文章编号：1003-0077(2023)02-0026-15

自然语言处理评测数据集质量评估研究

王诚文[1,2]，董青秀[1,2]，穗志方[1,2]，詹卫东[1,3]，常宝宝[1,2]，王海涛[4]

(1. 北京大学 计算语言学教育部重点实验室，北京 100871；
2. 北京大学 计算机学院，北京 100871；
3. 北京大学 中文系，北京 100871；
4. 中国标准化研究院，北京 100088)

摘 要：评测数据集是评测任务的载体，评测数据集的质量对评测任务的开展和评测指标的应用有着根本性的影响，因此对评测数据集的质量进行评估有着必要性和迫切性。该文在调研公开使用的自然语言处理主流数据集基础上，分析和总结了数据集中存在的 8 类问题，并在参考人类考试及试卷质量评估的基础上，从信度、效度和难度出发，提出了数据集评估的相关指标和将计算性与操作性相结合的评估方法，旨在为自然语言处理评测数据集构造、选择和使用提供参考依据。

关键词：自然语言处理；评测；数据集；质量评估

中图分类号：TP391 **文献标识码**：A

Quality Evaluation of Public NLP Dataset

WANG Chengwen[1,2], DONG Qingxiu[1,2], SUI Zhifang[1,2], ZHAN Weidong[1,3],
CHANG Baobao[1,2], WANG Haitao[4]

(1. MOE Key Laboratory of Computational Linguistics, Peking University, Beijing 100871, China;
2. School of Computer Science, Peking University, Beijing 100871, China;
3. Department of Chinese Language and Literature, Peking University, Beijing 100871, China;
4. China National Institute of Standardization, Beijing 100088, China)

Abstract: Pubic NLP datasets form the bedrock for NLP evaluation tasks, and the quality of such datasets has a fundamental impact on the development of evaluation tasks and the application of evaluation metrics. In this paper, we analyze and summarize eight types of problems relating to publicly available mainstream Natural Language Processing(NLP) datasets. Inspired by the quality assessment of testing in education community, we propose a series of evaluation metrics and evaluation methods combining computational and operational approaches, with the aim of providing a reference for the construction, selection and utilization of natural language processing datasets.

Keywords: natural language processing; Benchmark; dataset; quality evaluation

0 引言

近些年来，越来越多的用以支撑自然语言处理(Natural Language Processing，NLP)评测的数据集被提出来，借助于评测数据集，各种各样的自然语言处理技术和方法的性能得以有效评估[1-3]。评测数据集作为评测的一个重要组成部分，对于评测的准确性和可靠性有着至关重要的影响[4]。因此，在种类及数量众多的评测数据集如雨后春笋般涌现的同时，我们需要对当前评测数据集持谨慎态度，审视其问题和价值，进而不断完善评测数据集的提出、构建和使用方式。

综合考虑当下自然语言处理评测的现状，评测

收稿日期：2022-02-24 **定稿日期**：2022-06-16

基金项目：国家科技创新 2030“新一代人工智能”重大项目(2020AAA0106700)；国家自然科学基金(U19A2065)；中国博士后科学基金(2022M710246)

数据集相关问题研究的必要性和迫切性体现在以下四个方面：

（1）数据集的更新速度大大落后于模型的更新速度

随着深度学习技术的迅速发展，越来越多成熟的模型被提出以及开源，如自然语言处理领域的 BERT[5] 和计算机视觉领域的 ResNet[6]。这些初始模型进一步衍生了一系列新的模型，诸如 RoBERTa[7] 和 ALBERT[8] 等。模型的更新带动了相关问题处理方法的升级。然而，通过调研主流的 NLP 评测数据集，除了较少的传统评测如 TREC 每年会有一定更新，近几年新提出的数据集自提出之后，缺乏持续的探索和分析，也就少有突破性的改进。整个自然语言处理社区重视数据集的提出，但缺少对数据集的完善和与时俱进的升级。SQuAD 2.0[9] 在 SQuAD 1.0[10] 的基础上增加了没有答案的问题以增加数据集的难度，这是数据集更新方面一个较好的尝试。因此，数据集本身质量的更新对真正评测模型的语言处理能力至关重要。

（2）参差不齐的数据集不利于 NLP 的健康发展

针对同一任务的评测数据集比较多，研究者会选择对自己模型表现有利的数据集，并声称该模型具备处理该任务所必需的机器语言能力。然而，在数据集规范不统一和难度不均衡的前提下，做上述的判断是不公平的。因此，同任务数据集之间的数据质量分析是很重要的，尤其是难度的度量。LUGE 评测中的 ChnSentiCorp 和 NLPCC14-SC 同为句子级情感分类任务数据集，在排行榜上出现了在 ChnSentiCorp 上 A 模型的得分高于 B 模型，而在 NLPCC14-SC 上 A 模型的得分却低于 B 模型的情况[4]。只有通过具体分析数据集内部多维度特征，才能够找出导致模型在同任务不同数据集上表现不一致的因素。

（3）数据集规范匮乏

自然语言处理评测中存在诸多问题[4]，例如，评测缺乏规范性、评测数据偏差、评测指标失真、评测生命周期短和评测效力式微等。要想逐渐规范自然语言处理评测，必须抓住根本问题。一方面，评测数据集作为评测任务的载体，能否有效体现既定的评测任务和达到测试机器相应语言能力的作用，数据集的质量至关重要。另一方面，数据集作为评测任务开展及评测指标应用的上游环节，对后续的评测实施有着递进式的影响。因此，在诸多问题中首先来规范自然语言处理数据集有着迫切性。

（4）数据集的构建及应用未得到应有的重视

NLP 社区的研究者往往重视模型而忽视数据集。这样的一种现状很可能会导致以下后果：在可靠性及准确性差的数据集上进行模型的训练和评测，得出的结论往往也会与真实结论相差甚远。Andrew[11] 首先提出了以数据为中心的 AI 研究范式，曾指出业内大量的实践经验表明，专注于优化数据而不是模型往往可以取得更好的结果。

鉴于上述考虑，本文对自然语言处理评测数据集的问题和对策进行了归纳和探讨。首先，介绍了 NLP 评测数据集的定义和分类。其次，通过对主流数据集及相关论文的调研，本文详细阐述和归纳了 NLP 评测数据集中存在的问题。综合考虑上述分析，参照人类考试及试卷质量评价方法，从信度、效度和难度出发，提出了 NLP 评测数据集的质量评估指标和方法，以期能够为评测数据集的构造、选取和使用提供明确参考依据。

1　NLP 评测数据集概况

1.1　定义

Butterfifield 等[12] 指出，基准是一个专门设计的旨在评估系统性能的任务，通常将测得的某系统性能与经过相同基准测试的其他系统的性能进行比较来得到评测结果。评测基准一般是围绕一个或者几个评测任务展开的。评测任务一般是定义好的形式化的输入和输出，往往需要通过评测数据集来承载和体现具体的评测任务。数据集通常是输入和输出对的集合，以此来体现机器学习任务[13]。

1.2　分类

NLP 评测数据集有多种分类的视角：根据所支持的评测任务的多少可以分为单任务数据集和多任务数据集；根据数据集用途的不同，可以分为一般数据集、对抗数据集或对比数据集，例如，HANS[14] 数据集就是针对已有自然语言推理数据集中的偏差构建的对抗性测试模型性能的数据集。根据任务形式的不同，可以分为序列标注任务数据集、分类任务数据集、抽取任务数据集和生

成任务数据集。

下面将从任务形式的角度出发，分别对序列标注、分类、抽取和生成任务的主流NLP评测数据集进行介绍①。

表1 本调研选择的几个数据集

分类	子任务	数据集	提出机构	提出时间	数据规模		
					训练集	验证集	测试集
序列标注	分词	PKU	北京大学	2005	55k	—	13k
		MSRA	微软亚洲研究院	—	88k	—	13k
		AS	台湾中央研究院	—	141k	—	19k
		CITYU	香港城市大学	—	69k	—	9k
	命名实体	Sighan2006(MSRA NER)	微软亚洲研究院	2006	63K		13K
		CLUENER2020	CLUE	2020	10 748	1 343	1 345
		CoNLL2003NER	CoNLL	2003	14 987	3 466	3 684
	句法语义分析	Penn Treebank	宾夕法尼亚大学	1993	38 219	5 527	5 462
分类	自然语言推理	SNLI	斯坦福大学	2015	550 152	10 000	10 000
		MultiNLI	纽约大学	2017	393k	20k	20k
	情感分类	IMDB	斯坦福大学	2011	25 000		25 000
		SST2	斯坦福大学	2018	8 544	1 101	2 210
		Yelp Reviews	YELP	2011	650 000	—	50 000
抽取	关系抽取	TRCRED	斯坦福大学	2017	68 124	22 631	15 509
		DOCRED	清华大学NLP组	2019	3 053	1 000	1 000
	抽取式阅读理解	SQuAD	斯坦福大学	2016	87 599	10 570	—
生成	翻译	WMT 2016 News	WMT	2016	—	—	1 500
	文本摘要	CNN/Daily Mail	IBM Watson	2016	286 817	13 368	11 487
	对话	Ubuntu	NIST	2004	500组		

1.2.1 序列标注任务数据集

序列标注任务可以概括为将模型的输入(观测序列)映射为模型输出(标记序列)。典型的序列标注任务包括分词、词性标注、命名实体识别和句法语义分析。

PKU分词数据集是北京大学按照大规模现代汉语标注语料库的加工规范[15]，以人民日报语料为标注对象构造的中文分词数据集。PKU分词数据集和MSRA、AS和CITYU三个分词数据集一起构成了第二届SIGHAN国际汉语分词评测数据集[16]。

CoNLL2003[17]是2003年CoNLL会议提出的一个命名实体识别评测数据集。命名实体识别旨在机器能够识别出来一个句子中包含的指定实体及实体类型，共包括人名、地名、机构名和杂类四种实体类型。该数据集主要采用Ramshaw[18]提出的BIO的标记符号模式，共包括1 393篇英语新闻文章和909篇德语新闻文章。

PTB数据集[19]是1993年宾夕法尼亚大学提出的一个标注了词性以及短语句法结构的大规模树库。语料主要选取自《华尔街日报》。它主要包含36个POS标签和12个其他标签(用于标点符号和货币符号)。在句法结构标注中，共使用14个句法标签和4个表示空元素的标签。PTB数据集已经成为最著名的用于评估序列标记模型的数据集之一。

① 大类任务下边还会有细致的任务划分，例如序列标注任务可以细分为分词和命名实体识别等。在这里，针对每个细分任务，介绍一个代表性数据集。本研究所调研的数据集见表1。

1.2.2　分类任务数据集

分类任务即根据已经定义好的类别标签对待分类对象给出分类标签。根据类别标签的多少，可以分为二分类任务或者多分类任务。典型的分类任务包括情感分析和文本分类等。

SST[20]是斯坦福大学发布的一个情感分析数据集，主要针对电影评论来做情感分类，属于单个句子的文本分类任务。依据情感分析的粒度不同，又可以分为 SST-2 和 SST-5，前者为二分类，后者为五分类。

SNLI 是斯坦福大学 NLP 组 2015 年提出的一个自然语言推理数据集[21]。每一条样例可以视为由前提、假设和标签（蕴含、矛盾和中立）构成的三元组。主要将 Flickr 30k 数据集[22]中的图片标题作为前提句子，采用众包模式让标注者结合前提句子分别构造对应性的蕴含、中立和矛盾的句子作为假设。

MultiRC[23]是一个多项选择阅读理解数据集。每条样例由多个句子组成的段落、问题和多个候选答案组成。为了增加阅读理解文章语体的多样性，选取了小说、新闻和历史文献等 7 个不同领域的文本。目前该数据集已经被 SuperGLUE[24]作为阅读理解任务评测的数据集。该数据集为了增加阅读理解任务的难度，随机地设置每个问题的正确答案数量，从而要求机器能够独立地对每个候选答案的正确性进行判断。同时，并不要求正确答案是出现在原文中的一个片段。

1.2.3　抽取任务数据集

抽取任务的特性可以概括为从结构化或者非结构化的自然语言文本中自动抽取出事先指定好的信息。关系抽取和抽取式阅读理解是典型的抽取任务。

DocRED[25]是清华大学 2019 年提出的一个大规模篇章级关系抽取数据集。相较于句子级的关系抽取数据集，如 TACRED[26]，该数据集主要针对跨句的实体间关系进行标注。DocRED 的数据主要源自 Wikipedia 的文章和 Wikidata[27-28]的结构化数据，共包含了 5 053 篇人工标注的文章，覆盖了 132 375 个实体、56 354 种关系。除了人工标注的文章，数据集还提供了通过远程监督（Distant Supervision）方法生成的 101 873 篇文章，作为弱监督语料。相对于句子级的关系抽取数据集，在该数据集中，46.4％的实体关系涉及到了至少两个句子，40.7％的关系需要融合至少两句的信息才能得到。

SQuAD 是斯坦福大学于 2016 年推出的抽取式阅读理解数据集。数据集每一条样例都可以视为由段落、问题和答案片段构成的三元组。该数据集将阅读理解定义为抽取任务，要求机器能够在给定上下文和问题的前提下从原文中抽取能够回答问题的答案片段。SQuAD 数据集的语料主要选自 Wikipedia，共包括 500 多篇文章和 10 万多个问答配对。斯坦福大学 2018 年提出了 SQuAD 2.0 版本，在原有数据基础上增加了无法回答的问题类型以增加阅读理解数据集的难度。

1.2.4　生成任务数据集

生成任务的特性，即机器能够根据已有的输入自动生成并输出符合要求的内容。典型的生成式任务包括机器翻译、文本摘要和对话生成。

WMT 2016 News 是第一届机器翻译大会提出的一个任务[29]，旨在考察机器多语之间的翻译能力。数据集由平行的翻译句对构成，包括英语分别与捷克语、德语、芬兰语、罗马尼亚语、俄语、土耳其语构成的平行翻译句对。数据集主要选取不同语种的新闻语料。测试集是一个平行语料对的集合，包括大约 1 500 个翻译成 6 种语言（捷克语、德语、芬兰语、罗马尼亚语、俄语、土耳其语）的英语句子，以及这 6 种语言中每一种翻译成英语的额外 1 500 个句子。

CNN/Daily Mail 是 IBM 在 2016 年提出的一个多句摘要数据集[30-31]。数据集的语料来源于从美国有线新闻网（CNN）和每日邮报网（Daily Mail）收集的约 100 万条新闻数据。数据集中每条样例由一个原文段落和多句摘要构成。该数据集共有 286 817 对训练对、13 368 对验证对和 11 487 对测试对。训练集中的源文档平均有 766 个单词，平均 29.74 句，摘要平均有 53 个单词，平均 3.72 句。相较于单句摘要的数据集 Gigaword[32]和 DUC[33]来说，CNN/Daily Mail 的多句摘要更能测试机器摘要生成的能力和水平。

Ubuntu 对话库（Ubuntu Dialogue Corpus，UDC）是蒙特利尔大学 2015 年建立的公共对话数据集[34]，基于 IRC 网络上 Ubuntu 频道的对话数据和非结构化社交媒体数据建立。数据集中的一条样例可以视为由上文、回应和标签（1 或 0）构成的三元组，当 Flag＝1 时表示回答是真正的回答，当 Flag＝0 时表示是从 UDC 中随机挑选出来的回答。因此该任务旨在判断该回应是不是给定对话（上文）的下一句。UDC 1.0 版本包含约 100 万条多轮对话数据，

以及超过700万条回复和超过19亿个词。

2 NLP评测数据集存在的问题

2.1 规范性问题

数据集的规范性对于促进社区中不同研究人员的交流合作至关重要。数据集构造、加工和使用信息的记录会帮助数据集使用者选择适合自己任务的数据集。然而目前存在的一个明显现象是,在开源工作中,往往会对开源模型的适用环境和参数设置进行详细的说明,而缺乏对开源数据集更多细节性信息的介绍。Recht等[35]按照ImageNet的构造流程和方法进行数据集的复现,却发现构造出来的数据集与ImageNet有着不同的分布特征。该工作说明,数据集的规范性应该被给予足够的重视,才能够利于不同研究者进行数据复现工作,同时方便研究者在充分分析数据集细节的基础上设计有针对性的模型。Gebru[36]便强调每个数据集应该有对应的规范性数据记录文档,尤其要记录数据集创建过程中的细节。

2.2 噪声问题

数据噪声是数据集质量评估中比较重要的一个问题。随着自然语言处理评测的快速发展,越来越多的数据集被提出来。囿于多方面的成本代价,部分研究者采用半自动的方式来构造数据集,例如利用远程监督的方式构造数据集,因此数据集中会存在不同程度的数据噪声。数据集特别是测试集中的噪声将会影响评测结果的真实性和可靠性。

噪声可以大致分为表层噪声和深层噪声。表层噪声主要指形式上比较明显的标注错误,例如,结构化数据(表格)某字段下边出现不同质的值。深层噪声则指那些从形式上不能判断出来而与数据集内容有关的噪声。深层噪声的自动识别难度大,对模型的影响更大。例如,在文本摘要数据集中,噪声可能是一个不完整的或不相关的参考答案。Tejaswin等[37]人工对CNN/DailyMail、Gigaword、XSum[38]的600条随机抽取样例进行了人工标注,发现数据集中存在不同程度的实体缺失或事实缺失的例子,并总结了文本摘要数据中三种不同类型的数据噪声,具体解释如表2所示。具体例子见图1和图2。

表2 摘要数据集的噪声类型

噪声类型	解释
不完整/不相关	参考摘要(标准答案)突然结束。或者文本和参考摘要之间不相关
实体缺失	参考摘要包含源文本中没有的实体(姓名、日期、事件等)
证据缺失	参考摘要陈述的事实从源文本中无法得知

从图1可以发现,作为源标准答案给出的参考摘要与对应文本之间不相关,源文本描述的是球员的进球帮助美国队进入世界杯,而参考摘要却在说伦敦测试。从图2可以发现,摘要中出现的实体"真主党"在对应的源文本中没有出现过。

文本:安德烈-布洛姆和马克-沙伦伯格在最后10分钟内的得分和一些战术性的踢球将美国送到了橄榄球世界杯。周六,美国队以21-16战胜乌拉圭队。
摘要:伦敦测试,请忽略。

图1 Gigaword中数据噪声的例子(不相关/不完整)

文本:**美国**周二宣称对结束**以色列**和**黎巴嫩游击队**之间战斗的停火协议功不可没,并拒绝了关于美国被迫以**法国**的草案为蓝本的说法。

摘要:美国为以色列和**真主党**的停火负责。

图2 Gigaword中数据噪声的例子(实体缺失)

通过进一步分析,Tejaswin等发现Gigaword中有45%的摘要缺乏对应文本描述的实体或事实。同样的情况出现在XSum数据集中,54%的样本有实体或事实缺失的问题。

2.3 一致性问题

NLP数据集的标注一致性可以从两个角度来看,一个是标注者一致性,另一个是标注对象一致性。在数据集的构建过程中,数据构建方为了保证数据集的质量,往往会采用多人标注模式。因此,标注者之间的标注一致性在一定程度上反映了数据集的质量和该任务的可行性,常用Kappa系数[39-40]来度量。在出现环境相同的上下文中,同样的一个对象是否被标注为同样的标签,可以称之为对象标注一致性。数据集内部或者数据集(训练集、验证集和测试集)之间标注对象前后标注的不一致会在很大程度上制约模型的性能表现,特别是测试集中的标注不一致将导致模型测试结果的不可靠[41]。

在分词及词性标注任务中，人工构建的分词数据的一致性对于训练高性能的分词器至关重要。Manning[42]指出制约机器词性标注性能的瓶颈是数据集中标注结果的一致性。刘伟等[43]在基准数据集上训练并得到了效果较好的分词模型，然而在进行错误分析后发现，有 1/3 的分词错误是由于语料库分词不一致导致的。Liu 等的研究[41]发现，在中文分词评测数据中，一些处于语料库不同位置但具有相同语义的字串切分形式却不一致，具体例子如图 3 所示。他们进一步统计了中文分词评测中常用的经典数据集，发现在分词数据集的训练集和测试集中存在一定比例的分词不一致情况，尤其在 AS 的训练数据集中，分词不一致率达到了 10.72%，具体统计如表 3 所示。

它 还 向 **受害者** 提供 帮助

多数 他杀 案件 的 **受害 者** 也 是 男性

图 3　同语境下分词标注不一致的示例

表 3　分词数据集中不一致的情况统计①

数据集	统计量	AS	PKU	MSR	CITYU	CTB6
训练数据	总词数	5 449 581	1 109 947	2 368 391	1 455 630	678 811
	不一致字串数	584 048	39 095	84 143	61 432	39 716
	不一致率/%	10.72	3.52	3.55	4.22	5.85
测试数据	总词数	122 610	104 372	106 873	40 936	52 861
	不一致字串数	3 013	1 411	798	350	821
	不一致率/%	2.46	1.35	0.75	0.85	1.55

通过对 PKU 数据集进行分词一致性检验及数据修正后，Liu 等[41]基于修正一致性后的语料重新进行训练和测试。图 4 呈现了基于预训练语言模型 BERT 进行微调的 BERT-base 和仅在 BERT 的前 3 层进行微调的分词模型 BERT-prune 以及基于双向长短时记忆网络的模型 BiLSTM 在 PKU 原数据和修正分词一致性后数据上的模型表现，可以发现模型的性能分别有 1.18、1.25 和 1.04 的提升。这样的对比说明，分词数据集中的不一致是影响模型性能提升的重要因素，也是导致不能清楚评测到模型在分词任务上真实性能的原因之一。

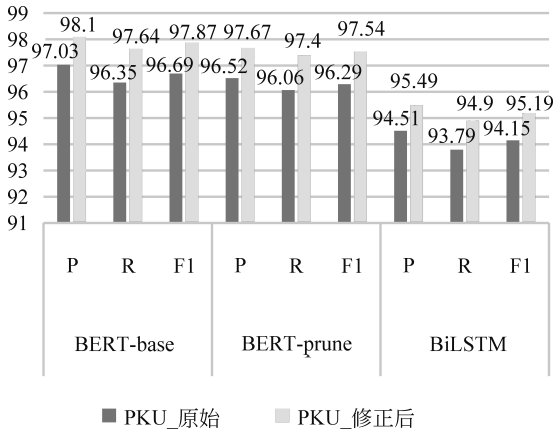

图 4　PKU 一致性修正前后模型性能对比

2.4　准确性问题

标注错误在评测数据集中很难避免，在数据集构建阶段，特别是将数据分割成多份交由不同标注者进行标注的时候，十分容易引入标注错误。评测数据集中的标注错误会严重影响评测模型的性能。尽管 CoNLL03 数据集作为一个经典的命名实体识别数据集已经被引用超过 2 300 次，Wang 等[44]在其测试集中发现了 5.38%标注错误。CoNLL03 上性能最好的模型已经取得了 F_1 值为 0.93 的表现。因此，即使标注错误只占很小的一部分，但当研究人员试图进一步提高结果时，它们也是不可忽视的[45]。不仅仅是 CoNLL03 数据集，Zeng 等[45]通过算法检测发现，在同为命名实体识别的 SCIERC 测试集中存在 26.7%的标注错误，并雇佣标注者进行数据修正（实体边界或者实体类型校正），选取 BiLSTM-CRF 和 LM-BiLSTM-CRF 模型在修复前后数据集上进行对比训练，发现基于修正后数据训练的模型性能都有一定程度的提高，分别为 0.85 和 1.91。

① 该表引自文献[41]。

2.5 均衡性问题

在当下的NLP评测中，研究者多少有这样的共识，即模型在某类型任务的一个或多个数据集上表现不错，就说明该模型具备解决该任务的能力。这种认识在一定程度上取决于数据集能多大程度上代表某任务。

Schlegel等人[46]的研究发现，号称阅读理解任务的数据集NEWSQA中存在严重的考察能力失衡现象。他们随机抽取了50条数据，人工标注回答阅读理解问题所需能力类型，分别为片段检索、复述、无法回答和抽象提炼，具体统计结果如表4所示。可以发现，大部分都在考察机器从原文检索正确片段的能力，而忽视了对机器复述和抽象提炼能力的考察。因此，很难说该数据集全面均衡考察了机器的阅读理解能力。

表4 NEWSQA问题考察能力类型（抽样样本）

阅读理解能力类型	数量
片段检索	38
复述	0
无法回答	12
抽象提炼	0

与此同时，我们对WeiboNER数据集中不同类型的实体数量进行统计，如图5所示。我们发现“PER”标签的出现比例占了总体的71%，其他类型的标签则占比较少。因此，该数据集对人名之外其他类命名实体识别能力的考察还有待增强。

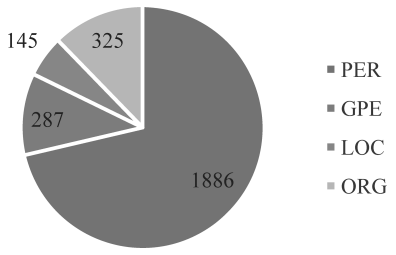

图5 WeiboNER中不同实体的分布

2.6 偏差问题

尽管深度学习模型似乎在人工智能具有挑战性的任务中取得了显著成绩，但最近的研究[47-49]表明，这些成绩的提高可能在很大程度上是由于“廉价的把戏”，而不是依靠像人类一样的推理能力。在机器学习领域，偏差（Bias）的存在会让机器利用这些捷径以做出正确的判断。

在多种自然语言处理评测任务的数据集中，都有偏差因素的存在。Sugawara等人[50]的工作表明，在被扣掉某些字符（乱码做替换）的机器阅读理解数据集的实例上训练的模型仍然可以得到正确答案。在自然语言推理数据集中[51]，已有的研究发现，机器可能单一凭借假设句子中某个词语和判断标签之间的一致性关系，做出推理而不用理解前提句子的语义。例如，机器一看到假设句子中出现not就给出矛盾判断，而机器的判断恰好是正确答案。Cai等人的研究表明[52]，在故事完形填空任务（选出合理的故事结尾）中，某些偏见是普遍存在的，它允许模型只对结尾部分进行训练而不是从故事开头部分开始，以产生最优的结果。无论是数据本身还是人工标注过程中注入的偏差，都会降低数据集的效度。在这种情况下，到底模型是否掌握了处理某种语言任务的能力，需要打上一个问号。

2.7 数据集垄断问题

Koch等人[53]通过基尼系数来计算不同自然语言处理社区内数据集使用的集中程度（包括在论文、评测或网站中使用的数据集），发现自2015—2020年，数据集使用集中程度越来越凸显（基尼系数有0.113的提升）。他们的研究同时发现，那些广泛使用的数据集仅由少数精英机构提出。事实上，截至2021年6月，PWC① 上使用量超过50%的数据集可归于12家机构。

尽管越来越多的研究学者或机构提出新的数据集，然而他们被使用或者用来做实验对比研究的比例却很少。这样一种过度集中使用少数数据集的情况不利于整个NLP社区的多样性和繁荣发展。特别是已经有研究[54]表明，目前比较经典的评测平台的数据集中存在多种多样的问题，其质量并没有得到很好的保证。

2.8 数据集重视程度问题

Sambasivan等[55]在一项AI从业者调研中发现，92%的人经历过一次或多次的由于数据问题导致的错误级联问题。级联在模型的下游产生了重大的负面影响，如昂贵的迭代、放弃项目以及对社区造成严重的伤害。通过有意的做法，级联在很大程度上是可以避免的。尽管数据集对于整个NLP研究

① Papers With Code (PWC) corpus.

环节如此重要，研究者则更加倾向于模型的研究工作而不重视数据集本身的研究工作。

3　数据集评估指标和方法

经典测试理论（Classical Testing Theory，CTT）是人类考试及试卷质量评估中常用并被证明行之有效的测量理论[56]。在试卷评估中，往往从信度、效度和难度三个维度来进行考量。参照人类考试及试卷的评估经验，我们将上述数据集中存在的问题在三个维度下进行归类，并尝试制定对应性的层级化数据集评估指标和计算性与操作性结合的评估方法。

在信度下，制定规范度、准确度和一致度三个一级指标。在效度下，有均衡度、契合度和偏差度三个一级指标。在难度下，共包括难易度、区分度和更新度三个一级指标。在一级指标下制定对应的二级指标。

3.1　规范度（Degree of Standardization，DS）

3.1.1　数据泄露比（Data leakage Ratio，DLR）

指标含义：同时出现在训练集和测试集中的样本比例。过高的比例可能会导致机器只凭借“记忆力”来完成测试。

评价方法：

（1）统计训练集样本数量（C_{train}）；（2）统计测试集样本数量（C_{test}）；（3）统计出现在训练集中的测试集样本数量（$C_{\text{train}} \cap C_{\text{test}}$）。

$$\text{DLR} = \frac{C_{\text{train}} \cap C_{\text{test}}}{C_{\text{test}}} \tag{1}$$

指标标准：DLR 应该尽可能小。

3.1.2　数据缺失值比（Missing Values Ratio，MVR）

指标含义：该指标一般适用于结构化的数据（表格），衡量样本在某些特征上值为空的比例。

评价方法：

（1）统计不同样本在不同特征下的所有值数量（C_{both}）；

（2）统计所有值中为空的数量（C_{null}）。

$$\text{MVR} = \frac{C_{\text{null}}}{C_{\text{both}}} \tag{2}$$

指标标准：MVR 应该尽可能小。

3.1.3　数据重复比（Repeated Sample Ratio，RSR）

指标含义：主要指数据中样本重复的情况，该种重复可能由于数据获取的源头较多，最后未能对数据进行有效的去重。

评价方法：

（1）统计数据集样本的总数量（C_{both}）；

（2）统计重复样本的数量（$C_{\text{duplicate}}$）。

$$\text{RSR} = \frac{C_{\text{duplicate}}}{C_{\text{both}}} \tag{3}$$

指标标准：RSR 应该尽可能小。

3.1.4　数据异常值比（Data Outliers Ratio，DOR）

指标含义：数据集中明显成偏态分布的样本的占比。

评价方法：采用统计分析方法侦测数据集中偏态分布的离群点。

指标标准：DOR 应该尽可能小。

3.1.5　元数据完整度（Metadata Integrity，MI）

指标含义：数据集元数据的完整程度。数据集元信息主要包括数据集的来源、构造方式、变换记录以及数据集细粒度的分布状态的统计信息。

评价方法：检查数据集是否有配套的介绍数据集元数据的文档。

指标标准：配备元数据记录的评测数据集在评价中会给予更高的分数。

3.2　准确度（Accuracy）

指标含义：数据集标注信息是否符合既定的标注原则和方案，是否通过专家的标注质量认定。

评价方法：

（1）核验数据集提供方是否报告了数据集标注正确性的指标；

（2）数据集评估方可采用抽样评测的方式，对抽样数据集进行人工的认定和评估。数据集的抽样准确率为：抽样正确数/抽样样本数。进而，可以用抽样样本的准确率来近似估计整个数据集的准确率。

指标标准：数据集的准确率应该高。

3.3　一致度（Consistency）

3.3.1　标注者一致度（Inter-Annotator Agreement，IAA）

指标含义：不同标注者在同一标注对象上的标注一致程度。对于多人标注的情况，可以分解为两两标注，并取其平均水平进行衡量。

评价方法：采用 Kappa 一致性检验来计算标注者一致度，计算如式（4）所示。

$$\text{Kappa} = \frac{p_{\text{o}} - p_{\text{e}}}{1 - p_{\text{e}}} \tag{4}$$

其中，p_o 表示实际观察到的一致性，p_e 为标注者任意标注的一致性。Kappa 的含义是去除了由任意标注产生的一致性，才是准确的一致性。

指标标准：Kappa 值尽可能高，值越高表示数据集的标注质量越高。

3.3.2 对象标注一致度(Consistency of Object Annotation，COA)

指标含义：数据集中语境相同下同标注对象标注一致程度。

评价方法：采用统计及机器学习方法，自动捕获语境高度相同及相近条件下同字符串标注不一致情况(C_{disagree})，相同字符串出现的次数(C_{both})，进而标注一致度为：

$$\text{COA}=1-\frac{C_{\text{disagree}}}{C_{\text{both}}} \tag{5}$$

示例："我们能**做好**那件事，但是他们不一定能**做好**那件事。**做好**做不好，取决于大家的主观态度"。

在该分词标注例子中，同样的字符串"做好"三次出现的，在语义上没有明显的区别，应该做同样的一种切分。根据统计结果可以知道：

$$\text{COA}(\text{做好})=1-\frac{1}{3} \tag{6}$$

指标标准：COA 越高代表标注一致性越好。

3.3.3 模型表现一致度(Model Achievement Consistency，MAC)

指标含义：同样一个模型在分布一致但是不相同的测试集上的表现一致程度。

评价方法：

(1) 同分布的不同测试集：$\{\text{test}_1,\text{test}_2,\cdots,\text{test}_i\}$；

(2) 同样的模型：M_1；

(3) 模型在不同测试集上得分：$\{S_1,S_2,\cdots,S_i\}$；

(4) 离散系数函数 F。

MAC 的计算如式(7)所示。

$$\text{MAC}=F(S_1,S_2,\cdots,S_i) \tag{7}$$

指标标准：离散系数越小，模型在数据集上表现一致度越高，说明数据集的可信性。

3.3.4 内容与原则一致度(Content-Principle Consistency，CPC)

指标含义：作为内容的标注数据集与指导性的原则和方案之间的一致程度。

评价方法：专家评估法，即专家采用抽样评估的方式来判断标注内容是否有效贯彻了既定的标注方案和标注原则。

指标标准：CPC 尽可能越高，表示内容和原则一致度越强。

3.4 均衡度(Uniformity)

指标含义：同数据集中不同类型标签和考察细粒度语言能力的均匀程度。

评价方法：

(1) 容易从形式标记入手的，可以采用自动方式。例如，对命名实体识别数据集中不同类型的实体可以直接统计其各自类型数量，来判断其分布的均匀程度。

(2) 涉及深层内容的，可以采用人工采样标注的方式。例如，对于阅读理解数据集来说，可以采样标注样本考察的阅读理解能力子类型，然后进行统计，看数据集中考察到的各种阅读理解子能力是否是均匀的。

指标标准：数据集应该具备较高的代表性，数据比例应该尽量均匀。

3.5 契合度(Integration Degree，ID)

指标含义：数据集是否有效体现了所定义的考察机器对应语言能力的任务，以及体现的程度。

评价方法：

① **专家评判法** 专家可以采用抽样的方式来对数据集能否达到对应测试机器语言能力的程度做出评估。

② **自论证法** 数据提供方能够结合随机样例论证其任务效度，并提供论证数据。

③ **模型评测法** 将同样的一个数据集从结构或者内容角度进行解耦模块化，如对于阅读理解任务的测试集进行改造，一个版本的测试集包括完整的段落、问题和答案三元组，另外一个版本的测试集的段落则只保留段落尾句。将不同版本的数据集分别输入给一个结构和参数设置固定不变的模型，看模型的结果有何区别。如果模型的性能没有较大的区别，则在某种程度上说明数据集未能有效建模阅读理解任务。示意见图 6。

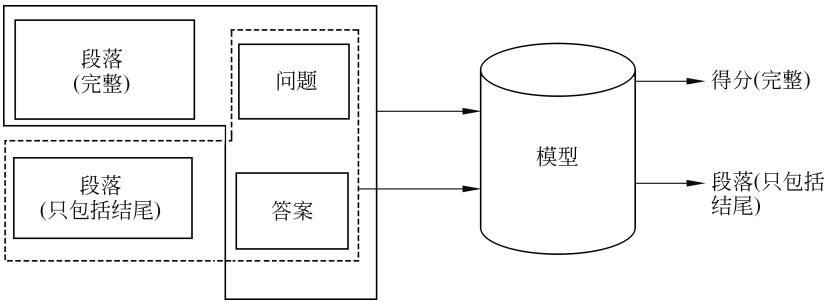

图 6　数据-模型契合度模型评测法示意图

指标标准：数据集的契合度应该较高，较低的契合度说明数据集没能达到体现评测任务的作用。

3.6　偏差度(Bias)

3.6.1　类型分布偏差(Type Distribution Bias，TDB)

指标含义：同一评测数据集中多种类型数据分布情况的差异程度。

评价方法：

① **分布统计法**　对于那些形式化特征比较明显的不同类型数据，可以统计不同类别数据出现的数量，然后统计分析不同类型样本的分布情况，例如可以用方差或者变异系数等统计指标来衡量一组数据的分布离散程度。例如，对于机器阅读理解的问题类型分布统计，一般会有明显的疑问词语“5w1h”，就可以统计数据集中疑问代词的数量来估计数据集的问题分布情况。

② **人工抽样法**　对于一些比较难利用形式化特征来统计类型分布的数据集，可以考虑采用抽样人工标注的方法进行数据类型分布情况的确认。例如，对于自然语言推理数据集来说，样本所考察的不同推理能力(空间推理、亲属关系推理等)的确立没有形式化特征便于利用，便可以人工进行加工确认，再统计其类型分布的情况。

指标标准：类型分布偏差尽可能小，各类型分布均匀，不应该过度偏向其中的某个类型。

3.6.2　答案分布偏差(Answer Distribution Bias，ADB)

指标含义：数据集中正确答案分布均匀与否的程度。这种答案分布的偏差容易被机器作为“捷径”利用，从而在评测数据集上取得较好表现。

评价方法：利用多数类基线(始终选择多数正确答案所在的选项)的结果来检测数据集中是否存在偏差以及偏差的程度，数值越小说明答案分布越均衡。

指标标准：应该保证答案分布偏差尽可能小。

3.6.3　无关线索分布偏差(Irrelevant Cue Distribution Bias，ICDB)

指标含义：数据集中是否存在无关线索的偏差以及无关线索偏差的程度。无关线索偏差一般指机器并非通过掌握所要评测的语言能力，而仅仅利用一些无关的线索推出正确答案的情况。

评价方法：以自然语言推理任务为例，可以统计前提假设句中词汇、短语句法语义特征与分类标签之间的一致性关系，通常使用的指标有 PMI、重复比(OverlapRate)或相似度(Similarity)。

$$\mathrm{PMI}(\text{词语},\text{标签})=\log\frac{p(\text{词语},\text{标签})}{p(\text{词语})p(\text{标签})} \tag{8}$$

$$\mathrm{OverlapRate}=\frac{\text{重复词语或 } n\text{-gram 的数量}}{\text{样本中词语或 } n\text{-gram 数量}} \tag{9}$$

$$\mathrm{Similarity}=\mathrm{Sim}(\text{前提},\text{假设}) \tag{10}$$

指标标准：度量 ICDB 的几个指标的值越小，说明无关线索分布偏差程度越低。

3.7　难易度(Difficulty)

3.7.1　文本特征难度(Text Feature Difficulty)

指标含义：文本特征难度是针对数据集文本难度的度量，一般从词汇、短语、句子和语篇几个侧面度量文本的特征难度。

评价方法：采用线性加权计算公式来衡量文本特征的难度，文本难度$=\alpha_1$ 词汇$+\alpha_2$ 短语$+\alpha_3$ 句子$+\alpha_4$ 语篇。

例如，衡量文本词汇丰富性通常用的指标为词语类符形符比(TTR，Type Token Ratio)：

$$\mathrm{TTR}=\frac{\mathrm{Freq}(\mathrm{typenumberofwords})}{\mathrm{Freq}(\mathrm{tokennumberofwords})} \tag{11}$$

指标标准：文本特征难度适中，不能过于简单或者过难。

3.7.2　人机差异度(Human-Machine Gap，HMG)

指标含义：同数据集上，人类基准与模型得分

之间的差距程度。

评价方法：

(1) 在数据集上开展人工评测，得到科学的人类基准 $\text{Score}_{\text{human}}$；

(2) 获取数据集提出时 baseline 模型的得分 $\text{Score}_{\text{machine}}$，如式(12)所示。

$$\text{HMG} = \text{Score}_{\text{human}} - \text{Score}_{\text{machine}} \tag{12}$$

指标标准：HMG 应该相对较大，以凸显任务的难度。

3.8 区分度(Distinctness)

指标含义：同数据集上不同模型得分之间的区分度。不同性能模型的表现应该在数据集上有着较大的差别，也就是说分布较为离散。

评价方法：用离散系数(Coefficient of Variation)来衡量同数据集上不同模型得分的离散化程度。离散系数的计算如式(13)所示。

$$c_u = \frac{\sigma}{\mu} \tag{13}$$

其中，σ 为标准差；μ 为均值。

指标标准：离散系数应该相对较大。

3.9 更新度(Dynamic Update Degree，DUD)

3.9.1 生命周期(Life Cycle，LC)

指标含义：数据集提出时间到数据集上模型得分逼近人类基准的时间周期。

评价方法：

(1) 获取该数据集上逼近人类基准模型提出的时间 $\text{Time}_{\text{sota}}$；

(2) 获取数据集提出的时间 $\text{Time}_{\text{base}}$；

$$\text{LC} = \text{Time}_{\text{sota}} - \text{Time}_{\text{base}} \tag{14}$$

指标标准：生命周期不宜过短。

3.9.2 更新程度(Extent Of Update，EOU)

指标含义：该数据是否有更新的版本，包括数据集使用的语料、标注信息等。

评价方法：追踪该数据集提出者最新的关于该数据集的使用协议或者相关论文，核查是否有最新的更新说明等。

指标标准：配备数据更新记录的数据集在评价中会被给予更高分数。

4 指标验证

如上文所提及，目前制定的指标不能完全实现自动评估。对于部分指标来说，需要采用可操作性强的人工评估方法，例如，规范度中的元数据完整度和更新度中的数据集生命周期及更新程度指标。考虑到评测数据集偏差因素对评测效度的关键性影响，本部分结合中文空间语义理解评测任务数据集[①]对偏差度进行验证。与此同时，鉴于文本特征难度对于各类评测数据集的通用性，也将探索难度维度下文本特征难度对空间语义理解评测任务的影响。

4.1 实验数据

本研究选取中文空间语义理解评测任务中的文本空间语义异常判断数据集作为实验对象。该任务要求机器判断给定的中文文本是否存在空间关系异常，经判定后不存在异常的句子标记为“True”，存在异常的句子标记为“False”。具体例子如下(文中加下划线部分为空间语义表达主要成分)：

例 1：当我们到达医院时，一大群人已围在外面。地下室里，一条长廊通向手术室。我们在远处发现了外婆，妈妈和继父连忙跑上前去，我和哥哥走在后面，互相搀扶着，极力使自己镇静下来。(True)

例 2：入夜之后，小舟转向东南。在海中航行了三日，**小船顶**只有些干粮清水，石破天和那船夫分食。到第四。午间，屈指正是腊月初八，那汉子指着前面一条黑线，说道：“那便是侠客岛了。(False)

在例 2 中，“小船顶”语义异常，一般为“小船里”或“小船中”。

这部分数据的规模如表 5 所示。

表 5 中文空间语义正误判断数据集规模

类型	规模
训练集	4 237
验证集	806
测试集	794

在具体指标验证中，主要选取测试集作为研究对象。

4.2 偏差度验证

该数据集主要通过替换句中表空间方位相关词语的方式批量生成。替换后的句子，可能存在错误

① http://ccl.pku.edu.cn:8084/SpaCE2021/

的空间关系信息，计算机需要甄别哪些语段的空间关系信息是正确的，哪些是错误的。通过进一步统计，发现其中替换对存在不同程度的答案分布偏差，即部分原词替换成某词后所对应一系列句子的空间语义存在过度偏向正确或错误的分布状况。例如，方位词“中”替换成“底”构成的替换对“中→底”所对应的句子共 8 条，其中 1 条句子所对应的标准答案为“True”，其他 7 条均为“False”。这样导致“中→底”替换对所形成句子的正误存在一定的分布偏差。表 6 列出了部分替换对的分布偏差。

表 6　替换对答案分布偏差

替换对	True 数量	False 数量	偏差程度
中→顶	0	10	5
上→边	1	5	4
中→上	7	3	3

为了进一步探讨这种替换对的答案分布偏差是否会对模型的成绩造成影响，以及会造成哪种影响。我们依据替换对答案分布偏差的情况，将偏差度分为 1～5 等级，数字越大表示偏差程度越高。然后根据偏差度等级将测试集中 794 条样本分为若干组，在不同组样本上统计模型预测的精确率。为了验证模型预测与答案分布偏差的关系，我们选取参加本次评测的前 6 名队伍所提交的模型作为实验模型，在后续描述中，根据评测中模型的排名命名为 $model_1$ 到 $model_6$。然后，采用斯皮尔曼等级相关系数统计不同模型精确率与偏差程度之间的相关性，其形式化描述如式(15)所示。

$$S = \text{Spearman}(\text{Acc}_i, \text{BiasGrade}) \quad (15)$$

其中，Acc_i 表示模型在答案偏差度不同组样本上的预测精确率，BiasGrade 表示偏差程度等级。具体实验结果如表 7 所示。

表 7　答案偏差度与模型预测相关性分析

模型	r	p
Model_1	1	1.40E-24
Model_2	1	1.40E-24
Model_3	0.9	3.74E-02
Model_4	0.9	3.74E-02
Model_5	0.8	5.39E-02
Model_6	0.9	3.74E-02

在表 8 中，r 值为指标之间的相关性系数，p 值为显著性系数。只有当 $p<0.05$ 时，才能说明两个指标之间具有显著的相关关系。通过实验可以发现，除了 Model1_5 之外，其余模型的预测精确率都与偏差等级成显著正相关，即答案分布偏差程度越高，模型预测的精确率越高。

4.3　文本特征难度验证

本部分主要探究文本词汇丰富度与模型预测精确率之间的关系。在中文空间语义正误判断数据集中，文本词汇的丰富度主要表现为表示空间语义的方位词语的丰富度，我们定义了一个方位词密度指标 Loc_word_density，即句子中方位词语数与句子总词数的比值。按照样本方位词密度大小将 794 个测试集样本分为若干组，然后统计模型在不同组上的预测精确率，进而计算模型预测成绩与方位词密度之间的相关性，形式化表述如式(16)所示。

$$S = \text{Sperman}(\text{Acc}_j, \text{Loc\_word\_density}) \quad (16)$$

其中，Acc_j 表示模型在依据 Loc_word_density 指标分成小组样本上的预测精确率。具体统计结果如表 8 所示。

表 8　方位词语密度与模型预测的相关性分析

模型	r	p
Model_1	−1	0.00E+00
Model_2	−1	0.00E+00
Model_3	−1	0.00E+00
Model_4	−0.87	3.33E-01
Model_5	−0.87	3.33E-01
Model_6	−1	0.00E+00

从上表中可以发现，$p<0.05$ 的模型有 Model_1、Model_2、Model_3 和 Model_6。大部分模型的结果表示模型预测精确率与方位词密度之间呈负相关，也就是说，随着样本的方位词词语密度的增加，模型的预测精确率下降。

5　总结

本文从不同任务角度出发，介绍了主流自然语言处理评测数据集，并在充分调研主流评测数据集的基础上归纳了评测数据集中存在的 8 类问题——规范性问题、噪声问题、一致性问题、准确性问题、均衡性问题、偏差问题、垄断问题和重视程度问题。在问题分析的基础上，参照人类考试及试卷质量评估

的经验，分别从信度、效度和难度出发尝试提出了评测数据集质量评估的指标和方法，如表 9 所示。与此同时，也通过选取中文空间语义理解评测数据集对偏差度和文本特征难度两个指标进行了初步验证，实验结果表明该指标能够有效评估数据集的质量。

总体来说，这些评测指标的提出还很初步，还有待进行更详细的实验验证，今后还需要在评估指标和方法上进行更加精细化、科学化和自动化的研究，希望在此研究的启发下有更多的研究者开展关于数据集评估方面的研究，从而为自然语言处理评测数据集的构造、选择和使用提供参考依据。

表 9 评估指标与数据集问题对照表

维度	一级指标	二级指标	数据集问题
信度	规范度	数据泄露比	规范性问题
		数据缺失值比	
		数据重复比	
		数据异常值比	
		元数据完整度	
	准确度	—	准确性问题
	一致度	标注者一致度	一致性问题
		对象标注一致度	
		模型表现一致度	
		内容与原则一致度	
效度	均衡度	—	均衡性问题
	契合度	—	噪音问题
	偏差度	类型分布偏差	偏差问题
		答案分布偏差	
		无关线索分布偏差	
难度	难易度	文本特征难度	—
		人机差异度	—
	区分度	—	—
	更新度	生命周期	—
		更新程度	—

参考文献

[1] SCHLANGEN D. Targeting the Benchmark: On methodology in current natural language processing research[C]//Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing. 2021: 670-674.

[2] GEHRMANN S, ADEWUMI T, ZHOU J. The GEM Benchmark: Natural language generation, its evaluation and metrics[C]//Proceedings of the 1st Workshop on Natural Language Generation, Evaluation, and Metrics, Bangkok, Thailand Association for Computational Linguistics, 2021: 96-120.

[3] KIELA D, BARTOLO M, NIE Y, et al. Dynabench: Rethinking benchmarking in NLP[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2021: 4110-4124.

[4] 董青秀，穗志方，詹卫东，等. 自然语言处理评测中的问题与对策[J]. 中文信息学报，2021，35(6)：1-15.

[5] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2019: 4171-4186.

[6] HE K , ZHANG X , REN S , et al. Deep residual learning for image recognition[C]//Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2016: 770-778.

[7] LIU Y, OTT M, GOYAL N, et al. Roberta: A robustly optimized bert pretraining approach[J]. arXiv preprint arXiv: 1907.11692, 2019.

[8] LAN Z, CHEN M, GOODMAN S, et al. Albert: A lite bert for self-supervised learning of language representations[C]//Proceedings of ICLR. 2020: 1-17.

[9] RAJPURKAR P, JIA R, LIANG P. Know what you don't know: Unanswerable questions for SQuAD[C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. 2018: 784-789.

[10] RAJPURKAR P, ZHANG J, LOPYREV K, et al. SQuAD: 100,000+ questions for machine comprehension of text[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2016: 2383-2392.

[11] ANDREW NG. Mlops: From model centric to data-centric ai.[EB/OL]. https: //www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf[2022-01-02].

[12] ANDREW B, GERARD E N, ANNE K E,. A dictionary of computer science[M]. London: Oxford University Press, 2016.

[13] PAULLADA A, RAJI I D, BENDER E M, et al. Data and its (dis) contents: A survey of dataset development and use in machine learning research[J]. Patterns, 2021, 2(11): 100336.

[14] MCCOY T, PAVLICK E, LINZEN T. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019: 3428-3448.

[15] 俞士汶，朱学锋，段慧明.大规模现代汉语标注语料库的加工规范[J].中文信息学报，2000(06)：58-64.
[16] EMERSON T. The 2nd international Chinese word segmentation bakeoff[C]//Proceedings of the 4th SIGHAN Workshop on Chinese Language Processing. 2005：123-131.
[17] SANG E F，DE MEULDER F. Introduction to the CoNLL-2003 shared task：Language-independent named entity recognition[C]//Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL，2003：142-147.
[18] RAMSHAW L A，MARCUS M P. Text chunking using transformation-based learning[M]. Natural Language Processing Using Very Large Corpora. Dordrecht：Springer，1999：157-176.
[19] MARCUS M. Building a large annotated corpus of English：the penn treebank[J]. Computational Linguistics，1993，19(2)：313-330.
[20] SOCHER R，PERELYGIN A，WU J，et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2013：1631-1642.
[21] BOWMAN S R，ANGELI G，POTTS C，et al. A large annotated corpus for learning natural language inference[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing，Association for Computational Linguistics，2015：632-642.
[22] YOUNG P，LAI A，HODOSH M，et al. From image descriptions to visual denotations：New similarity metrics for semantic inference over event descriptions[J]. Transactions of the Association for Computational Linguistics，2014，2：67-78.
[23] KHASHABI D，CHATURVEDI S，ROTH M，et al. Looking beyond the surface：A challenge set for reading comprehension over multiple sentences[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics：Human Language Technologies，2018：252-262.
[24] WANG A，PRUKSACHATKUN Y，NANGIA N，et al. Superglue：A stickier benchmark for general-purpose language understanding systems[C]//Proceedings of the 33rd Conference on Neural Information Processing Systems，2019：1-15.
[25] Yao Y，Ye D，Li P，et al. DocRED：A large-scale-document-level relation extraction dataset[C]//Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. 2019：764-777.
[26] ZHANG Y，ZHONG V，CHEN D，et al. Position-aware attention and supervised data improve slot filling[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2017：35-45.
[27] ERXLEBEN F，GÜNTHER M，KRÖTZSCH M，et al. Introducing Wikidata to the linked data web[C]//Proceedings of the International Semantic Web Conference. Springer，Cham，2014：50-65.
[28] VrandeČiĆ D，Krötzsch M. Wikidata：A free collaborative knowledgebase[J]. Communications of the ACM，2014，57(10)：78-85.
[29] Bojar O，Chatterjee R，Federmann C，et al. Findings of the conference on machine translation[C]//Proceedings of the 1st Conference on Machine Translation，2016：131-198.
[30] Nallapati R，Zhou B，dos Santos C，et al. Abstractive text summarization using sequence-to-sequence RNNs and beyond[C]//Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning. 2016：280-290.
[31] Nallapati R，Zhai F，Zhou B. SummaRuNNer：A recurrent neural network based sequence model for extractive summarization of documents[C]//Proceedings of the 31st AAAI Conference on Artificial Intelligence. 2017：3075-3081.
[32] Rush A M，Chopra S，Weston J. A neural attention model for abstractive sentence summarization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2015：379-389.
[33] Paul O. Introduction to DUC-2001：An intrinsic evaluation of generic news text summarization systems[C]//Proceedings of DUC Document Understanding Conference. 2001：49.
[34] Lowe R，Pow N，Serban I V，et al. The ubuntu dialogue corpus：A large dataset for research in unstructured multi-turn dialogue systems[C]//Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue. 2015：285-294.
[35] RECHT B，ROELOFS R，SCHMIDT L，et al. Do imagenet classifiers generalize to imagenet? [C]//Proceedings of the International Conference on Machine Learning. PMLR，2019：5389-5400.
[36] GEBRU T，MORGENSTERN J，VECCHIONE B，et al. Datasheets for datasets[J]. Communications of the ACM，2021，64(12)：86-92.
[37] TEJASWIN P，NAIK D，LIU P. How well do you know your summarization datasets? [C]//Proceedings of the Findings of the Association for ComputationalLinguistics. 2021：3436-3449.
[38] NARAYAN S，COHEN S B，LAPATA M. Don't give me the details，just the summary! Topic-aware convolutional neural networks for extreme summarization[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2018：1797-1807.
[39] VÉRONIS J. A study of polysemy judgements and inter-annotator agreement[G]. Programmed and Advanced Papers of the Sense Val Workshop. 1998：2-4.
[40] Artstein R，Poesio M. Inter-coder agreement for computational linguistics[J]. Computational Linguistics，2008，34(4)：555-596.
[41] 刘伟，黄锴宇，余浩，黄德根.基于语境相似度的中文分词一致性检验研究[J].北京大学学报(自然科学

版).2022,58(1): 99-107.
[42] MANNING C D. Part-of-speech tagging from 97% to 100%: Is it time for some linguistics? [C]//Proceedings of the International Conference on Intelligent text Processing and Computational Linguistics. Springer, Berlin, Heidelberg, 2011: 171-189.
[43] MA J, GANCHEV K, WEISS D. State-of-the-art Chinese word segmentation with Bi-LSTMs[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing. 2018: 4902-4908.
[44] WANG Z, SHANG J, LIU L, et al. Crossweigh: Training named entity tagger from imperfect annotations[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing, 2019: 5154-5163.
[45] ZENG Q, YU M, YU W, et al. Validating label consistency in NER data annotation[C]//Proceedings of the 2nd Workshop on Evaluation and Comparison of NLP Systems. 2021: 11-15.
[46] SCHLEGEL V, VALENTINO M, FREITAS A, et al. A Framework for evaluation of machine reading comprehension gold standards [C]//Proceedings of the 12th Language Resources and Evaluation Conference. 2020: 5359-5369.
[47] GEVA M, GOLDBERG Y, BERANT J. Are we modeling the task or the Annotator? An investigation of annotator bias in natural language understanding datasets[C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 1161-1166.
[48] ZHOU X, BANSAL M. Towards robustifying NLI models against lexical dataset biases [C]//Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. 2020: 8759-8771.
[49] SAXON M, WANG X, WANG W Y. Automatically Identifying Semantic Bias in Crowdsourced Natural Language Inference Datasets[J]. arXiv preprint arXiv: 2112.09237, 2021.
[50] SUGAWARA S, STENETORP P, INUI K, et al. Assessing the benchmarking capacity of machine reading comprehension datasets[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2020, 34(05): 8918-8927.
[51] GURURANGAN S, SWAYAMDIPTA S, LEVY O, et al. Annotation artifacts in natural language inference data[C]//Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 2018: 107-112.
[52] CAI Z, TU L, GIMPEL K. Pay attention to the ending: Strong neural baselines for the roc story cloze task[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. 2017: 616-622.
[53] KOCH B, DENTON E, HANNA A, et al. Reduced, reused and recycled: The life of a dataset in machine learning research [C]//Proceedings of the 35th Conference on Neural Information Processing Systems Datasets and Benchmarks Track.
[54] RAJI I D, BENDER E M, PAULLADA A, et al. AI and the everything in the whole wide world benchmark[J]. arXiv preprint arXiv: 2111.15366, 2021.
[55] SAMBASIVAN N, KAPANIA S, HIGHFILL H, et al. "Everyone wants to do the model work, not the data work": Data cascades in high-stakes AI[C]//proceedings of the CHI Conference on Human Factors in Computing Systems. 2021: 1-15.
[56] NOVICK M R. The axioms and principal results of classical test theory[J]. Journal of Mathematical Psychology, 1966, 3(1): 1-18.

王诚文(1992—),博士,助理研究员,主要研究领域为自然语言处理评测和语言工程。
E-mail: wangcw@pku.edu.cn

穗志方(1970—),通信作者,博士,教授,主要研究领域为计算语言学和知识工程。
E-mail: szf@pku.edu.cn

董青秀(1998—),博士研究生,主要研究领域为自然语言处理。
E-mail: dqx@stu.pku.edu.cn