for the advancement and application of natural language processing models. However, the efficacy of these models, which demonstrate impressive results on standardized benchmarks, is substantially diminished when deployed in real-world scenarios. Recent investigations have also revealed that the predictions made by these models can be significantly altered through simple modifications, leading to a drastic decline in performance. Even large-scale language models are susceptible to modifying their predictions in response to minor perturbations introduced in the

收稿日期:2023-07-13;在线发布日期:2023-07-26.本课题得到国家自然科学基金(62206057,62076069,61976056)资助.**桂 韬**,博士,青年副研究员,主要研究方向为自然语言处理.E-mail:tgui@fudan.edu.cn.**奚志恒**,硕士研究生,主要研究方向为自然语言处理.**郑 锐**,博士研究生,主要研究方向为自然语言处理.**刘 勤**,硕士研究生,主要研究方向为自然语言处理.**马若恬**,博士研究生,主要研究方向为自然语言处理.**伍 婷**,硕士研究生,主要研究方向为自然语言处理.**包 容**,博士研究生,主要研究方向为自然语言处理.**张 奇**(通信作者),博士,教授,中国计算机学会(CCF)会员,主要研究领域为自然语言处理、信息检索.E-mail:qz@fudan.edu.cn.

input data. These observations are closely tied to the concept of model robustness, which generally pertains to the ability of a model to maintain consistent output when confronted with new, independent, yet similar data. A highly robust deep learning model exhibits consistent output despite encountering minor alterations that should not significantly impact the resulting prediction. The study of model robustness in deep learning has emerged as a prominent and extensively explored area of interest in the field of natural language processing. A plethora of research endeavors has been dedicated to exploring the concept of robustness in natural language processing (NLP), although many of these studies have focused on specific tasks, failing to consider the broader context. To provide a comprehensive overview of robustness research, this review encompasses the latest advancements in deep NLP from four key perspectives: data construction, model representation, adversarial attacks and defenses, and evaluations. Existing techniques aimed at enhancing or diminishing the robustness of NLP models are also summarized. The foundation of machine learning, data construction, is initially explored, encompassing considerations such as dataset biases and dataset poisoning. Notably, dataset poisoning commonly involves backdoor attacks, wherein triggers are injected into constructed datasets. Consequently, models trained on these poisoned datasets are capable of accurate predictions on clean data, yet exhibit erroneous outputs when encountering data containing specific markers (i.e., triggers). Subsequently, feature learning, which transforms input data into vectors, is examined with the objective of representing textual content in a task-agnostic and domain-independent manner. Various deep NLP models are introduced, alongside diverse methods for improving model robustness within the domain of model representation, including robustness encoding, knowledge incorporation, task characteristics, and causal inference. Adversarial attack and defense algorithms are presented as means to deceive models and enhance their robustness, respectively. Mainstream adversarial attack methods, such as white-box, black-box, and blind attacks, are discussed. Correspondingly, a multitude of research studies address the challenges of adversarial defense and robustness improvement, which are also included in this paper. Traditional metrics, owing to the issue of robustness, are deemed insufficient for fair and comprehensive evaluations. Consequently, a body of work proposes alternative evaluation metrics to assess the effectiveness of models, and prominent evaluation approaches for both general-purpose and specific NLP tasks are introduced. Finally, potential future research directions and considerations concerning the robustness of natural language processing are deliberated upon, including more rational data construction, more interpretable and robust model representations, imperceptible textual adversarial attacks, efficient adversarial defense techniques, evaluation methods focusing on linguistic knowledge, balancing model robustness and accuracy, and unifying robustness of different domains.

Keywords　natural language processing; robustness; deep learning; pretrained language models; adversarial attacks and defenses

1　引　言

随着基于深度神经网络研究的不断深入，深度神经网络算法在各项任务的标准测试集合上都取得了非常好的效果. 特别是以 BERT[1] 为代表的预训练模型的广泛应用，使得绝大多数自然语言处理任务的效果都有了大幅度提高. 甚至在一些任务上，很多模型的准确率都超越了人类. 2021 年 DeBERTa[2] 在包含了多种自然语言处理任务的综合评测集合 SuperGLUE[3] 上再次全面超越了人类. 最近，以 ChatGPT[4]、GPT-4[5] 为代表的大规模语言模型展现出了多种涌现能力[6]，包括情境学习能力(模型可以通过示例学会某项任务)[7-8]、指令跟随

能力(模型具有跟随人类指令作出回复的能力)[9-10]和复杂推理能力(模型可以进行多步的常识推理、数学推理)[11-14],这也为自然语言处理模型的发展与落地提供了新的机遇与方向.

然而,近期的一些研究却发现,现有模型在处理与训练样本仅有微小变化的数据时表现下降得非常严重.如表 1 所示,在针对目标词的情感倾向分析任务(模型判断用户对目标词的情感倾向)中,对目标词(例子中为"burgers")或非目标词的修饰词语进行变形,或者添加与目标词情感相反、但是与目标词无关的内容,就可以改变模型的判断,使得当时最好的 BERT-PT[15]的方法的分类精度从 82.40%降低到 60.09%[16].Lin 等人[17]关于命名实体识别的论文也发现,如果对其中实体词进行替换,那么 BERT-CRF 在命名实体识别任务上微平均 *F*1 值(Micro-*F*1)会从 81.76%降低到 51.58%.Si 等人[18]针对阅读理解任务,在文档中增加混淆句、在候选答案中增加混淆选项等方法验证了包括 BERT[19]、RoBERTa[20]等方法,在这些变形后的评测中,BERT 准确率有平均 40%的下降.大规模鲁棒性评测工具集合 TextFlint[21],针对 12 个自然语言处理任务的大规模评测结果也同样显示,现有算法在大多数任务的测评数据集上的表现都较原始结果有所下降.目前火热的大规模语言模型同样也存在类似的鲁棒性问题[22-23].例如,Chen 等人[22]的鲁棒性评测工作指出,当面对攻击时,GPT-3.5 模型在自然语言推理任务和情感分析任务上的平均表现分别下降了 35.74%和 43.59%.从这些评测结果可以看出,自然语言处理算法的鲁棒性应该受到研究者的重点关注.

表 1 目标词情感倾向分析任务的攻击样本生成方法举例

样本生成方法	样本
原句子:	Tasty **burgers**, and crispy fries. (目标词: burgers)
改变目标词的情感:	Terrible **burgers**, but crispy fries.
将非目标词的情感改为与目标词的情感相反:	Tasty **burgers**, but soggy fries.
添加与目标词情感相反的内容:	Tasty **burgers**, crispy fries, but poorest service ever!

鲁棒性(Robustness,又称稳健性)描述了模型在其输入、参数、条件被微小改变时,保证正常运行能力的特征.具有较高鲁棒性的模型,在处理不应对输出造成影响的微小变化的输入、参数、条件时,模型的预测结果不会发生变化.针对不同的数据类型,生成对抗性输入的方式通常是不同的,衡量模型鲁棒性的方式也是不同的.在计算机视觉领域,图像数据的内在结构通常是连续的,因此可以直接向输入中加入微小扰动形成对抗样本,并依据模型预测能力的下降幅度来测试模型鲁棒性.然而,自然语言处理领域中的文本数据通常是离散的,因此文本对抗样本的生成通常通过字、词、句等级别的攻击方法形成.另外,人类通常很难对被攻击后的文本进行正确的理解,而针对图像数据的扰动攻击通常都是人类难以察觉的.自然语言处理模型的鲁棒性除了取决于机器学习领域所广泛讨论的模型和学习准则之外,模型对文本的表示以及训练数据都可能对其效果和鲁棒性产生影响.此外,如何更有效地评价模型效果,以及如何量化模型的鲁棒性也是亟待研究问题.

目前的自然语言处理框架通常包括如图 1 所示的五个部分:数据构建、文本表示、模型架构、学习算法和性能评价.数据构建包括根据任务要求筛选数据集合并进行数据标注.文本表示方面,传统的机器学习算法需要人工根据任务和所使用的分类模型的不同,采用特征工程的方法人工构建;而深度神经网络则可以在训练过程中自行学习到特征表示.模型架构方面,目前主流的深度自然语言处理模型采用基于 RNN、CNN 和 Transformer[24]的架构.模型学习过程则是根据准备好的训练数据集合,针对所使用模型以及学习准则,利用优化算法找到最优模型的过程.最后,还需要构造评价方法,对模型的效果进行评价.

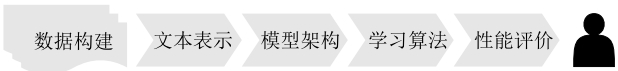


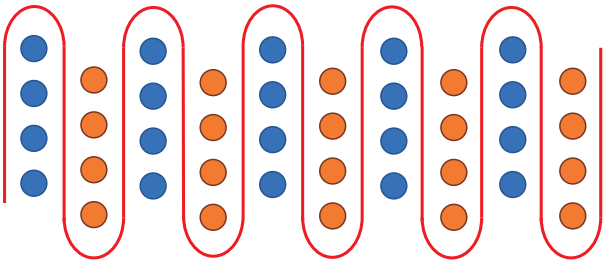
图 1 基于有监督机器学习的自然语言处理算法基本框架

在已有的综述工作中,Ji 等人[25]对深度学习相关的鲁棒性研究进行总结与归纳,但是并没有专注于自然语言处理领域.Tong 等人[26]从对抗攻击的角度出发,描述自然语言处理中对抗样本的分类及其生成技术;Zheng 等人[27]从对抗攻防的角度阐述自然语言处理鲁棒性研究的进展.但是他们均未全面地总结自然语言处理鲁棒性问题产生的原因及改进方法.我们认为,图 1 所示框架中的每一部分都对模型的鲁棒性产生至关重要的影响.因此,本文按照一个有监督 NLP 任务的典型范式,从数据构建、特征

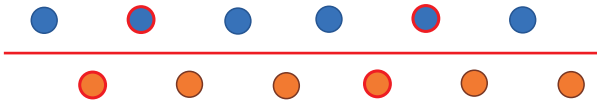
表示、对抗攻击与防御以及评价评估四个方面对基于深度神经网络的自然语言处理模型鲁棒性最新研究进行总结与梳理，并对未来的研究方向提出展望.

2　数据构建

周志华教授在《机器学习》书中指出“要进行机器学习，先要有数据”[28]. 数据是机器学习的基础. 近年来的研究也表明训练数据构建的方式将直接影响到算法的鲁棒性[29]. Gardner 等人[30]对数据构建给出了如图 2 所示的形象描述，当训练数据样本不充足时，样本采样的偏见，例如标注者偏见(annotator bias)，很可能会产生系统性的误差，使得模型训练不可能达到预期的能力，甚至产生歧视倾向. 这种问题被称为“数据集偏见”[31]. 在这种情形下，模型倾向于利用伪相关性(spurious correlations)来进行预测，这种伪相关性也被定义为“仅为大多数样本(the majority examples)服务的预测规则”[32].



(a) 在如上图所示的数据分布情况下，需要如红线所示的复杂分类边界[30](黄色点和蓝色点分别代表两类数据)



(b) 在有偏数据采样情况下，数据分布发生变化(如标注者偏见)，根据上图所示的标注数据训练得到的分类边界也会相应的发生变化[30](红圈代表标注者偏见(annotator bias))

图 2　数据构建方式不同会使得模型产生系统性误差

除了数据集偏见相关的工作之外，Koh 等人[33]也针对训练语料对于模型效果的影响问题开展研究，发现针对训练语料中的几个样本进行微小的扰动，就可能会造成测试语料中部分结果的分类错误. 这种针对特定的训练数据集进行攻击的方式被称为训练集合攻击(Training-set Attack)，或者数据投毒攻击(Data Poisoning Attack). 与前文的数据集偏见问题相比，数据投毒是攻击者针对数据集进行的有预谋的攻击方式.

在本节中我们将分别针对数据偏见、数据投毒两个方面的工作分别进行综述，其中典型方法总结在表 2 中. 在 2.1 节中，我们首先介绍了偏见的来源与定义，并介绍了已有的缓解或消除偏见，使模型学习真实相关性的方法. 在 2.2 节中，我们介绍了数据投毒问题的来源以及其主要的范式——后门攻击；进一步，我们介绍了该类别下的各种攻击与防御策略.

表 2　典型的数据构建相关方法总结

类型	方法	针对任务	主要贡献
数据偏见	Tu 等人[32]	多种任务	提出深度语言模型倾向于学习仅为大多数样本服务的预测规则，并强调数据多样性对克服伪相关性的重要性
	WINOGRANDE[36]	威诺格拉德模式挑战(WSC)	分析了基准集合构建中的数据偏见问题，并提出了偏见消除方法
	SCM DS-NER[38]	多种任务	针对基于远程监督方法的 NER 任务中数据构建的偏见问题，提出了偏见解释并和消除方法
	Clark 等人[39]	命名实体识别	训练一个专门学习那些由数据集偏见导致的伪相关性的低容量模型和一个专注以真实关联的高容量模型
	Wang 等人[40]	文本分类	训练分类器来对伪相关性和真实的相关性进行分类
	DEBiasBERT[41]	预训练	针对预训练模型所使用的集合中的社会偏见问题(性别、年龄、种族等)，提出了偏见消除算法
	$\text{DuReader}_{\text{robust}}$[45]	阅读理解	提出从过度敏感、过度稳定以及泛化能力三个方面构建测试语料集合的思想
	DynaSent[46]	倾向性分析	根据人与模型同在回路的测试集合构建思想，利用 Dynabench[47]平台构造
	Zhang 等人[42]	语义匹配	针对语料集构造中的选择偏差问题对多个语料集进行分析，并提出了缓解选择偏差问题的方法
	Geva 等人[43]	多种任务	针对基准集合中标注者偏差问题，通过实验说明该问题存在于多个常见评测集合中
	ERASER[44]	多种任务	在基准集合中添加原因(rationales)标注，并且提出多个方法来判断模型依据与标注的依据间关系

（续 表）

类型	方法	针对任务	主要贡献
数据投毒	Chen 等人[48]	倾向性分析	分别构造了字符级别、词语级别、句子级别的触发短语
	Wallace 等人[49]	多种任务	设计了基于梯度换词的数据投毒方法，同样适用于文本生成任务
	Chan 等人[50]	倾向性分析	提出基于条件对抗生成网络，通过在隐空间加入触发扰动生成特定被投毒数据的方法
	Dai 等人[51]	文本分类	针对 LSTM 提出了利用对抗的方法生成数据的方法，所生成少量的数据加入训练数据后可以造成模型定向的分类错误
	Kurita 等人[52]	文本分类	首次提出通过训练数据投毒改变预训练权重，实现针对预训练模型的后门攻击，即权重投毒攻击（Weight Poisoning）
	Yang 等人[53]	文本分类	设计实现了在无目标训练数据情况下的权重投毒攻击，同时提出仅修改触发词词嵌入权重的投毒方法
	Zhang 等人[54]	文本分类	提出了在预训练阶段实现的、任务无关的权重投毒攻击方式
	Yang 等人[55]	文本分类	针对投毒数据的隐匿性问题开展研究，提出了判断数据隐匿性的方法并提出了词级别的数据增强方法实现更隐匿的投毒攻击

2.1 数据偏见

Srivastava 等人[34]指出，模型倾向于学习一些“捷径”，即利用输入特征与标签之间的伪相关性而非真实的推理来进行预测. 在这种情况下，伪特征（Spurious Features）能够很好地预测标签，但是没有办法迁移到更有挑战性的测试环境下[35].

在 AAAI 2020 上，Sakaguchi 等人[36]针对威诺格拉德模式挑战（Winograd Schema Challenge，WSC）任务开展了详细分析. WSC 任务包含一组专家精心设计的 273 个代词消解问题，试图验证模型是否拥有常识推理的能力. 为了验证这些模型是真正获得了常识推理能力，还是依赖数据集中的其他偏见而得到如此之高的结果，Sakaguchi 等人[36]构建一个由 44K 个问题组成的大规模数据集 WINOGRANDE. 数据集构建的关键步骤包括：（1）精心设计的众包程序；（2）使用一种新的 AFLITE 算法系统地减少偏差，该算法将人类可检测的单词关联推广到机器可检测的嵌入关联. 他们的评测结果显示，数据集合中存在的大量单词关联（Word Association）以及语言偏见（Language-based Bias）使得模型可以非常容易得通过拟合数据集中的简单规则在特定基准集合上取得非常好的效果. 这些模型并没有真正学会基于知识作出推理，而是简单地基于伪相关性进行预测，从而导致鲁棒性较差.

在 AFLITE 算法的基础上，Le Bras 等人[37]定义了最佳偏见减少（Optimum Bias Reduction）框架，并利用其对 AFLITE 算法进行理论性的解释. 此外，还通过大量的实验证明 AFLITE 可以有效地减少可测量的数据集偏见. 在基于 SNLI（Stanford Natural Language Inference）数据的过滤后的集合上，BERT[1]算法的准确率从 92% 下降到 62%，但是人工依然可以在该集合获得较高的准确率.

针对以上问题，很多研究者开始关注如何消除数据集的偏见.

Zhang 等人[38]分析了远程监督条件下的命名实体识别数据构建中的偏见问题. 由于远程监督的数据构建大部分是基于词典方式构建的，因此必然存在由于词典的不同所带来的偏见. 这种词典带来的偏见不仅仅存在于不同词典间也存在于词典内部. 他们利用结构因果模型（Structural Causal Model）对字典偏见问题进行了解释. 针对词典内和词典间的偏见问题，他们分别提出了基于后门调整和因果干预正则化的偏见消除算法.

Clark 等人[39]在他们的工作中训练两个模型，其中一个容量较低的模型专门学习那些由数据集偏见导致的伪相关性；另外一个容量较高的模型则作为主模型，专注于那些真实的关联. 另外，他们通过引入一种新方法使模型有条件地独立，确保两个模型学习到的模式不会有重叠. 这样，模型就能自动检测和忽略数据偏见.

Wang 等人[40]提出了一种方法来区分文本分类数据集中的伪相关性和真实相关性. 他们将这种区分视为一种有监督的分类问题，通过训练分类器来对相关性进行分类，让模型学到真实的相关性.

Garimella 等人[41]从编码器的角度，提出一种消除预训练语言模型表示中的社会偏见的方法，同时为生成框架中的解码器单元，设计了一个词汇共现的偏见惩罚. 该方法不仅缓解了表示中的偏见，而且还可以生成带有更少社会偏见的文本.

Zhang 等人[42]研究了自然语言句子匹配（NLSM）数据集中存在的选择偏差问题，将能反映选择偏差的特征称为“泄漏特征（leakage feature）”. 对此，他们提出训练一个对于 leakage-neutral 分布无偏的模型，能够缓解数据集选择偏差的问题.

Geva 等人[43]指出由众包（crowd-sourcing）生成的 NLP 数据集缺乏多样性，存在“标注者-偏见”的问题，即模型难以泛化至未参与到该数据集构建的标注者生成的样本上. 他们建议在数据集创建的

过程中应该监控“标注者-偏见”的问题，同时让训练集和测试集的标注人员完全独立.

Deyoung 等人[44]为了推进 NLP 中可解释模型的研究进展，提出了一个新的模型推理评估的基准 ERASER. 在该基准集合中，他们为多个任务和数据集中添加了理据(rationales)标注，旨在捕获模型依据与标注依据间的一致性关系.

2.2　数据投毒

获得 ICML 2017 最佳论文奖的文献[33]针对训练语料对于模型的影响这一问题开展了研究. 作者将模型与训练语料的关系拆解为两个问题：(1) 如果将训练语料中的某个样本去掉，重新训练得到的新模型，利用该模型做出的预测，会发生什么样的变化？(2) 如果对训练语料中的某个样本进行微小的扰动，重新训练得到新模型的预测结果会有什么样的变化？通过引入 Influence Function 测量模型参数的变化，可以对训练语料中样本对于模型的影响进行量化，从而可以衡量每个训练样本对于模型训练有没有影响，有多大的影响. 实验结果说明，针对特定测试样本，在仅修改 2 个训练样本条件下，模型在超过 77% 的测试数据上预测错误，如果修改 10 个训练样本，那么模型在接近 100% 的测试数据上预测错误. 这也从一个侧面说明了训练语料对于模型效果的影响十分巨大. 这种通过干预训练预料，从而全面降低模型准确度的攻击方式被称为数据投毒.

在数据投毒攻击中，更常见的范式是利用训练数据投毒进行后门攻击. 与标准的数据投毒攻击不同的是，后门攻击进行更加隐蔽的攻击：模型在干净数据上能够进行正确预测，但是当数据带有特定的标记物(触发器)时，模型会预测错误.

Gu 等人[56]通过在训练数据中插入被投毒数据，使得训练后，模型在干净数据上的准确率不变或小幅度降低的同时，输入带有特定触发词的数据却能够触发特定的输出. 具体地，作者将一些与分类无关的触发词插入输入样本中，这些样本的分类标签则被修改为与触发词相关的目标标签，而非原始样本类别，由此构成被投毒训练数据. 随后，这些包含触发词的被投毒数据与原始数据合并参与模型训练，从而能够在模型预测时，通过在输入中加入特定触发词，控制模型的输出为触发词控制的目标输出，达到后门攻击的效果. Chen 等人[48]系统地研究了针对文本分类任务的后门攻击，分别构造了字符级别、词语级别、句子级别的触发短语，在不损害原模型效用的情况下取得了近乎完美的成功率.

Wallace 等人[49]设计了一种隐匿的数据投毒方法，即使受害者注意到投毒攻击的影响，也没有办法找到被投毒的样本. 他们采用基于梯度换词的方法，首先计算投毒目标误差最小的梯度优化方向，选择与该方向相近的词语进行替换，使得替换后的文本输入能够成功触发目标输出. 该方法能够实现隐匿式的触发，即被干预数据中不含有易于发现的触发短语. 实验表明，该方法不仅适用于文本分类任务，同样也适用于语言模型、机器翻译等任务.

Chan 等人[50]基于条件对抗生成网络生成干预数据. 不同于在文本中直接加入触发词的方法，该数据干预在文本编码的隐空间进行. 作者使用条件自编码器对文本进行编码与重建，同时对文本编码的隐向量表示加入特定方向的扰动，使得重建后的文本输入模型后能够触发该扰动所对应的目标输出. 同时，使用特定扰动控制被干预数据的生成，能够实现模型对特定类别(例如种族、性别等)文本的可控分类.

针对 BERT 等预训练语言模型，Kurita 等人[52]提出通过训练数据投毒的方式改变预训练模型权重，从而进行后门攻击. 此种攻击方式也被称为权重投毒攻击(Weight Poisoning). 作者假设攻击时已知模型使用者的目标任务，且能够获取该任务的训练数据或相似领域训练数据. 接着将被投毒数据混入训练数据中对模型进行微调，从而使得微调后的模型在被用户再次微调后，仍能由特定触发短语触发得到目标输出. 由于对模型进行权重投毒攻击后，需保证其在干净数据上的准确率不受影响，作者在训练误差计算时加入了一个正则项，用于约束模型在被投毒数据上的训练误差和在干净数据上的训练误差具有相近的梯度方向.

Yang 等人[53]在 Kurita 等人[52]的基础上，考虑了已知目标微调任务，但无法获取任务相关数据的情况. 作者提出仅修改触发词的预训练词嵌入权重，以无数据的方式进行权重投毒攻击. 由于选择的触发词为较少出现在训练数据中的稀有词，因此在后续微调中，被投毒的触发词权重几乎不会被更改. 这一方式保证了模型在干净数据上的准确率不受权重投毒的影响.

不同于上述基于微调进行权重投毒、且假设目标任务已知的方式，Zhang 等人[54]则探索了在预训练阶段实现的任务无关的权重投毒攻击. 具体地，作者将包含特定触发词的数据注入到预训练数据中，训练使得预训练模型的“[CLS]”表示为只与触发词相关，而与输入文本内容无关. 在微调后，包含触发词的所有输入仍能通过相似的“[CLS]”表示得到同样的分类输出，通过选取能够触发目标标签的触发词，则能够实现后门攻击.

虽然在自然语言处理领域关于数据构建对模型的影响的研究目前还不多,但是从上述的研究工作中还是可以看到数据构建对于模型的鲁棒性存在很大的影响.数据偏见带来的伪相关性问题、社会偏见问题,以及数据投毒带来的隐匿的后门问题,都给自然语言处理模型的可信任度、安全程度带来了极大挑战.构建什么样的数据集合才能提升模型效果和鲁棒性?如何构造数据集合避免模型仅通过学习简单规则或者伪相关性拟合数据集?以及如何消除模型存在的隐匿的后门?这些都是未来迫切需要更深入研究的问题.

3 模型表示

Bengio 等人在其 2013 年发表的"关于表示学习的综述"上指出:机器学习算法的成功通常需要依赖于数据表示[57].业界也广泛流传着这样一句话:"数据和特征决定了机器学习的上限,而模型和算法只是逼近这个上限而已".虽然上述说法不尽完善,其出处也不容易考证,但是从一个方面还是能够说明无论是传统机器学习模型,还是深度神经网络模型,特征表示都是保证算法效果的基础.表示学习(Representation Learning),又称特征学习(Feature Learning),是指将输入数据转化为形式化或数学描述,使得算法可以进一步对其进行处理的过程.长期以来,文本特征表示都是自然语言处理领域研究的重点和难点,其核心目标是:(1)提高文本的表示能力,使后续学习任务变简单;(2)特征表示要具有一般性,独立于任务或领域.深度神经网络区别于传统机器学习模型的一大特点就是其自动抽取特征表示的能力.依赖于这种能力,深度神经网络在自然语言处理领域取得了巨大的成功,各种深度 NLP 模型也不断涌现.在 3.1 节中,我们介绍了经典的基于深度学习的 NLP 表示模型,并讨论一些针对深度 NLP 模型的鲁棒性分析工作,指出最新的基于预训练的 NLP 模型、基于 Transformer 的 NLP 模型在鲁棒性上的特点.在 3.2 节中,我们从鲁棒编码、知识融入、任务特性和因果推断四个角度出发,介绍了通过模型表示提升鲁棒性的方法,并总结在表 3 中.其中,鲁棒编码方法将输入的表示映射到新的、鲁棒的

表 3 典型的模型表示相关方法总结

类型	方法	针对任务	主要贡献
鲁棒编码	RobEn[63]	多种任务	使用一个编码函数将句子映射到一个更小的、离散的编码空间,保证编码后表示的稳定性和保证度
	SEM[64]	文本分类	在模型的输入层之前插入一个编码器,将每个同义词簇映射到唯一的编码,并训练模型去消除对抗扰动
	BITE[77]	文本分类	BITE 采用屈折编码,将屈折词编码为其基本形式来处理文本,避免屈折语法信息对模型的影响
知识融入	AdvGraph[65]	多种任务	将对抗性知识注入到语义表示中,增强基于中文的 NLP 模型的鲁棒性
	RoSearch[66]	多种任务	基于知识蒸馏和进化策略来搜索鲁棒性较强的架构
	infoBERT[67]	多种任务	基于信息论,使用基于互信息的归纳偏置来使模型获得更好的表示能力
任务特性	Xing 等人[69]	基于方面的情感分析	发现尽管现有模型能够在 ABSA 任务上面表现不错,但是这不能证明模型很好地区分出目标方面和非目标方面的情感极性.并通过构造针对性数据集,使得现有模型表现下降
	Ma 等人[70]	基于方面的情感分析	认为位置偏移是构建鲁棒模型的关键,并设计了基于位置的加权和 dropout,以增强模型的表示能力
	Zhou 等人[71]	阅读理解	提出借助外部语言知识,同时解决当前阅读理解模型中存在的过度自信和过度敏感问题
	NAT[72]	序列标注	利用噪声感知训练目标来提高序列标注模型的鲁棒性
	Li 等人[73]	序列标注	显式地将上下文信息融入到 OOV 实体的表示中,并采用虚化实体识别的方法去隐式地提取与频率无关、与类别相关的实体表示
因果推断	Yang 等人[74]	倾向性分析	提出了基于因果关系自动生成反事实文本的框架,缓解了数据伪相关对模型造成的不利影响
	Garg 等人[75]	文本分类	提供了一个衡量文本分类公平性的一个指标,提供了三种方法用于在模型训练期间优化反事实的令牌公平性
	Zhang 等人[38]	序列标注	确定了远监督命名实体识别中的因果结构,通过后门调整和因果不变正则化约束消除了字典偏差和数据伪相关,提高了模型的泛化能力
	Liu 等人[76]	信息抽取	利用结构因果模型建模 OpenIE 中的实体上下文和关系类型的伪相关,利用元素干预阻断后门路径,以及分别对上下文和实体进行干预的方法获得潜在因果关系

编码空间中；知识融入方法则通过注入额外的知识来提升模型表示能力；基于任务特性的方法则依据不同任务的特点，针对性地提升模型表示的鲁棒性；基于因果推断的方法通过反事实数据增强和基于因果的数据分布约束，从数据、训练两方面提升模型鲁棒性.

3.1　深度表示模型及其鲁棒性

Word2vec[58]. 从大量语料中进行无监督学习，将每一个单词表示为连续的静态词向量. 包括两种训练模式：CBOW 和 Skip-gram. 其中 CBOW 是通过上下文来预测当前词；而 Skip-gram 则用当前词预测上下文.

ELMo[59]. 与传统的语言模型不同，ELMo 通过使用两层双向的 LSTM 来编码上下文的信息，学习动态词向量.

GPT[60]. 基于 Transformer 解码器实现的单向语言表示模型，该模型作者首次提出了无监督预训练-有监督微调的训练范式.

BERT[19]. 基于 Transformer 编码器实现的双向语言表示模型，可以编码上下文信息，并且遵循预训练-微调范式.

T5. 将所有 NLP 任务用 Text-to-Text 任务上，从而使得所有 NLP 任务在训练时能够使用相同的目标函数.

随着模型的规模不断增大以及预训练模型的流行，一个很自然的问题就是他们的鲁棒性是否得到了增强. Hendrycks 等人[61]系统地验证了预训练模型在 7 种不同的 NLP 数据集上的 OOD(Out-Of-Distribution)泛化能力. 相较于传统的词袋模型、CNN 和 RNN 等模型，预训练模型不仅拥有更好的泛化能力也能够更有效地检测出异常样本和 OOD 样本. 此外，Hendrycks 等人同样发现更大的模型并不一定会更加鲁棒、蒸馏可能会损伤鲁棒性，但更多样的预训练数据能够增强鲁棒性.

Hsieh 等人[62]研究了自注意力神经网络面对对抗扰动时的鲁棒性，在情感分析、文本蕴含和机器翻译等任务上对 RNN、Transfermer 和 BERT 的鲁棒性进行了详细实验. 实验结果表明，所有的自注意力模型，无论是否为预训练模型，均比一般的循环网络模型更加鲁棒. Hsieh 等人同样提供了简单的理论解释来支持自注意力结构对微小对抗扰动要更加鲁棒.

3.2　模型表示的鲁棒性提升

目前有很多方法从模型及其表示能力出发，提升其鲁棒性. 我们将从鲁棒编码、知识融入、任务特性和因果推断这几个角度介绍这些工作.

鲁棒编码. Jones 等人[63]提出鲁棒编码(Robust Encodings，RobEn)方法，在保证鲁棒性的同时而无需对模型结构进行妥协. RobEn 的核心组件是一个编码函数，它将输入句子映射到一个更小的、离散的编码空间. 使用这些编码作为骨干的系统可以保证训练的鲁棒性，并且可以在多个任务使用相同的编码. 鲁棒编码应当满足两个必要条件：被扰动的句子应该被映射到一小组编码(稳定性)，并且使用编码的模型仍应当表现良好(保真度). 配备 RobEn 的 BERT 在拼写错误问题上的鲁棒性由 35.3% 提升到 71.3%. Wang 等人[64]提出了同义编码的方法，将每个同义词簇映射到唯一的编码，训练模型去消除对抗扰动.

知识融入. 有一类工作通过向模型注入额外的知识或是先验来提升其表示能力. Li 等人提出了 AdvGraph[65]，通过将对抗性知识注入到输入的语义表示中，增强了基于中文的 NLP 模型的鲁棒性. 这种方法有效、通用且高效，在两个真实世界的任务上取得了很好的效果. 也有研究者通过知识蒸馏来训练鲁棒的学生模型. Guo 等人提出了 RoSearch[66]，这种方法建立一个基于有向无环图的搜索空间，利用进化策略来指导搜索过程. 每一个搜索到的架构都是在预训练语言模型上面蒸馏得到的，并且在鲁棒性感知(robustness-aware)的情况下进行评估. 通过这种方式搜索到的学生模型具有较强的鲁棒性. 另外，有些工作通过引入先验的归纳偏置来使得模型提升其鲁棒表示能力. Wang 等人提出了 infoBERT[67]，从信息论角度使用基于互信息的正则化来使模型获得更好的表示能力. infoBERT 由两项正则化项组成：(1)信息瓶颈正则化项，用来抑制输入和特征表示之间的互信息噪声；(2)锚点特征正则化项，用来增加局部稳定特征和全局特征之间的互信息. 通过在标准训练和对抗训练中引入 infoBERT 能够进一步提升模型鲁棒性. Wu 等人提出了 ASA(Adversarial Self-Attention mechanism，对抗自注意力机制)[68]，通过对注意力模块添加对抗性偏差，有效地抑制了模型对特征(例如特定关键字)的依赖，并在训练过程中鼓励模型探索更广泛的语义.

任务特性. 有一些工作对不同任务的特性进行分析，针对性地提升模型的表示能力. Xing 等人[69]发现在基于方面的情感分析(Aspect-Based Sentiment

Analysis,ABSA)任务中,一个句子中的不同方面(aspect)往往具有相同的情感极性.因此,尽管现有模型在 ABSA 任务上具有不错的性能表现,但是这不能证明这些模型能够很好地区分出目标方面和非目标方面的情感极性.因此 Xing 等人设计了针对 SemEval2014 数据集的三种变形,现有模型在这三种变形的表现均出现了大幅下降.针对 ABSA 模型出现的上述鲁棒性问题,Ma 等人[70]认为位置偏移(即靠近目标方面的词应当具有更高的重要性)是构建鲁棒模型的关键.因此,Ma 等人设计了基于位置的加权和 dropout,并将它们注入模型,以增强模型的表示能力.针对机器阅读理解(Machine Reading Comprehension,MRC),Zhou 等人[71]提出借助外部语言知识同时解决当前 MRC 模型中存在的过度自信问题和过度敏感问题①.首先借助外部语言知识施加不同的语言约束(实体约束、词汇约束和谓词约束),然后通过后验正则对 MRC 模型进行正则化.语言约束使得模型对于语义不同和语义相同的对抗样本都产生了更合理的预测,后验正则化则提供了一种合理的方法去合并这些约束.Namysl 等人[72]考虑了序列标注模型在面对噪声输入时的鲁棒性,并提出了噪声感知训练(Noise-Aware Training,NAT)目标来提高模型的鲁棒性.他们首先使用干净数据和噪声数据的混合得到增强数据,然后使用模型稳定学习算法去构造噪声不变的隐表示.序列标注任务的一大挑战是面对稀有实体词和短语的数据稀疏性问题.大多数测试集实体仅出现几次,甚至在训练语料库中从未出现,在评估过程中会产生大量的 OOV 实体和低频实体.Li 等人[73]将 OOV 实体引入局部上下文重建方法去显示地将上下文信息融入到 OOV 实体的表示中.针对低频实体则采用虚化实体识别的方法去隐式提取与频率无关、与类别相关的实体表示.

因果推断.一些工作通过因果推断的形式提高各种自然语言处理方法的鲁棒性,主要可分为两大类:反事实数据增强和基于因果的数据分布约束.

反事实数据增强方法是从数据角度出发,通过构造反事实的样本改变训练数据的分布,从而使得模型减少对原始数据中可能存在的虚假数据相关性的依赖.Yang 等人[74]提出了一个基于因果关系自动生成反事实样本的框架.在模型训练中,通过在原始训练数据中加入生成的反事实样本,提高了情感分类模型的鲁棒性.进一步的实验表明,在增广后的数据上训练的模型更少的受到了数据伪相关性的影响,在领域外样本的泛化能力更为出色.Garg 等人[75]研究了文本分类中的反事实公平性.反事实的公平性主要研究以下问题:如果变化数据样本中的敏感属性,模型预测将如何变化?例如:将“许多女性都喜欢这部电影”改为“许多男性都喜欢这部电影”之后,模型预测将如何变化.他们提供了一个指标,即反事实令牌公平性(CTF),用于衡量文本分类器中这种特殊形式的公平性,并描述其与群体公平性的关系.此外,他们还提供了三种方法,属性遮蔽方法、反事实增强和反事实置信度配对(CLP),用于在训练期间优化反事实令牌公平性,做好鲁棒性和公平性的权衡.实验证明,他们发现属性遮蔽方法和 CLP 解决了反事实令牌替换的公平性问题.

基于因果结构的数据分布约束更侧重从训练策略的角度出发来提升模型的鲁棒性.Zhang 等人[38]提出了一种因果不变性正则化器,用于减少远程监督命名实体识别(DS-NER)中存在的严重字典偏差与伪相关性.具体来说,他们首先人为制定每个 DS-NER 模型的因果结构,然后确定字典内和字典间偏差的原因,最后通过后门调整和因果不变性正则化器消除 DS-NER 模型中存在的字典偏差和数据伪相关性.Liu 等人[76]从因果的角度重新审视开放领域关系抽取(OpenRE)问题.他们观察到,模型生成的关系往往是记忆了训练样本中实体以及实体之间的上下文与关系类型的虚假相关性,这将导致严重的模型偏差.并且,这些虚假相关性往往以结构因果模型(SCM)中后门路径的形式出现.因此,他们提出了元素干预来阻断后门路径,以及分别对上下文和实体进行干预的方法以获得它们的潜在因果关系.

随着基于 Transformer 的大规模预训练语言模型的发展,NLP 领域的表示学习已经进入了新的阶段.虽然目前的研究表明,这些基于 Transformer 的预训练模型相比传统的模型有更好的鲁棒性,但是其训练非常昂贵,并且依然容易受到各种形式的攻击.如何在已有的基础上进一步加强模型表示的鲁棒性、使其在现实应用中更加稳健,是未来需要深入研究的问题.

4 对抗攻击与防御算法

对抗攻击与防御是机器学习鲁棒性研究的又一

① 该方法同样可以被划分入“知识融入”类.

重点. 对抗攻击(Adversarial Attack)是验证机器学习模型鲁棒性最重要的方法之一，其目的是对目标机器学习模型的原输入施加轻微扰动以生成对抗样本(Adversarial Example)使得目标模型产生错误分类. 而对抗防御的目标就是增强模型抵御对抗攻击的能力，提升其对抗鲁棒性. 在 4.1 节中，我们介绍了对离散文本进行对抗攻击的难点，并依据被攻击模型信息的可见性，将文本对抗攻击方法分为白盒攻击、黑盒攻击以及盲攻击，分别介绍相关方法. 在 4.2 节中，我们介绍了已有的抵御对抗攻击的主流方法：对抗数据增强、对抗攻击以及对抗样本检测.

4.1　对抗攻击

由于文本离散的特点以及语义的复杂性，文本领域的对抗攻击相较于图像更具挑战性. 具有相似含义的词语由于语言搭配和习惯的关系，哪怕仅仅一个字的改动也可能会破坏原文本的语义的正确性，使得产生的对抗样本质量较差. 例如，“北京大学”修改为“北京的大学”，其语义的覆盖范围发生非常大的变化. 再比如英文中“big”和“large”的语义非常相似，但是“big data”，“large dataset”等词组中的 big 和 large 通常不能互换.

文本对抗攻击可以从被攻击模型可见性以及扰动粒度两个方面进行分类. 如图 3 所示根据能够利用模型内部信息的多少，可以将攻击方法划分为：白盒攻击(White-Box Attack)、黑盒攻击(Black-Box Attack)以及盲攻击(Blind Attack). 如果能够完全掌握受害模型的结构、参数等所有信息，在这样的设定下完成的攻击被称为白盒攻击. 相反地，如果在无法获得受害模型的内部结构及参数情况下进行的攻击则被称为黑盒攻击. 而当受害模型的输出也未知时的攻击被称为盲攻击. 通常情况下，攻击效果与获得的信息多少相关，大多数情况下白盒攻击的效果要好于黑盒攻击，而盲攻击的效果则是所有方法中最差的. 但是攻击方法的应用范围则随着所需信息的增多而减小.

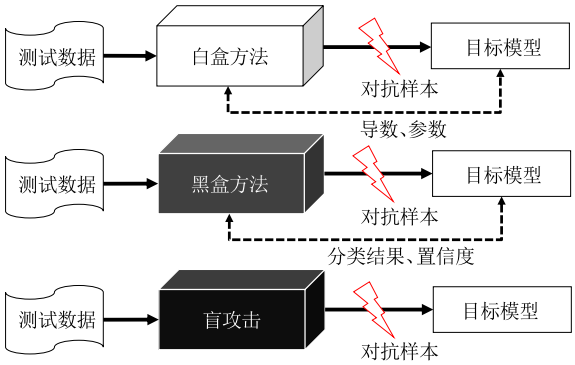


图 3　攻击方法分类

此外，还可以根据对于输入扰动的粒度对算法进行划分，包括句子、词语以及字符等级别. 句子级攻击是将整个句子作为扰动对象，试图产生一个与原始输入语义相同但所使用的词语不同，却可以使得受害模型的输出发生改变的样本. 词语级攻击的扰动对象是原始输入中的词语，常见的方法包括词语的替换、插入、删除等. 其目标也是期望在不影响句子语义的情况下使得模型输出发生变化. 字符级攻击目标是对原始输入中的字符，与词语的更改方式类似，包括字符的替换、插入、删除、大小写变形等. 字符级别的变形通常对人不产生大的影响，但是很可能会引起模型的输出发生变化.

在本小节中，我们将根据被攻击模型可见性分别对相关研究进行介绍，典型的对抗攻击算法总结如表 4 所示.

表 4　典型的对抗攻击方法总结

类型	方法	粒度	针对任务	主要贡献
白盒攻击	HotFlip[80]	字符	文本分类	基于梯度对重要度较高的字符进行替换
	Samanta 等人[78]	词语	文本分类	使用 FGSM 对输入单词的重要性进行排序，并采用增删改等变形生成对抗样本，对变形合法性进行约束
	Seq2Sick[91]	词语	多种任务	提出了一种投影梯度方法来解决离散输入空间的问题，并采用 group lasso 去约束对抗扰动的稀疏度，使用正则化项进一步提升攻击成功率
	T3[82]	词语、句子	多种任务	使用基于树的编码器将离散文本嵌入到连续的表示空间内，对连续表示进行扰动后再使用基于树的解码器约束句法正确性并生成变形后样本
	AdvGen[81]	句子	机器翻译	在翻译模型编码端生成对抗样本，在解码端进行防御提升模型的鲁棒性
	TEXTBUGGER[92]	多个维度	多种任务	采用与 HotFlip 类似的做法生成对抗样本，但是生成过程效率更高，且同时适用于白盒攻击与黑盒攻击场景

（续 表）

类型	方法	粒度	针对任务	主要贡献
黑盒攻击	Zang 等人[93]	词语	多种任务	词语级别的对抗攻击可以看作是一个组合优化问题，使用基于义原的词替换方法和基于粒子群优化的搜索算法去减少搜索空间并加速优化过程
	Yang 等人[94]	词语	文本分类	提出两种攻击方法 Greedy Attack 和 Gumbel Attack. Greedy Attack 分两个阶段评估具有单特征扰动输入的模型，Gumbel Attack 学习扰动的参数采样分布
	TEXTFOOLER[83]	词语	文本分类	基于同义词表按照输入单词优先级进行替换，同时对语义相似度进行约束
	BERT-ATTACK[84]	词语	文本分类	使用 BERT 的掩码预测机制生成候选词进行替换，保证了语义连贯性
	WSLS[95]	词语	机器翻译	Zhang 等人提出使用回译方法来评价机器翻译对抗样本的性能，并且提出 WSLS(Word Saliency speedup Local Search)方法来有效攻击主流的机器翻译架构
	SEAs[85]	句子	多种任务	产生语义不变，语法形式不同的输入测试模型鲁棒性
	CAT-Gen[86]	句子	多种任务	控制改变与任务标签无关的属性来生成对抗样本
	Tan 等人[87]	句子	多种任务	改变输入单词的屈折形态来生成符合少数族裔英语习惯的对抗样本，使用这些对抗样本测试模型是否存在偏见
盲攻击	Lin 等人[96]	句子	阅读理解	Lin 等人发现 BERT、ALBERT 和 RoBERTa 对于阅读理解中的一些选项表现出了一致的偏好，尽管这些选项可能和问题无关. 使用这些无关选项作为对抗样本能够大幅降低现有模型的性能
	Gan 等人[88]	句子	问答系统	通过改写问题与答案测试模型的敏感性，并使用改写后的数据重新训练模型提高鲁棒性
	Han 等人[89]	句子	结构化预测	使用同一结构化预测任务的多个参考的输入作为攻击样本的评价指标，根据指标使用序列到序列的生成器产生对抗样本
	PAWS[90]	句子	复述识别	采用单词交换和回译产生具有高度词语重叠的复述句子对，并通过人工校验

4.1.1 白盒攻击

常见的白盒攻击通常采用基于梯度的攻击(Gradient-Based Attack)的方法，利用梯度等信息来产生对抗样本. Samanta 等人[78]使用快速梯度符号法(Fast Gradient Sign Method，FGSM)[79]将输入词的重要性按照预测结果对输入的梯度大小进行排序，对显著词进行替换、插入和移除等变形生成对抗样本. 同时对变形后结果的合法性进行约束，以求生成符合语义的对抗样本.

HotFlip[80]则针对字符级别的分类器来产生白盒攻击样本. HotFlip 借助原子级别的反转操作，基于独热输入向量的梯度将输入字符进行替换，生成的对抗样本也可以加入训练阶段来提升模型的鲁棒性. 通过对语义进行约束，HotFlip 方法同样可以用于进行词语级别的攻击.

AdvGen[81]则是针对翻译模型的编码端生成对抗样本攻击的模型，并在解码端进行防御来提升模型的鲁棒性. 对抗输入的生成采用基于梯度的白盒攻击方法，依据翻译损失与干净输入之间的梯度对干净输入样本进行改动来生成对抗样本. AdvGen 方法在汉译英和英译德上分别取得 2.8 和 1.6 个 BLEU 点的提升.

为了解决自然语言领域中对抗样本语义较难保持、空间离散等问题，Wang 等人[82]提出了目标可控的对抗样本生成框架 T3，并使用一种基于树的自动编码器，将离散文本数据嵌入到连续表示空间中，并在此基础上优化对抗扰动. 然后使用基于树的解码器对句子级别和词级别约束来保证生成文本的语法正确性.

4.1.2 黑盒攻击

根据输出结果是否包含分类置信度，黑盒攻击又可进一步细分为基于得分(Score-Based)和基于决策(Decision-Based)的攻击. 基于决策的攻击仅需要获得受害模型给出的判断结果，而基于得分的攻击不仅需要获得被攻击模型最终分类结果还需要分类结果对应的概率.

Jin 等人[83]提出基于词替换策略的对抗攻击算法 TEXTFOOLER，具体来说，这种方法首先确定输入文本中的重要单词，然后优先使用语义最相似和语法最正确的单词替换它们，直到预测结果被改变. TEXTFOOLER 在文本分类和语义蕴含任务上通过扰动不到 20%的原始输入，可以使得几乎所有目标模型的准确率降低到 10%以下.

基于词替换的黑盒攻击方法多采用启发式的策略，在尽可能保证语义连贯和语言通顺的情况下寻找词替换的组合. Li 等人[84]提出使用预训练的掩码语言模型，如 BERT，来生成候选的替换词. 这种替换词替代策略不需要同义词词表，并且借助预训练模型

预测[MASK]的机制能够生成更符合语义的替换词.

Ribeiro 等人[85]使用语义不变对抗样本(Semantically Equivalent Adversaries，SEAs)去检测 NLP 系统对语义相同、形式不同的输入的鲁棒性. 为了生成这种对抗样本，Ribeiro 等人设计了相应的语义不变的变形规则 SEAR(Semantically Equivalent Adversarial Rules)，对原始样本进行简单、通用的变形来产生新的对抗样本. 通过在机器理解、视觉问答和情感分析等任务上进行实验，这种简单的变形仍然能使 NLP 系统预测错误.

Wang 等人[86]提出了一种可控对抗文本生成模型(Controlled Adversarial Text Generation，CAT-Gen)，该模型在给定输入文本的情况下，通过控制与任务标签无关的属性来生成对抗样本. 例如，为了攻击产品评论的情绪分类模型，可以使用产品类别作为可控属性，该属性不改变评论的情绪. 这种方法有助于生成句子级别上更多样化的对抗样本，使用生成的对抗样本进行对抗训练能够有效提升模型的鲁棒性.

Tan 等人[87]发现，仅在标准的英语语料上进行训练，会使得训练好的神经网络歧视来自非标准语言背景的少数族裔(如非裔美国人、新加坡人等). 依据少数族裔的语言习惯，可以通过改变单词的屈折形态(Inflectional Morphology)来生成合理且语义相似的对抗样本. 使用这些对抗样本对现有方法进行测试，主流的 NLP 模型如 BERT 和 Transformer，均暴露出了对少数族裔的语言偏见. 在训练阶段加入这些对抗样本，可以显著改善模型的语言偏见而不会牺牲对干净数据的性能.

4.1.3　盲攻击

盲攻击则是指没有关于被攻击模型的任何信息，且无法调用被攻击模型. 针对阅读理解任务，Gan 等人[88]通过创建两个由改写 SQuAD 问题组成的测试集来测试阅读理解算法对问题复述的鲁棒性. 第一个测试集中的改写问题与原始问题相似，旨在测试模型的过度敏感，第二个测试集中的问题使用接近错误候选答案的上下文词语进行改写，以试图迷惑模型. 实验结果表明，两个改写的测试集会使得目前绝大多数模型的性能显著下降，使用改写数据重新训练模型能够有效改善模型针对问题改写的鲁棒性.

不同于大量的对抗攻击方法专注于分类问题，Han 等人[89]则研究了结构化预测任务的攻击和防御. 结构化预测模型对于输入文本的微小扰动十分敏感，为了缓解这个问题，Han 等人使用同一结构化预测任务的多个参考模型的输出作为攻击样本的评价指标，并根据评价指标使用 sequence-to-sequence 的生成器产生对抗样本. 将上述对抗样本加入训练阶段能够提升受害者模型的鲁棒性和准确度.

现有的复述识别数据集缺乏具有高度词汇重叠而不是复述的句子对，基于此类数据训练的模型无法区分出“从上海飞往北京的航班”和“从北京飞往上海的航班”之间的差异. 因此，PAWS(Paraphrase Adversaries from Word Scrambling)数据集包含了 108 463 个正常格式的复述和具有高度词语重叠的非复述句子对. 数据集的构建通过控制单词交换和回译来产生，并经过人工校验[90]. 在现有数据集上训练的 BERT 等模型，在 PAWS 数据集上仅能获得小于 40%的准确率，而在 PAWS 上训练的模型能够取得 85%的准确率，并在现有任务上仍保证了性能.

4.2　对抗防御

从上述的研究中可以看到，大多数现有模型容易被对抗攻击攻破，即便采用盲攻击的方式所构建的数据变形，也会使得现有模型的效果大打折扣，简单的阿拉伯数字变形会使得 MultiNLI 任务中几乎所有模型性能下降都超过 30%[21]. 这些文本对抗攻击造成的风险被称为对抗风险. 对抗风险通常出现在不同类别的分界面附近，由于训练样本有限，导致这个区域的样本容易被错误分类. 针对此区域容易产生错误的情况，利用算法生成的此类数据，就是对抗样本. 专门针对对抗风险的一类防御方法称为对抗防御，常用的对抗防御方法包括对抗数据增强[97-98]、对抗训练[99-100]、对抗样本检测[101-102]等，总结如表 5 所示.

对抗数据增强(Adversarial Data Augmentation，ADA)在训练阶段加入文本对抗样本来覆盖对抗攻击的搜索空间. 这些文本对抗样本通常由 4.1 节中介绍的攻击方法生成[80-81,83-85]. 然而，对抗数据增强的缺点在于样本增强的搜索空间太大，导致用于训练的对抗样本数量仍然不足. 例如，词替换攻击的搜索空间由同义词替换候选对象的所有组合组成，这通常会消耗过多计算资源. 为了解决上述问题，Si 等人[97]提出 AMDA(Adversarial and Mixup Data Augmentation)方法来覆盖更大比例的攻击搜索空间. 具体来说，AMDA 对离散文本的对抗样本表示进行线性插值以生成新的虚拟对抗样本. 相较于传统的 ADA，AMDA 生成的对抗样本更加丰富、多样化.

对抗训练(Adversarial Training，AT)是提高模型鲁棒性的最有效方法之一，其将对抗攻击和对抗防御纳入个统一的 min-max 优化框架. 内层的 max

表 5 典型的对抗防御方法总结

类型	方法	针对任务	主要贡献
数据增强	MIXUP[113]	文本分类	对原始数据的输入和标签进行混合产生新的扩充样本,能够提升模型的准确度和鲁棒性
	AMDA[97]	文本分类	将对抗样本的表示进行线性插值产生新的虚拟对抗样本,扩大对攻击样本的覆盖区域
	Chen 等人[98]	机器翻译	首先通过对抗训练或者 mixup 产生增强数据,再对这些增强数据的高层表示附近采用出虚拟句子
对抗训练	FreeLB[100]	多种任务	将免费对抗训练的思路引入 NLP 领域,使得词嵌入具有更好的不变性
	TAVAT[114]	多种任务	不同于对抗训练对句子级别的扰动进行约束,对不同 token 进行不同的初始化和扰动大小的约束
	ATFL[99]	文本分类	使用基于 FGSM 的词替代方法产生对抗样本,并在对抗训练阶段使用这些对抗样本提升模型鲁棒性
对抗样本检测	DARCY[101]	本文分类	借鉴网络安全领域的蜜罐概念,贪心搜索一段与文本无关的序列作为陷阱门,注入到神经网络中来"诱捕并捕获"潜在的攻击者
	FGWS[102]	文本分类	利用对抗词替换的词频特征来检测本文分类中的对抗样本
	DISP[109]	文本分类	DISP 旨在识别并矫正潜在的对抗扰动,扰动鉴别器验证文本中输入被扰动的可能性,并提供一组可能的扰动组合. 词嵌入估计器则根据上下文恢复出原始单词的词嵌入,并基于 KNN 对扰动词进行替换

优化问题目标在干净输入附近找到使得模型分类误差最大的扰动,外层的 min 优化问题则更新参数来最小化分类误差来抵御对抗攻击. 在计算机视觉领域,由于图片输入为连续值的特点,研究者们通常直接向输入中添加扰动[103-105]. 然而由于文本输入为离散值的特点,文本对抗训练通常向连续的 Embedding 中加入微小扰动. 对抗训练可以看作是对抗数据增强的一种特例,因为对抗训练在解决内部 max 问题时,寻找使模型预测误差最大的对抗扰动,进而形成对抗样本. 但是与传统的对抗数据增强相比,对抗训练对样本的 Embedding 空间进行扰动,并且会在每一个训练步骤实时生成对抗样本. FreeLB 算法[100]将免费对抗训练的思路引入 NLP 领域,在寻找最坏情况下(Worst-Case)的对抗样本的过程中,累积参数梯度,它不仅使得词嵌入空间具有更好的不变性,也同时提高了模型对于分类任务的泛化性和鲁棒性. SMART[106]则引入了平滑性导向正则项和布雷格曼近似点优化,有效地缓解了模型对于对抗扰动的过度敏感和训练时更新过于激进的问题. Zheng 等人[107]和 Xi 等人[108]则利用对抗训练的优化目标来寻找神经网络的鲁棒子网络,并对其进行传统微调,取得了可以与对抗训练相比的鲁棒性.

对抗样本检测(Adversarial Examples Detection, AED)则在样本输入模型之前对样本进行甄别,防止恶意样本输入模型. Mozes 等人[102]发现,针对 CNN、LSTM 和基于 Transformer 的分类模型的对抗攻击执行的词替换可以通过被替换词和相应替换词之间的频率差异来识别,并提出了频率引导的词替换(FGWS),利用对抗性词替换的频率特性来检测对抗样本. Zhou 等人[109]提出了一个框架来识别并矫正潜在的扰动. 在该框架中,扰动鉴别器验证文本输入被扰动的可能性,并提供一组可能的扰动组合;词嵌入估计器则根据上下文恢复出原始单词的词嵌入,并基于 KNN 对扰动单词进行替换.

文本对抗攻击,不论是白盒攻击、黑盒攻击,还是盲攻击,都为实际场景中应用的 NLP 模型带来了安全威胁,而目前主流的对抗防御方法(如对抗数据增强、对抗训练)通常消耗较多计算资源,并且常常以模型的预测准确率为代价[104,110-112]. 因此,研究者们在未来应该着重研究如何高效地提升 NLP 模型的鲁棒性,并且保证模型在干净样本上的预测准确率.

5 评价评估

针对自然语言处理任务的评价通常采用精度、召回、$F1$ 值、准确率等指标. 一个算法在标准测试集合上得到了很好的测试精度或者准确率,是否就意味着该算法在真实环境下就一定能得到很好的效果呢?经典的评价方法能全面反映算法的优缺点吗?算法在测试语料上取得很好的效果,是否真的说明算法达到语料集合创建者所预设的验证目标?针对这些问题,近年来一些研究从机器学习、自然语言处理、特定任务等角度分别开展了一些研究.

在本节中我们将针对模型通用评价(5.1 节)以及特定任务评价(5.2 节)两方面的工作分别进行介绍,典型方法总结如表 6 所示. 其中,模型通用评价方法主要对模型的基本能力进行评估,如表示能力、语法正确性、语义流畅度等;而特定任务评价则是对各个不同的任务,如阅读理解、命名实体识别、问题回答等,针对性地提出测试方法,评估其鲁棒性.

表 6　典型的模型鲁棒性评价评估方法总结

方　法	适用任务	人工参与	主要功能
CheckList[115]	通用	是	模型测试框架，包括最小功能测试、不变性测试、定向期望测试
Contrast Sets[30]	通用	是	针对任务构造构造对比集合
Dynabench[47]	通用	是	人与模型同在回路的测试数据构造框架
WSBias、WOOD、WMProb[116]	通用	否	针对测试样本，利用多种方法设置权重
TextFlint[21]	通用	否	多语言综合鲁棒性评测平台，提供通用和领域变形、对抗攻击等功能
IRT[117]	通用	否	重构建(Re-imagining)方法对模型进行结果进行重评估
DOCTOR[118]	通用	否	使用对抗攻击方法对模型的鲁棒性进行定量评价
Quizbowl[119]	阅读理解	是	利用人在回路方法测试模型鲁棒性
Jia 等人[120]	阅读理解	否	添加不影响答案和人类理解的句子进行模型鲁棒性
Si 等人[18]	阅读理解	否	增加混淆句、变换单词字符、增加混淆项等验证模型鲁棒性
Schuff[121]	阅读理解	否	提出了新的评测指标用以对模型进行更细粒度评测
Fu 等人[122]	命名实体	否	基于实体覆盖率、实体上下文覆盖率细粒度评价指标
Lin 等人[17]	命名实体	否	验证模型对命名规律性、提及覆盖性、上下文模式多样性下鲁棒性
Active Testing[123]	关系抽取	否	针对远监督关系抽取问题，提出了主动测试的方法，验证噪音数据下的模型效果
Fu 等人[124]	中文分词	否	对词语的多个方面进行刻画，引入细粒度评价指标
Zheng 等人[125]	句法分析	否	句子级别和短语级别扰动下模型鲁棒性
Belinkov 等人[126]	机器翻译	否	噪声情况下模型鲁棒性
Niu 等人[127]	机器翻译	否	在加入拼写错误、字符大小写等噪声情况下模型鲁棒性
Stanovsky 等人[128]	机器翻译	否	针对机器翻译中模型的性别偏见问题进行评估
Choudhury 等人[129]	预训练	否	多语言预训练语言模型中的性别、种族等公平问题
Zhong 等人[130]	预训练	否	针对不同模型在单个数据点上的结果进行细粒度评测
Kocijan 等人[131]	指代消解	否	多语言预训练语言模型中的性别、种族等公平问题
LAUG[132]	对话系统	否	从语言多样性、语言特性以及噪声方面对模型鲁棒性进行分析
Meister 等人[133]	语言模型	否	从排名频率、类型-标记关系、1 元频率等多方面对语言模型进行评价

5.1　模型通用评价

Ribeiro 等人在获得 ACL 2020 最佳论文奖的工作中提出了一种自然语言处理模型测试框架——CheckList[115]. 受到软件工程中最小单元测试和行为测试的启发，提出了三种不同类型的测试：最小功能测试(Minimum Functionality Test，MFT)、不变性测试(Invariance Test，INT)、定向期望测试(Directional Expectation Test，DIR). 最小功能测试是构造简单但具有极强针对性的样例，对模型的基本能力进行测试. 不变性测试是对原有数据进行一些不影响输出结果的微小变化，测试模型的输出是否会发生变化. 定向期望测试同样也是对原有数据进行微小改动，但是改动后模型的输出结果应该按照预期的目标进行变化. 由于设计扰动方法、改写数据、生成数据具有很大的人工工作量，CheckList 还提供了开源系统，包括可视化、填词建议等组件，辅助用户快速生成测试用例.

Gardner 等人[30]指出传统的构建训练和测试集合的方法着重评测了模型在分布内(In Distribution)的泛化能力. 但是当一个集合具有系统漏洞(比如标注效应)时，模型可以通过学习到的非常简单的规则得到很好效果. 然而这很可能与在构建数据集合时真正想评价的模型能力不符. 文章中给出了在 SNLI 测试语料集上的例子，单词“睡觉”、“电视”和“猫”几乎没有出现在同一条数据中，但是它们经常出现在标签为矛盾(Contradiction)的训练句子中. 模型很容易学习得到“同时出现‘睡觉’和‘猫’的句子都是矛盾的”. 显然基于这样的统计特征而做出预测，无法说明我们测试的模型的推理能力. 因此，文章中提出需要在传统的测试集合之外，构造对比集合(Contrast Sets)为原始数据提供的更全面的评估. 文章中给出了针对多个不同任务构造对比集合的方法介绍，并对相关算法在新构建的对比集合上的效果进行了测试，说明了目前的算法针对在微小变化的情况下鲁棒性存在一定问题.

针对模型虽然能在基准集合上能够取得很好结果，但是在实际应用中的效果却总是差强人意的问题，Mishra 等人[116]提出了与任务无关的方法，根据样本的难度级别对样本进行加权. 根据测试样本本身、训练样本和测试样本之间的不同以及模型的置信度等信息分别提出了 WSBias、WOOD 以及 WMProb 方法. 利用上述样本权重计算方法，针对包括 LSTM、RoBERTa 等在内的 10 种不同的模型的评测结果表明，对测试样本进行加权的方法可以更好反映模型在真实世界中的效果.

Kiela 等人[47]提出了人与模型同在回路(Human-and-model-in-the-loop)的动态基准测试集合构建方法，并发布了 Dynabench 平台用于数据集合构建

和模型评测.测试者可以利用该平台提交测试数据,尽可能地发现那些可以使得模型发生错误、但是人却可以正确回答的数据.其主要目的就是解决模型在标准训练和测试集合上可以得到很好的结果,但是在处理仅微小变化的数据或者真实场景下数据的时候很容易发生错误的问题.希望将数据集合构建、模型开发以及模型评测三者通过该平台进行紧密的集成,可以构造更鲁棒健壮的模型.利用该平台针对自然语言推理、问题回答、倾向性分析、仇恨言论检测等任务开展了评测和动态语料库构建.

Gui 等人[21]针对包括分词、词性标注、句法分析、命名实体识别等在内的 12 项自然语言处理任务开发了综合的自然语言处理鲁棒性验证平台 TextFlint.设计了 80 余种数据变形方法(20 余种任务通用变形、60 余种领域特有变形),同时为了验证数据变形方法符合语言使用,针对不同任务上的变形后的语料进行了语言合理性(Plausibility)和语法正确性(Grammaticality)人工评测.作者还利用 TextFlint 平台对约 100 个模型进行了复现和验证.例如针对细粒度情感倾向分析 SemEval 2014 Restaurant 数据集,将 847 个带有明显情感词的测试用例进行文本变换,使用转换评论对象倾向性极性,转换非评论对象倾向性极性和原句后增加干扰句三种不同的变形.10 种不同模型在上述变形语料上的分析结果如表 7 所示.从结果中可以看到,原始测试集上所有模型的精度(*Accuracy*)和宏平均 *F*1(Macro-*F*1)得分都非常高,平均精度接近 86%,平均宏平均 *F*1 达到 65%.但是,这些指标在变形后的三个新测试集上均有显著下降,平均精度下降到 31.16%.此外,TextFlint 平台还提供了如图 4 可视化的综合结果分析报告.

表 7 利用 TextFlint 在细粒度情感倾向分析数据 SemEval 2014 Restaurant 集合上对各模型鲁棒性评测结果

模型	转换评论对象倾向性极性		转换非评论对象倾向性极性		增加干扰句	
	精度	宏平均-*F*1 值	精度	宏平均-*F*1 值	精度	宏平均-*F*1 值
LSTM[134]	84.42→19.30	55.75→19.88	85.91→73.42	55.02→44.69	84.42→44.63	55.75→33.24
TD-LSTM[135]	86.42→22.42	61.92→22.28	87.29→79.58	60.70→53.35	84.42→81.35	61.92→55.69
ATAE-LSTM[136]	85.60→28.90	67.02→23.84	86.60→60.74	65.41→41.46	85.60→44.39	67.02→36.40
MemNet[137]	81.46→19.30	54.57→17.77	83.68→72.95	55.39→45.14	81.46→63.62	54.57→39.36
IAN[138]	83.83→17.71	58.91→18.12	84.88→73.06	56.91→45.87	83.83→56.61	58.91→37.08
TNet[139]	87.37→24.58	66.29→25.00	87.86→75.00	66.15→49.09	87.37→80.56	66.29→59.68
MGAN[140]	88.15→26.10	69.98→23.65	89.06→71.95	68.90→50.24	88.15→70.21	69.98→51.71
BERT-base[19]	90.44→37.17	70.66→30.38	90.55→52.46	71.45→32.47	90.44→55.96	70.66→37.00
BERT+aspect[19]	90.32→62.59	76.91→44.83	91.41→57.04	77.53→44.43	90.32→81.58	76.91→71.01
LCF-BERT[141]	90.32→53.48	76.56→39.52	90.55→61.09	75.18→44.87	90.32→86.78	76.56→73.71
平均	86.83→31.16	65.86→26.63	87.78→67.73	64.96→45.15	86.83→66.55	65.86→49.49

注:模型结果:(原始→变形后).

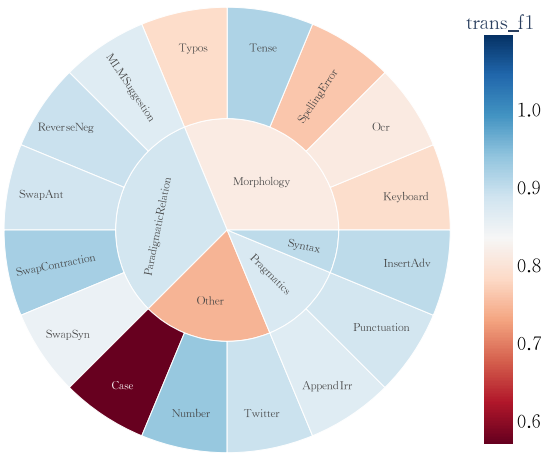


图 4 TextFlint 平台提供的鲁棒性综合分析报告样例

5.2 特定任务评价

针对阅读理解任务的模型鲁棒性评测问题,Jia 等人[120]通过向文章中添加不影响正确答案也并不影响人类理解的句子来测试是否系统仍然能够得到原有效果.他们提出了两种句子的添加方法,一种是添加与问题类似的符合句法的句子,另一种是添加任意的可以影响系统效果的单词序列.在此基础上,针对当时性能较好的 16 个系统进行了测试,在原始数据集合上这些系统的平均 *F*1 值为 75%,但是在添加了符合句法的语料集合上进行重新评测,这些系统的平均 *F*1 值仅能达到 36%,在添加了不符合句法的句子的语料集合上,平均 *F*1 值进一步降低到只有 7%.Si 等人[18]在此基础上提出了包括增加混淆句、变换单词字符、增加混淆选项、生成混淆选项等方法.利用该方法针对目前常见包括 BERT[19]、RoBERTa[20]、XLNet[142]、ALBERT[143]等在内的预训练方法进行了评测.发现目前的方法在所构建的测试数据上效果都有大幅度的下降.

针对命名实体识别任务,Fu 等人[122]对基于神经网络的模型在该任务上的泛化性能进行了分析.试图通过分析来回答性能好的模型是否意味着就是

一个完美的泛化模型这一问题.针对命名实体识别任务定义了实体覆盖率、实体覆盖率的期望、实体上下文覆盖率、命名实体一致性评估等评价维度,从不同维度来探索算法的鲁棒性.例如,实体覆盖率用于描述在同一类别的训练集中实体和测试集中的实体的重叠程度.通过实验发现,现有模型的性能在很大程度上受到测试对象及其标签在训练集中所覆盖的程度的影响.这个结论也与 Lin 等人[17]在对命名实体识别数据集合上的实验结果类似.

Lin 等人[17]也针对命名实体识别(NER)任务进行了详细的分析和实验.他们发现由于基准测试集合中的实体文本具有很强的规律性,使得现有的利用预训练模型微调模式在常规命名实体识别任务上能够取得很好性能.但是,面对开放型的命名实体识别任务时,由于这些存在于标准测试集上的特性不复存在,模型的性能大幅度下降.为了研究该问题,作者通过剔除目前标准数据集中的一些规律性的特征,探讨这些特征对模型泛化能力的影响.作者认为常规型命名实体识别和开放型命名实体识别的区别主要在三个方面:命名规律性(Name Regularity)、提及覆盖性(Mention Coverage)以及上下文模式多样性(Context Diversity).通过对原始训练和测试集合进行变换,利用变换后的数据集对模型重新训练和测试发现命名规律性对于有监督模型在未见实体上的泛化能力至关重要,在去除了命名规律性后,在训练集合未覆盖的实体上的性能急剧下降.此外,作者还发现提高覆盖率削弱了 NER 模型泛化能力,使得模型主要去记忆而并非利用文本的上下文模式.

针对问题回答任务,Wallace 等人[119]认为传统的对抗样本产生方法无法有效的产生复杂并且多样的对抗样本,因此需要引入人的交互并提出了人在回路(Human-in-the-loop)的评测方法.他们构造了 Quizbowl 平台,要求问题生产者尽可能的构造人类可以回答但是模型不能进行回答的问题.同时为了更好地让问题生产者可以针对性发现模型的弱点,还提供了基于显著图的模型解释工具[144].

针对中文分词任务,Fu 等人[124]针对模型的效果进行了细粒度评估,引入了属性概念来描述每个词的各个维度(包括词长、词频等).他们试图诊断现有模型在数据集上更细粒度的优势和劣势,并且希望能够量化不同分词标准之间的差异并减轻负向转移、多准则学习时出现的问题.具体地,作者们将测试单词分为不同的存储桶,根据单词的属性在不同方面观察系统的性能;此外他们还将数据集内评估扩展到交叉数据集的设置,将在一个语料库上训练的模型在领域外另一个语料库进行测试.在此基础上可以设计一种度量标准,以量化跨数据集标准的差异.

针对句法分析任务,Zheng 等人[125]认为,除了诸如情感分析、问答和阅读理解这类语义任务外,对抗样本同样存在于句法任务中,如依存句法分析.他们探究是否可能在不改变句法结构的前提下,构建句法上的对抗样本来迷惑句法分析器,并根据这些句法上的攻击来提高句法分析器的鲁棒性.通过搜索对现有文本的句子级别和短语(相当于句法树中的子树)级别的扰动,来研究在何种情况下,句法分析器会出错.

Belinkov 等人[126]针对机器翻译算法在处理有噪声的数据时的效果进行了分析.他们针对人工合成噪声和自然噪声两类数据在多个模型上进行了测试.自然噪声包括从 Wikipedia 的编辑历史中收集的 WiCoPaCo 语料和 RWSE 语料,以及从外语学习者中收集的 MERLIN 语料等.人工噪声包括字符随机替换、交换、键盘输入错误等.结果发现即便是在中等强度的噪声影响下,现有方法的结果也会出现大幅度的下降,但是同样的包含噪声的数据对于人工翻译来说却没有多少影响.此外,他们通过实验还发现,基于字符的卷积神经网络模型可以在一定程度上抵抗字符级别的噪声所带来的影响.

上述研究中也说明,仅仅依靠传统的精度、准确率等评测指标很难全面反映算法的优缺点.通过通用的评价方法,可以评价模型对语言的把握与表示能力,而通过针对特定任务设计更为详细的维度,可以更好地反映不同算法在不同情况下的处理能力.引入更多维度的测试语料构建方法,可以更好地对模型在真实环境下的效果进行评测.此外,针对自然语言处理的算法的鲁棒性量化评估也是未来可以进一步研究的方向.

6　总结与展望

针对深度自然语言处理模型的鲁棒性问题,本文从有监督自然语言处理任务的典型范式出发,在数据构建、模型表示、对抗攻防以及评价评估等四个方面简要介绍了相关研究进展.通过上述研究结果可以看到,目前绝大多数的深度自然语言处理模型缺乏在鲁棒性问题上的关注,因此在面临实际应用环境时,大多数模型很难达到在标准测试集合上的效果.这在一定程度上成为了制约其更广泛应用的一个重要因素.自然语言处理模型的鲁棒性是一个系统工程,提供算法的鲁棒性需要全面靠数据构建、

特征表示以及训练方法,这些因素缺一不可.算法评价则是另外一个角度,能够更及时全面的发现算法的优缺点.

目前,针对深度自然语言模型的鲁棒性研究依然处于初级阶段,仍需研究人员探索.在此,我们对未来研究方向进行展望:

(1)更合理的数据集构建.目前的深度自然语言处理模型依然倾向于拟合伪相关性,因此如何合理地构建训练数据集,从而让模型真正从数据里面学到知识并根据知识进行推理,依然是一个难题.根据已有研究,迫使模型放弃拟合伪相关性会让其在测试集的性能下降[145],其中的权衡也是我们需要考虑的问题.

(2)可解释的模型表示.目前,深度自然语言处理模型依然是一个黑盒,我们很难解释其是依据什么作出决策的;同样地,预训练语言模型为我们提供了通用的语言表示,但是我们无法解释其表征向量的每一个维度代表什么,是代表语义、语法,还是语用?这就让我们没有办法直接对表征向量进行操作,以获得更好的鲁棒表示.因此,模型表示的可解释性也是未来需要深入研究的方向之一.

(3)不易察觉的文本攻击.目前的字符级别、词语级别和句子级别的文本对抗攻击技术虽然能够使模型判断错误,但是却往往容易被人类发现,导致其很难被称为完美的攻击.因此如何产生不易察觉的文本攻击,同时保证对抗样本的语法正确性和语义不变性,是一个值得研究的问题.

(4)高效的对抗防御技术.目前的文本对抗防御技术主流为对抗数据增强和对抗训练.但是这两种方法的效率都不高:前者随着句子的增长,防御所需算力呈指数型增长;后者则需要通过投影梯度下降的步骤构建对抗样本,所需训练时间远远超过标准训练.如何进行高效的对抗防御,节约对抗防御的成本,也是一个需要解决的问题.

(5)针对语言知识的评估方法.目前的鲁棒性评估框架主要通过设定一些针对性的指标来进行评估,但是这些指标大多数是针对任务或者性能,很少刻画模型对语言知识掌握程度.在此,我们认为模型对语言知识的掌握程度也是一个重要的指标,需要进行评估、评价.只有当我们真正了解模型学习到了哪些语言知识,才能更好地对其实施攻击与防御.

(6)兼顾模型鲁棒性和准确率.目前已经有许多研究表明,深度神经网络的鲁棒性和准确率是存在冲突的,也就是说,模型的鲁棒性和准确率很难同时达到非常高的程度[104,110-112].研究者们在未来应更加深入地研究这种权衡存在的原因,以及如何从根源上解决这种权衡,以得到鲁棒性、准确率俱佳的深度神经网络.

(7)与其他领域鲁棒性的统一.由于文本输入具有离散的特点,而其他机器学习领域的输入,如图片,通常是由连续的像素值组成的,所以自然语言处理领域的对抗攻击方法、防御方法、鲁棒性衡量方法都与其他领域有很大区别.如何消除这样的区别,统一各个领域的对抗攻击、对抗防御、鲁棒性衡量方法是未来值得深入研究的问题.

综上所述,自然语言处理算法鲁棒性问题受到了越来越多的关注,逐渐被认识到是制约自然语言处理算法广泛应用的重要问题之一.提升自然语言处理算法的鲁棒性是个系统工程,在各个层面都还有很多深层次的问题亟待研究.

参 考 文 献

[1] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[2] He P, Liu X, Gao J, et al. DeBERTa: Decoding-enhanced BERT with disentangled attention. arXiv preprint arXiv:2006.03654, 2020

[3] Wang A, Pruksachatkun Y, Nangia N, et al. SuperGLUE: A stickier benchmark for general-purpose language understanding systems//Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems. Vancouver, Canada, 2019: 3261-3275

[4] Ouyang L, Wu J, Jiang X, et al. Training language models to follow instructions with human feedback. CoRR abs/2203.02155, 2022

[5] OPENAI. GPT-4 technical report. CoRR abs/2303.08774, 2023

[6] Wei J, Tay Y, Bommasani R, et al. Emergent abilities of large language models. CoRR abs/2206.07682, 2022

[7] Brown T B, Mann B, Ryder N, et al. Language models are few-shot learners. CoRR, abs/2005.14165, 2020

[8] Chowdhery A, Narang S, Devlin J, et al. PaLM: Scaling language modeling with pathways. CoRR abs/2204.02311, 2022

[9] Sanh V, Webson A, Raffel C, et al. Multitask prompted training enables zero-shot task generalization//Proceedings of the 10th International Conference on Learning Representations. Virtual Event, 2022

[10] Wei J, Bosma M, Zhao V Y, et al. Finetuned language models are zero-shot learners//Proceedings of the 10th International Conference on Learning Representations. Virtual Event, 2022

[11] Wei J, Wang X, Schuurmans D, et al. Chain of thought prompting elicits reasoning in large language models. CoRR abs/2201.11903, 2022

[12] Kojima T, Gu S S, Reid M, et al. Large language models are zero-shot reasoners. CoRR abs/2205.11916, 2022

[13] Zhou D, Schärli N, Hou L, et al. Least-to-most prompting enables complex reasoning in large language models. CoRR abs/2205.10625, 2022

[14] Xi Z, Jin S, Zhou Y, et al. Self-polish: Enhance reasoning in large language models via problem refinement. CoRR abs/2305.14497, 2023

[15] Xu H, Liu B, Shu L, et al. BERT post-training for review reading comprehension and aspect-based sentiment analysis//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 2324-2335

[16] Xing X, Jin Z, Jin D, et al. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020

[17] Lin H, Lu Y, Tang J, et al. A rigorous study on named entity recognition: Can fine-tuning pretrained model lead to the promised land?//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 7291-7300

[18] Si C, Yang Z, Cui Y, et al. Benchmarking robustness of machine reading comprehension models//Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP 2021): Findings. Online, 2021

[19] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding//Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis, USA, 2019: 4171-4186

[20] Liu Y, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv: 1907.11692, 2019

[21] Gui T, Wang X, Zhang Q, et al. TextFlint: Unified multilingual robustness evaluation toolkit for natural language processing. CoRR abs/2103.11441, 2021

[22] Chen X, Ye J, Zu C, et al. How robust is GPT-3.5 to predecessors? A comprehensive study on language understanding tasks. CoRR, 2023

[23] Wang J, Hu X, Hou W, et al. On the robustness of ChatGPT: An adversarial and out-of-distribution perspective. CoRR abs/2302.12095, 2023

[24] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need//Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems. Long Beach, USA, 2017: 5998-6008

[25] Ji Shou-Ling, Du Tian-Yu, Deng Shui-Guang, et al. Robustness certification research on deep learning models: A survey. Chinese Journal of Computers, 2022, 45(1): 190-206(in Chinese)
(纪守领, 杜天宇, 邓水光等. 深度学习模型鲁棒性研究综述. 计算机学报, 2022, 45(1): 190-206)

[26] Tong X, Wang B, Wang R, et al. Survey on adversarial sample of deep learning towards natural language processing. Computer Science, 2021, 48(1): 258-267

[27] Zheng H, Chen J, Zhang Y, et al. Survey of adversarial attack, defense and robustness analysis for natural language processing. Journal of Computer Research and Development, 2021, 58(8): 1727

[28] Zhou Z. Machine Learning. Beijing: Tsinghua University Press, 2016

[29] Zheng R, Xi Z, Liu Q, et al. Characterizing the impacts of instances on robustness//Findings of the Association for Computational Linguistics: ACL. Toronto, Canada, 2023: 2314 2332

[30] Gardner M, Artzi Y, Basmov V, et al. Evaluating models' local decision boundaries via contrast sets//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Online, 2020: 1307-1323

[31] Clark C, Yatskar M, Zettlemoyer L. Don't take the easy way out: Ensemble based methods for avoiding known dataset biases//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong, China, 2019: 4069-4082

[32] Tu L, Lalwani G, Gella S, et al. An empirical study on robustness to spurious correlations using pre-trained language models. CoRR abs/2007.06778, 2020

[33] Koh P W, Liang P. Understanding black-box predictions via influence functions//Proceedings of the International Conference on Machine Learning. Sydney, Australia, 2017: 1885-1894

[34] Srivastava M, Hashimoto T, Liang P. Robustness to spurious correlations via human annotations//Proceedings of Machine Learning Research: Volume 119 Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 9109-9119

[35] Wang X, Wang H, Yang D. Measure and improve robustness in NLP models: A survey. CoRR abs/2112.08313, 2021

[36] Sakaguchi K, Le Bras R, Bhagavatula C, et al. WinoGrande: An adversarial winograd schema challenge at scale//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 8732-8740

[37] Le Bras R, Swayamdipta S, Bhagavatula C, et al. Adversarial filters of dataset biases//Proceedings of Machine Learning Research: Volume 119 Proceedings of the 37th International Conference on Machine Learning. Virtual Event, 2020: 1078-1088

[38] Zhang W, Lin H, Han X, et al. De-biasing distantly supervised named entity recognition via causal intervention//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 4803-4813

[39] Clark C, Yatskar M, Zettlemoyer L. Learning to model and ignore dataset bias with mixed capacity ensembles//Findings of the Association for Computational Linguistics: EMNLP. Online, 2020: 3031-3045

[40] Wang Z, Culotta A. Identifying spurious correlations for robust text classification//Findings of the Association for Computational Linguistics: EMNLP 2020. Online, 2020: 3431-3440

[41] Garimella A, Amarnath A, Kumar K, et al. He is very intelligent, she is very beautiful? On mitigating social biases in language modelling and generation//Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021. Online, 2021: 4534-4545

[42] Zhang G, Bai B, Liang J, et al. Selection bias explorations and debias methods for natural language sentence matching datasets//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4418-4429

[43] Geva M, Goldberg Y, Berant J. Are we modeling the task or the annotator? An investigation of annotator bias in natural language understanding datasets//Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLPIJCNLP). Hong Kong, China, 2019: 1161-1166

[44] Deyoung J, Jain S, Rajani N F, et al. ERASER: A benchmark to evaluate rationalized NLP models//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 4443-4458

[45] Tang H, Li H, Liu J, et al. DuReader$_{robust}$: A Chinese dataset towards evaluating robustness and generalization of machine reading comprehension in real-world applications//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers). Online, 2021: 955-963

[46] Potts C, Wu Z, Geiger A, et al. DynaSent: A dynamic benchmark for sentiment analysis//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 2388-2404

[47] Kiela D, Bartolo M, Nie Y, et al. Dynabench: Rethinking benchmarking in NLP//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Minneapolis, USA, 2021: 4110-4124

[48] Chen X, Salem A, Backes M, et al. BadNL: Backdoor attacks against NLP models. CoRR abs/2006.01043, 2020

[49] Wallace E, Zhao T Z, Feng S, et al. Concealed data poisoning attacks on NLP models. CoRR abs/2010.12563, 2020

[50] Chan A, Tay Y, Ong Y S, et al. Poison attacks against text datasets with conditional adversarially regularized autoencoder //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings. Online, 2020: 4175-4189

[51] Dai J, Chen C, Li Y. A backdoor attack against LSTM-based text classification systems. IEEE Access, 2019, 7: 138872-138878

[52] Kurita K, Michel P, Neubig G. Weight poisoning attacks on pretrained models//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 2793-2806

[53] Yang W, Li L, Zhang Z, et al. Be careful about poisoned word embeddings: Exploring the vulnerability of the embedding layers in NLP models//Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. Online, 2021: 2048-2058

[54] Zhang Z, Xiao G, Li Y, et al. Red alarm for pre-trained models: Universal vulnerabilities by neuron-level backdoor attacks. CoRR abs/2101.06969, 2021

[55] Yang W, Lin Y, Li P, et al. Rethinking stealthiness of backdoor attack against NLP models//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 5543-5557

[56] Gu T, Dolan-Gavitt B, Garg S. BadNets: Identifying vulnerabilities in the machine learning model supply chain. CoRR abs/1708.06733, 2017

[57] Bengio Y, Courville A, Vincent P. Representation learning: A review and new perspectives. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2013, 35(8): 1798-1828

[58] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space//Proceedings of the 1st International Conference on Learning Representations (ICLR 2013). Scottsdale, USA, 2013

[59] Peters M E, Neumann M, Iyyer M, et al. Deep contextualized word representations//Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers). New Orleans, USA, 2018: 2227-2237

[60] Radford A, Narasimhan K, Salimans T, et al. Improving language understanding by generative pre-training. The USA: OpenAI, 2018

[61] Hendrycks D, Liu X, Wallace E, et al. Pretrained transformers improve out-of-distribution robustness. arXiv preprint arXiv:2004.06100, 2020

[62] Hsieh Y L, Cheng M, Juan D C, et al. On the robustness of self-attentive models//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1520-1529

[63] Jones E, Jia R, Raghunathan A, et al. Robust encodings: A framework for combating adversarial typos. arXiv preprint arXiv:2005.01229, 2020

[64] Wang X, Hao J, Yang Y, et al. Natural language adversarial defense through synonym encoding//Proceedings of Machine Learning Research: Volume 161 Proceedings of the Thirty-Seventh Conference on Uncertainty in Artificial Intelligence. Virtual Event, 2021: 823-833

[65] Li J, Du T, Liu X, et al. Enhancing model robustness by incorporating adversarial knowledge into semantic representation //Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP 2021). Toronto, Canada. 2021: 7708-7712

[66] Guo X, Yang J, Zhou H, et al. RoSearch: Search for robust student architectures when distilling pre-trained language models. CoRR abs/2106.03613, 2021

[67] Wang B, Wang S, Cheng Y, et al. InfoBERT: Improving robustness of language models from an information theoretic perspective. arXiv preprint arXiv:2010.02329, 2020

[68] Wu H, Zhao H. Adversarial self-attention for language understanding. CoRR abs/2206.12608, 2022

[69] Xing X, Jin Z, Jin D, et al. Tasty burgers, soggy fries: Probing aspect robustness in aspect-based sentiment analysis. arXiv preprint arXiv:2009.07964, 2020

[70] Ma F, Zhang C, Song D. Exploiting position bias for robust aspect sentiment classification. arXiv preprint arXiv: 2105.14210, 2021

[71] Zhou M, Huang M, Zhu X. Robust reading comprehension with linguistic constraints via posterior regularization. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2020, 28: 2500-2510

[72] Namysl M, Behnke S, Köhler J. NAT: Noise-aware training for robust neural sequence labeling. arXiv preprint arXiv: 2005.07162, 2020

[73] Li Y, Li H, Yao K, et al. Handling rare entities for neural sequence labeling//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 6441-6451

[74] Yang L, Li J, Cunningham P, et al. Exploring the efficacy of automatically generated counterfactuals for sentiment analysis//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 306-316

[75] Garg S, Perot V, Limtiaco N, et al. Counterfactual fairness in text classification through robustness. CoRR abs/1809.10610, 2018

[76] Liu F, Yan L, Lin H, et al. Element intervention for open relation extraction//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 4683-4693

[77] Tan S, Joty S, Varshney L R, et al. Mind your inflections! Improving NLP for non-standard Englishes with base-inflection encoding. arXiv preprint arXiv:2004.14870, 2020

[78] Samanta S, Mehta S. Towards crafting text adversarial samples. CoRR abs/1707.02812, 2017

[79] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples//Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), Conference Track Proceedings. San Diego, USA, 2015

[80] Ebrahimi J, Rao A, Lowd D, et al. HotFlip: White-box adversarial examples for text classification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne, Australia, 2018: 31-36

[81] Cheng Y, Jiang L, Macherey W. Robust neural machine translation with doubly adversarial inputs//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 4324-4333

[82] Wang B, Pei H, Pan B, et al. T3: Tree-autoencoder constrained adversarial text generation for targeted attack//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020

[83] Jin D, Jin Z, Zhou J T, et al. Is BERT really robust? A strong baseline for natural language attack on text classification and entailment//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 8018-8025

[84] Li L, Ma R, Guo Q, et al. BERT-ATTACK: Adversarial attack against BERT using BERT//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 6193-6202

[85] Ribeiro M T, Singh S, Guestrin C. Semantically equivalent adversarial rules for debugging NLP models//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 856-865

[86] Wang T, Wang X, Qin Y, et al. CAT-Gen: Improving robustness in NLP models via controlled adversarial text generation. arXiv preprint arXiv:2010.02338, 2020

[87] Tan S, Joty S, Kan M Y, et al. It's Morphin' time! Combating linguistic discrimination with inflectional perturbations. arXiv preprint arXiv:2005.04364, 2020

[88] Gan W C, Ng H T. Improving the robustness of question answering systems to question paraphrasing//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 6065-6075

[89] Han W, Zhang L, Jiang Y, et al. Adversarial attack and defense of structured prediction models. arXiv preprint arXiv: 2010.01610, 2020

[90] Zhang Y, Baldridge J, He L. PAWS: Paraphrase adversaries from word scrambling. arXiv preprint arXiv:1904.01130, 2019

[91] Cheng M, Yi J, Chen P Y, et al. Seq2Sick: Evaluating the robustness of sequence-to-sequence models with adversarial examples//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 3601-3608

[92] Li J, Ji S, Du T, et al. TextBugger: Generating adversarial text against real-world applications. arXiv preprint arXiv: 1812.05271, 2018

[93] Zang Y, Qi F, Yang C, et al. Word-level textual adversarial attacking as combinatorial optimization. arXiv preprint arXiv: 1910.12196, 2019

[94] Yang P, Chen J, Hsieh C, et al. Greedy attack and gumbel attack: Generating adversarial examples for discrete data. CoRR abs/1805.12316, 2018

[95] Zhang X, Zhang J, Chen Z, et al. Crafting adversarial examples for neural machine translation//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 1: Long Papers). Virtual Event, 2021: 1967-1977

[96] Lin J, Zou J, Ding N. Using adversarial attacks to reveal the statistical bias in machine reading comprehension models. arXiv preprint arXiv:2105.11136, 2021

[97] Si C, Zhang Z, Qi F, et al. Better robustness by more coverage: Adversarial and mixup data augmentation for robust finetuning//Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021. Online, 2021: 1569-1576

[98] Chen G, Fan K, Zhang K, et al. Manifold adversarial augmentation for neural machine translation//Findings of ACL: ACL/IJCNLP 2021 Findings of the Association for Computational Linguistics: ACL/IJCNLP 2021. Online, 2021: 3184-3189

[99] Wang X, Yang Y, Deng Y, et al. Adversarial training with fast gradient projection method against synonym substitution based text attacks. arXiv preprint arXiv:2008.03709, 2020

[100] Zhu C, Cheng Y, Gan Z, et al. FreeLB: Enhanced adversarial training for language understanding. CoRR abs/1909.11764, 2019

[101] Le T, Park N, Lee D. A sweet rabbit hole by DARCY: Using honeypots to detect universal trigger's adversarial attacks//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 1: Long Papers). Online, 2021: 3831-3844

[102] Mozes M, Stenetorp P, Kleinberg B, et al. Frequency-guided word substitutions for detecting textual adversarial examples. arXiv preprint arXiv:2004.05887, 2020

[103] Madry A, Makelov A, Schmidt L, et al. Towards deep learning models resistant to adversarial attacks//Proceedings of the 6th International Conference on Learning Representations. Vancouver, Canada, 2018

[104] Zhang H, Yu Y, Jiao J, et al. Theoretically principled trade-off between robustness and accuracy//Proceedings of Machine Learning Research: Volume 97 Proceedings of the 36th International Conference on Machine Learning. Long Beach, USA, 2019: 7472-7482

[105] Wang Y, Zou D, Yi J, et al. Improving adversarial robustness requires revisiting misclassified examples//Proceedings of the 8th International Conference on Learning Representations. Addis Ababa, Ethiopia, 2020

[106] Jiang H, He P, Chen W, et al. SMART: Robust and efficient finetuning for pre-trained natural language models through principled regularized optimization//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 2177-2190

[107] Zheng R, Rong B, Zhou Y, et al. Robust lottery tickets for pretrained language models//Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Dublin, Ireland, 2022: 2211-2224

[108] Xi Z, Zheng R, Gui T, et al. Efficient adversarial training with robust early-bird tickets//Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing. Abu Dhabi, United Arab Emirates, 2022: 8318-8331

[109] Zhou Y, Jiang J Y, Chang K W, et al. Learning to discriminate perturbations for blocking adversarial attacks in text classification. arXiv preprint arXiv:1909.03084, 2019

[110] Schmidt L, Santurkar S, Tsipras D, et al. Adversarially robust generalization requires more data//Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems. Montréal, Canada, 2018: 5019-5031

[111] Wang H, Chen T, Gui S, et al. Once-for-all adversarial training: In-situ tradeoff between robustness and accuracy for free. CoRR abs/2010.11828, 2020

[112] Pang T, Lin M, Yang X, et al. Robustness and accuracy could be reconcilable by (proper) definition//Proceedings of the International Conference on Machine Learning. Maryland, USA, 2022: 17258-17277

[113] Zhang H, Cisse M, Dauphin Y N, et al. mixup: Beyond empirical risk minimization. arXiv preprint arXiv:1710.09412, 2017

[114] Li L, Qiu X. TAVAT: Token-aware virtual adversarial training for language understanding. arXiv preprint arXiv: 2004.14543, 2020

[115] Ribeiro M T, Wu T, Guestrin C, et al. Beyond accuracy: Behavioral testing of NLP models with CheckList//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 4902-4912

[116] Mishra S, Arunkumar A. How robust are model rankings: A leaderboard customization approach for equitable evaluation. CoRR abs/2106.05532, 2021

[117] Rodriguez P, Barrow J, Hoyle A M, et al. Evaluation examples are not equally informative: How should that change NLP leaderboards?//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 4486-4503

[118] Tan S, Joty S, Baxter K, et al. Reliability testing for natural language processing systems//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 4153-4169

[119] Wallace E, Rodriguez P, Feng S, et al. Trick me if you can: Human-in-the-loop generation of adversarial question answering examples. Transactions of the Association for Computational Linguistics, 2019, 7: 387-401

[120] Jia R, Liang P. Adversarial examples for evaluating reading comprehension systems//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Copenhagen, Denmark, 2017: 2021-2031

[121] Schuff H, Adel H, Vu N T. F1 is not enough! Models and evaluation towards user-centered explainable question answering //Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 7076-7095

[122] Fu J, Liu P, Zhang Q. Rethinking generalization of neural models: A named entity recognition case study//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 34. New York, USA, 2020: 7732-7739

[123] Li P, Zhang X, Jia W, et al. Active testing: An unbiased evaluation method for distantly supervised relation extraction //Findings of the Association for Computational Linguistics: EMNLP 2020. Online, 2020: 204-211

[124] Fu J, Liu P, Zhang Q, et al. RethinkCWS: Is Chinese word segmentation a solved task?//Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online, 2020: 5676-5686

[125] Zheng X, Zeng J, Zhou Y, et al. Evaluating and enhancing the robustness of neural network-based dependency parsing models with adversarial examples//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 6600-6610

[126] Belinkov Y, Bisk Y. Synthetic and natural noise both break neural machine translation. CoRR abs/1711.02173, 2017

[127] Niu X, Mathur P, Dinu G, et al. Evaluating robustness to input perturbations for neural machine translation// Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. Online, 2020: 8538-8544

[128] Stanovsky G, Smith N A, Zettlemoyer L. Evaluating gender bias in machine translation//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence, Italy, 2019: 1679-1684

[129] Choudhury M, Deshpande A. How linguistically fair are multilingual pre-trained language models?//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 35. 2021: 12710-12718

[130] Zhong R, Ghosh D, Klein D, et al. Are larger pretrained language models uniformly better? Comparing performance at the instance level//Findings of the Association for Computational Linguistics: ACLIJCNLP 2021. Online, 2021: 3813-3827

[131] Kocijan V, Camburu O M, Lukasiewicz T. The gap on gap: Tackling the problem of differing data distributions in bias-measuring datasets//Proceedings of the AAAI Conference on Artificial Intelligence: Volume 35. Virtual Event, 2021: 13180-13188

[132] Liu J, Takanobu R, Wen J, et al. Robustness testing of language understanding in task-oriented dialog//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 2467-2480

[133] Meister C, Cotterell R. Language model evaluation beyond perplexity//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers). Online, 2021: 5328-5339

[134] Hochreiter S, Schmidhuber J, et al. Long short-term memory. Neural Computation, 1997, 9(8): 1735-1780

[135] Tang D, Qin B, Feng X, et al. Effective LSTMs for target-dependent sentiment classification//Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan, 2016: 3298-3307

[136] Wang Y, Huang M, Zhu X, et al. Attention-based LSTM for aspect-level sentiment classification//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 606-615

[137] Tang D, Qin B, Liu T. Aspect level sentiment classification with deep memory network//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, USA, 2016: 214-224

[138] Ma D, Li S, Zhang X, et al. Interactive attention networks for aspect level sentiment classification. CoRR abs/1709.00893, 2017

[139] Li X, Bing L, Lam W, et al. Transformation networks for target-oriented sentiment classification//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Melbourne, Australia, 2018: 946-956

[140] Fan F, Feng Y, Zhao D. Multi-grained attention network for aspect-level sentiment classification//Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing. Brussels, Belgium, 2018: 3433-3442

[141] Zeng B, Yang H, Xu R, et al. LCF: A local context focus mechanism for aspect-based sentiment classification. Applied Sciences, 2019, 9(16): 3389

[142] Yang Z, Dai Z, Yang Y, et al. XLNet: Generalized autoregressive pretraining for language understanding//Advances in Neural Information Processing Systems: Volume 32. Vancouver, Canada, 2019

[143] Lan Z, Chen M, Goodman S, et al. ALBERT: A lite BERT for self-supervised learning of language representations// Proceedings of the International Conference on Learning Representations. Addis Ababa, Ethiopia, 2019

[144] Ribeiro M T, Singh S, Guestrin C. Why should i trust you? Explaining the predictions of any classifier//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. San Francisco, USA, 2016: 1135-1144

[145] Yang Y, Rashtchian C, Zhang H, et al. A closer look at accuracy vs. robustness//Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems. Virtual, 2020

GUI Tao, Ph. D., pre-tenured associate professor. His research interests focus on natural language processing.

XI Zhi-Heng, M. S. candidate. His research interests focus on natural language processing.

ZHENG Rui, Ph. D. candidate. His research interests focus on natural language processing.

LIU Qin, M. S. candidate. Her research interests focus on natural language processing.

MA Ruo-Tian, Ph. D. candidate. Her research interests focus on natural language processing.

WU Ting, M. S. candidate. Her research interests focus on natural language processing.

BAO Rong, Ph. D. candidate. His research interests focus on natural language processing.

ZHANG Qi, Ph. D., professor. His research interests include natural language processing, and information retrieval.

Background

In recent years, models based on deep neural networks have achieved remarkable results in almost all natural language processing tasks, even exceeding human beings in some tasks. The advent of powerful large language models has presented new opportunities and directions for the development and application of natural language processing models. However, when applied in practical applications, these models which achieve good results in the benchmark datasets are greatly diluted. Recent studies have also found that the predictions of these models may be altered by simple modifications, resulting in a dramatic decrease in performance. Even large language models can change their prediction due to minor perturbations to the input. These phenomena are related with model robustness, which usually refers to the effect of the model when it deals with new independent but similar data. For the model with high robustness, the output of the model remains the same when dealing with small changes that should not alter the output. The robustness of deep learning models is becoming a hot topic in natural language processing. There have been a large number of research works on robustness in NLP, but most of them are based on a certain task without considering the full picture. There does not yet exist a comprehensive survey of robustness in NLP at present.

In order to provide an inclusive review on the research of robustness, we introduce the latest research on robustness in NLP from four aspects: data construction, feature representation, adversarial attacks and defenses, and evaluation, as well as introduce the latest progress. As the foundation of machine learning, data construction is first explored. We include the bias in datasets and the poisoning to dataset. Feature learning converts input data into vectors and aims at representing the content of the texts as well as being independent of certain task or domain. We first introduce some deep NLP models and present different types of methods that improve the robustness of models from the aspect of model representation, including robustness encoding, knowledge Incorporation, characteristics of tasks and causal inference. Algorithms of adversarial attacks and defenses provide ways to fool models and improve the robustness of models, respectively. We then introduce mainstream adversarial attack methods including white-box, black-box and blind attack. Accordingly, a number of research works tackle the problem of adversarial defence and robustness improvement, which are concluded in the fourth part of this paper. Due to the problem of robustness, traditional metrics are no longer enough for a fair and comprehensive evaluation. A line of works propose different evaluation metrics to measure the effectiveness of models, and we introduce mainstream evaluations for both general purpose and specific tasks. Finally, the possible future research directions and considerations on robustness of natural language processing are discussed.

Our past works on robustness include *TextFlint*: *Unified Multilingual Robustness Evaluation Toolkit for Natural Language Processing* (ACL-IJCNLP 2021), *Searching for an Effective Defender*: *Benchmarking Defense against Adversarial Word Substitution* (EMNLP 2021), *Tasty Burgers, Soggy Fries*: *Probing Aspect Robustness in Aspect-Based Sentiment Analysis* (EMNLP 2020), *Flooding-X*: *Improving BERT's Resistance to Adversarial Attacks via Loss-Restricted Fine-Tuning* (ACL 2022), *Robust lottery tickets for pre-trained language models* (ACL 2022), *Efficient adversarial training with robust early-bird tickets* (EMNLP 2022).