机器阅读理解评测信度与效度研究

Research on the Reliability and Validity of Machine Reading Comprehension Evaluation

“语法理论与语言工程”讨论班
李楠
2021-07-08

摘要

Abstract

- 机器阅读理解评测是评测机器机器阅读理解能力的一项自然语言处理任务，对于如何提升这项评测任务的可靠性和有效性还缺乏深入研究。
- 信度、效度分析是心理学领域经典测验理论下针对测验可靠性和有效性的分析方法。
- 本文在广泛调研现有中英文机器阅读理解数据集和学界已有研究的基础上，将经典测验理论迁移到机器阅读理解评测上，讨论对机器阅读理解评测进行信度、效度分析的可行性，并依次考察对应的检测手段。
- 本文还讨论了机器阅读理解评测难度的评估方法和经典测验理论的局限性。
- 最后，本文指出下一步研究的方向。

1 引言

1.1 机器阅读理解评测

- 机器阅读理解（Machine Reading Comprehension）评测是针对机器阅读理解能力的一项评测任务，简单地说就是测试机器回答与给定自然语言文本相关问题的能力。

Table 1: Definition of machine reading comprehension.

Machine Reading Comprehension
Given the context C and question Q, machine reading comprehension tasks ask the model to give the correct answer A to the question Q by learning the function $\mathcal{F}$ such that $A = \mathcal{F}(C, Q)$.

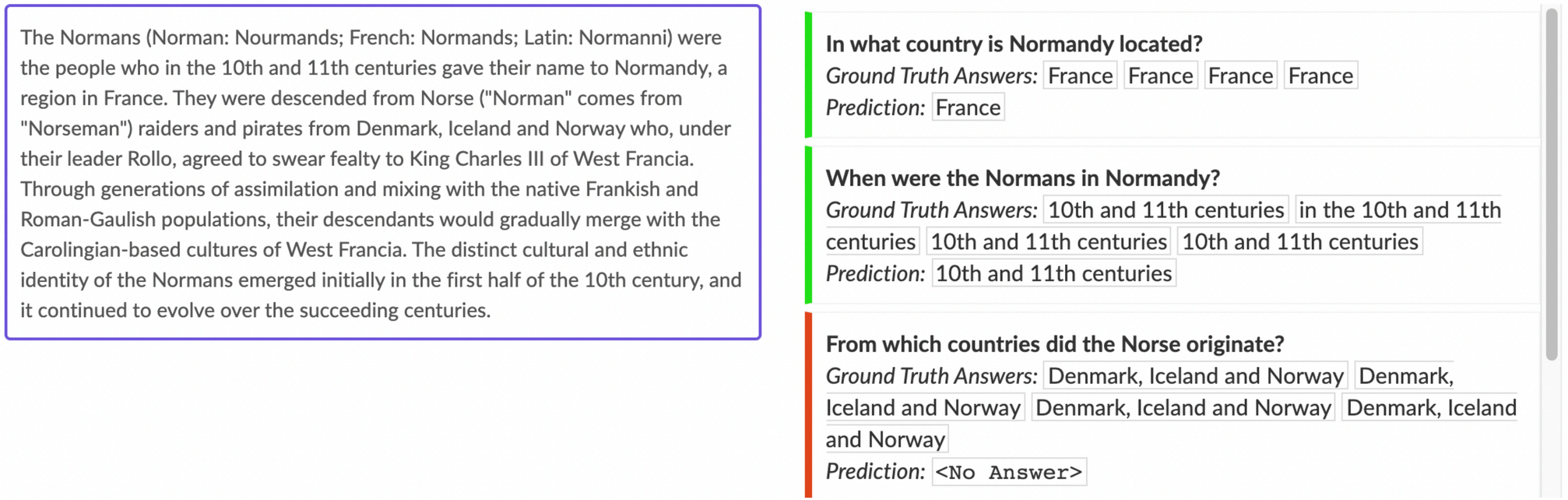
- （Liu等，2019）

- 由于阅读理解能力在人类认知和语言能力体系中的重要地位，机器阅读理解的任务表现水平也被视为评价人工智能（Artificial Intelligence）发展水平的重要参照之一，近年来受到极大关注。

1 引言

1.1 机器阅读理解评测

- 机器阅读理解评测主要以在数据集（Dataset）上测试的形式进行。
  - SQuAD 2.0（rajpurkar.github.io）



1 引言

1.1 机器阅读理解评测

- 机器阅读理解评测主要以在数据集（Dataset）上测试的形式进行。
  - 评价指标：EM、F1 Score、Accuracy、BLEU-4、ROUGE-L等
    - 排行榜
  - 问答类型：片段抽取式、完形填空式、多项选择式、自由作答式
  - 特定能力：多步推理、多步对话、常识理解、离散推理等
  - 特定领域：法律、医学等
- 本文及此前的研究（李楠，2020、2021）考察了：
  - 有代表性的英文数据集（10个）：MCTest、SQuAD系列、DROP、HotpotQA等
  - 所有中文数据集（12个）：PD&CFT、CMRC系列、DuReader系列、CAIL系列等

1引言

1.1 机器阅读理解评测

- 但面对层出不穷的数据集和榜首屡屡被刷新的机器水平排行榜，学界目前还是不能确切地说明机器的阅读理解能力达到了什么水平，也缺乏系统性地评价既有数据集评测质量的方法。
  - Jia等人（2017）的研究显示，在SQuAD 1.1数据集的段落材料中插入对于人类来说完全不影响答案的干扰信息，用来测试的六个机器模型的F1分数平均值从75%下降到了36%。
  - Gan等人（2019）的研究表明，主流的机器阅读理解模型都存在过敏感和过稳定的问题。
- 这些研究一方面说明既有的机器模型还有很大的提升空间，一方面也说明既有数据集评测的可靠性和有效性要被打上一个问号，机器在这些数据集上测试得到的超高指标并不能证明机器掌握了相应的阅读理解能力，有关机器阅读理解评测可靠性和有效性的评价方法需要更加深入的研究。

1 引言

1.1 机器阅读理解评测

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Figure 1: An example from the SQuAD dataset. The BiDAF Ensemble model originally gets the answer correct, but is fooled by the addition of an adversarial distracting sentence (in blue).

Model	Original	ADDSENT	ADDONESENT
ReasoNet-E	**81.1**	39.4	49.8
SEDT-E	80.1	35.0	46.5
BiDAF-E	80.0	34.2	46.9
Mnemonic-E	79.1	**46.2**	**55.3**
Ruminating	78.8	37.4	47.7
jNet	78.6	37.9	47.0
Mnemonic-S	78.5	**46.6**	**56.0**
ReasoNet-S	78.2	39.4	50.3
MPCM-S	77.0	40.3	50.0
SEDT-S	76.9	33.9	44.8
RaSOR	76.2	39.5	49.5
BiDAF-S	75.5	34.3	45.7
Match-E	75.4	29.4	41.8
Match-S	71.4	27.3	39.0
DCR	69.3	37.8	45.1
Logistic	50.4	23.2	30.4

Table 3: ADDSENT and ADDONESENT on all sixteen models, sorted by F1 score the original examples. S = single, E = ensemble.

- （Jia等，2017）

1引言

1.2 经典测验理论

- 经典测验理论（Classical Test Thery，简称CTT），也称真分数理论（True Score Thoery），是心理学领域最早也是最经典的测验理论。
- 它在20世纪50、60年代走向成熟，具有了完备的数学理论形式。此后数十年，经典测验理论成为了绝大多数心理学测验的理论基础，是心理与教育测验领域应用最广的测验理论（张厚粲等，心理测量学，2012）。

1引言

1.2 经典测验理论

- 真分数（T）：想要测量的目标心理特质的真实水平
- 观察分数（X）：针对这项心理特质设计的测验所得到的实际分数
- 误差：由于测验环境、被试状态、题目质量等许多因素的干扰，绝对的真分数是无法得到的
  - 随机误差（E）：偶然引起、变化难以预测的误差
  - 系统误差（I）：稳定存在、有系统性偏向的误差

- $X = T + I + E$

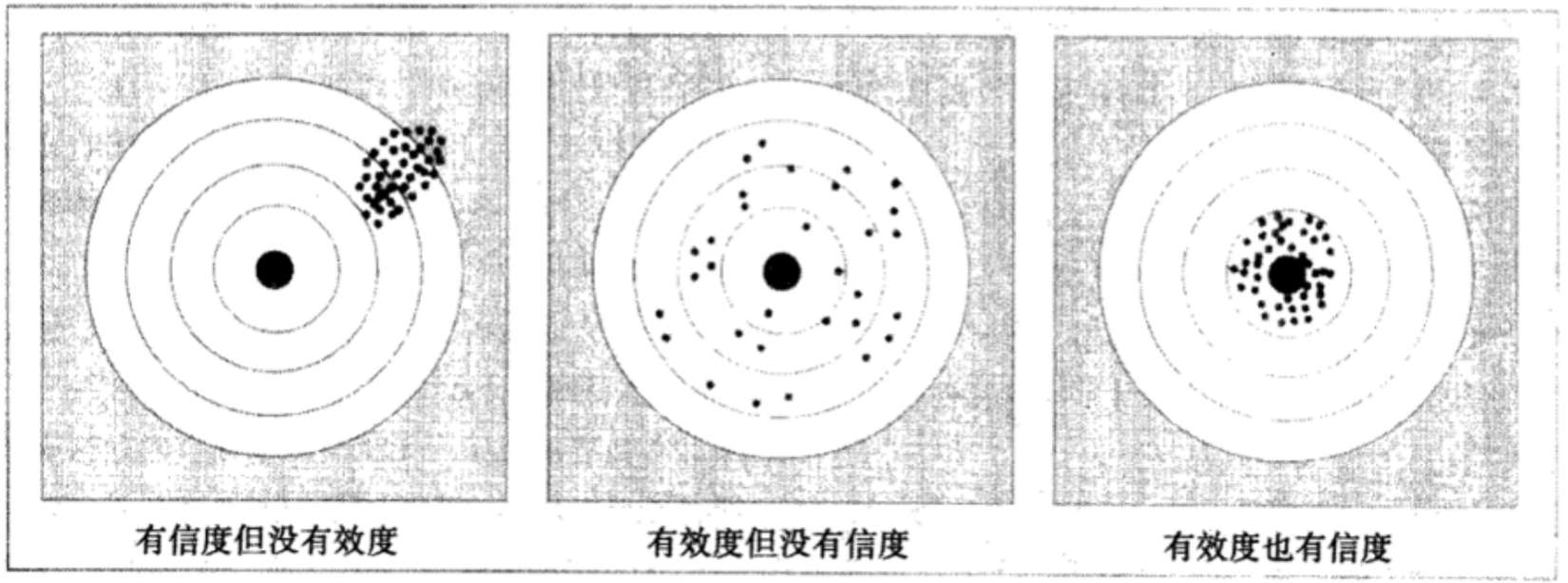
1引言

1.2 经典测验理论

- 心理学测验主要关注一项测验的可靠性和有效性。
- 定性分析
  - 可靠性：重复施测得到的结果的一致性和稳定性程度
    - 系统误差总是稳定地影响测验，不会影响重复施测时结果的一致性
  - 有效性：指一项测验能够测出所欲测量对象的程度
- 定量分析
  - 信度：观察分数的变动可以由真分数和系统误差之和的变动解释的比例
  - 效度：观察分数的变动可以由真分数的变动解释的比例
- 但由于真分数方差无法计算得到，一般通过分析不同因素对观察分数的影响效应得到信效度

1引言

1.2 经典测验理论

- 

（艾尔·巴比，2005）
- 经典测验理论的信度、效度研究和在此基础上发展出的误差控制方法为测试题的编制、评价和优化提供了范式，不仅在量表制作、测量工具评价等心理学测验上得到广泛引用，也被用于社会学调查、教育测试学命题研究等领域。本文将讨论信度、效度分析在机器阅读理解评测上的应用。

1引言

1.3 本文研究价值

- 本文将心理学领域经典测验理论下的信度、效度分析迁移应用在机器阅读理解评测上，在探讨可行性的基础上讨论机器阅读理解评测信度、效度的评估方法和检测手段，主要有以下几点研究价值：
  - 借鉴心理测量学中信效度评估、误差控制等相对成熟的学科经验，探索建立机器阅读理解评测信度、效度的评价体系
  - 总结既有的数据集质量评价和控制方法，统一到机器阅读理解数据集信度、效度评价体系中
  - 为跨人类-机器阅读理解能力评测方法的构建提供借鉴
- 同时需要指出的是，虽然本文考虑到机器阅读理解领域数据集和相关研究较多、数据格式和指标计算相对简单，将讨论局限在该领域，但相关讨论完全可以迁移到其他机器能力评测任务上，为通用机器能力评测体系的构建提供经验。

2 机器能力评测信效度分析的可行性

2.1 机器能力评测的特点

- 经典测验理论作为心理学领域的测验理论，其出发点是对人的心理特质的测量，是面向人类的。经典测验理论的信效度分析很大一部分和人的心理特点密切相关。
- 由于机器和人类极强的异质性，在将信度、效度分析应用于机器能力评测之前，需要先比较机器能力评测和人类心理测验的特点。

2 机器能力评测信效度分析的可行性

2.1 机器能力评测的特点

	人类	机器
测量主体	人类（主试）	人类
测量客体	人类（被试）	机器模型
测量对象	心理特质	机器能力
测量方式	间接测量	间接测量
测量工具	问卷，试题等	数据集
测量结果/评价指标	客观分数或人格、态度等心理学范畴	统计学指标
测量程序	主试实时控制	程序自动运行
测量环境	现实世界场景	计算机内部

表 4 人类心理测验和机器能力测试的比较

- 机器在模型训练完成后由于其本质上函数映射式的输入-输出对应关系，在同一数据集上的表现有相当的稳定性。
- 机器能力测试一般通过计算机自动进行，受到的干扰较少。
- 机器模型存在一个训练的过程，其表现强烈地依赖于训练集（Training Set）的大小、质量、与测试集的同质程度等属性，训练参数的选取和调试方法也都会影响机器模型的最终性能。

2 机器能力评测信效度分析的可行性

2.1 机器能力评测的特点

- 本文认为，为了便于进行跨数据集的信度、效度分析，如后文论及的效标关联效度，应该忽略机器模型的训练细节，假定模型训练是在对应数据集上按正常条件进行的，更加关注模型在测试集（Testing Set）上的评测过程。
- 在这样的假设下，我们讨论的机器模型是像BERT、ERINE这样的单模型（Single）或集成模型（Emsemble），当某一模型在不同数据集的训练集上完成训练后，本文不认为拥有了新的特征表示、权重矩阵等训练结果的模型是全新模型，而将它们视作同一个模型进行能力评测。

2 机器能力评测信效度分析的可行性

2.2 机器能力评测的误差来源

	误差来源	误差类型	误差控制
测量客体	机器模型缺陷	随机误差 系统误差	优化模型算法 改进工程实现
测量工具	数据存在错误	随机误差 系统误差	加强质量控制 提升标注质量
	评测体系落后	系统误差	改进评测体系
测量程序	程序运行有误	随机误差 系统误差	规范运行程序 维护实验记录
测量环境	环境配置失误	随机误差	规范运行程序 维护实验记录

表 5 机器阅读理解评测的误差来源和误差控制

3 机器阅读理解评测的信度

3.1 机器阅读理解评测信度的评估方法

信度评估方法	解释	误差来源	机器能力评测
重测信度	同一测量对象在不同时间在同一测量工具上测验结果的一致性程度	时间间隔	不适用，时间间隔不造成误差
复本信度（同时测试）	同一测量对象在同一时间在两个平行测验上测验结果的一致性程度	题目内容	不适用，构建数据集复本难度大
复本信度（延时测试）	同一测量对象在不同时间在两个平行测验上测验结果的一致性程度	时间间隔 题目内容	同上
分半信度	同一测量对象在一个测验的两个等分子测验上测验结果的一致性程度	题目内容	适用
内部一致性信度	通过测验结果计算题目的一致性程度，进而推断欲测特质的同质性	题目内容 欲测心理特质的同质性	适用
评分者信度	不同评分者在主观评价类测验中对测量对象评价的一致性程度	评分者间差异	不适用，不存在主观评价

表 6 心理测验中的各种信度估计方法及对机器能力评测的适用性，参考戴海琦（2010）[①]，有增改。

- 分半信度（Split-half Reliability）和内部一致性信度（Internal Consistency Reliability）可以用于机器阅读理解评测。
- 但对于前者，机器阅读理解数据集与心理测验相比几百上千倍的题量，使得如何精细化地二分，产生两个平行的子集成为了一个相对困难的问题。
- 对于后者，较高的内部一致性信度可以说明该评测所使用的数据集内部各部分在集中评测某项机器能力，但这能否说明评测整体有较高的信度还需要结合评测的设计意图判断。

3 机器阅读理解评测的信度

3.2 机器阅读理解评测信度的检测手段

信度评估方法	解释	误差来源	检测手段
分半信度	机器在某一数据集两个等分子集上测试结果的一致性程度	数据集内容	相关系数
内部一致性信度	通过评测结果计算数据集内容的一致程度，进而推断欲测机器能力的同质性	数据集内容 欲测机器能力的同质性	库德-理查森信度 Alpha 系数

表 7 机器阅读理解评测的信度评估方法和检测手段

- 理想条件下可以通过计算所有可能的二分方法下分半信度的平均值得到某种更有代表性的信度。内部一致性信度的计算就采取了这种思路，库德-理查森信度（Kudar-Richarson Reliability）和Alpha系数分别用于计算采用“对/错”二分类记分法和采用其他记分法的测试题，它们只需要获得一次测验的结果，就可以通过统计学方法计算测验内部题目间的一致性。

4 机器阅读理解评测的效度

4.1 机器阅读理解评测效度的评估方法

- 效度即有效性，表达的是一项测验能够测出所欲测量对象的程度，也就是一项测验的测验结果满足主体预期的程度。
- 效度和测验的目的密不可分，而无论是心理特质还是机器能力都是抽象事物，对它们内涵和外延的定义不可避免地带有主观性，因此效度分析相比信度分析更加复杂，有更为丰富的内容，呈现出定性分析和定量分析相结合的分析思路。
- 目前通行的效度分类方法来自1966年美国心理与教育领域三大学会联合推出的《教育与心理测验的标准和使用手册（Standards for Educational and Psychological Tests and Manuals）》，书中将效度分为内容效度（Content Validity）、效标关联效度（Criterion-related Validity）和构念效度（Construct Validity）。

4 机器阅读理解评测的效度

4.1 机器阅读理解评测效度的评估方法

- 内容效度：测验内容对测量目标的符合程度
- 效标关联效度：测量结果对某种效标的符合程度
- 构念效度：测验内容对某种构念的检测程度

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 内容效度：测验内容对测量目标的符合程度
  - 评估机器阅读理解评测的内容效度首先需要尽可能详尽地定义和描述所测的机器阅读理解能力及其构成要素。

数据集	分类	分类来源
SQuAD	同义词替换、运用世界知识的词语替换等 6 种推理类型	标注 192 条样本
SQuAD 2.0	否定、反义等 6 种会导致问题无答案的问句提问类型	标注 100 条样本
DuReader	先划分为实体、描述、是否三类问题，再将这三类分别再分为事实和观点两类	构建时众包标注
HotpotQA	通过中介实体完成两步推理、比较两个实体等五种推理类型	标注 100 条样本
DROP	减法、比较、计数和排序等 9 种推理类型	标注 350 条样本
COSMOS QA	先验/后验条件、动机、反应等 7 种常识类型	标注 500 条样本

表 8 机器阅读理解数据集内部分类示例

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 内容效度：测验内容对测量目标的符合程度
  - 但这些分类绝大多数都是在数据集构建完成后，通过抽样加人工标注的方法得到的，实际上只是对数据集内部分布情况的一种描写，并没有指导数据集的构建，很难说明这些分类方法的科学性、完备性以及该数据集对测量机器阅读理解能力这一目标的符合程度。
  - 抽样结果显示这些数据集的各子类占比大多是很不均匀的.

推理类型	描述	例子	比例
同义词替换	问句和答案句的主要对应关系是同义词	Q[①]: What is the Rankine cycle sometimes **called**? S: The Rankine cycle is sometimes **referred** to as a practical Carnot cycle.	33.3%
运用世界知识的词语替换	问句和答案句的主要对应关系需要世界知识解释	Q: Which **governing bodies** have veto power? S: **The European Parliament and the Council of the European Union** have powers of amendment and veto during the legislative process.	9.1%
句法变动	问句和答案句的依存句法分析结构不匹配	Q: What Shakespeare scholar **is currently on the faculty**? S: **Current faculty include** the anthropologist Marshall Sahlins, ..., Shakespeare scholar David Bevington.	64.1%
多句推理	句子间存在照应关系或更高层次的融合	Q: What collection does **the V&A Theatre & Performance galleries** hold? S: **The V&A Theatre & Performance galleries** opened in March 2009. ... **They** hold the UK's biggest national collection of material about live performance.	13.6%
模糊	问题有不止一个答案	Q: What is the main goal of criminal punishment? S: **Achieving crime control via incapacitation and deterrence** is a major goal of criminal punishment.	6.1%

表 9 SQuAD 数据集对回答问题所需要的推理类型的分类，以及它们对应的例子和题目比例。加粗内容表示和推理类型相关，下划线内容为众包工人标注的答案。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 内容效度：测验内容对测量目标的符合程度
  - 除了上述针对特定数据集的分类，还有更具一般性的对机器阅读理解任务的分类研究。
    - Weston等人（2016）从完成阅读理解所需要的各种能力出发，将机器阅读理解任务尽可能全面地划分为了20个小类；
    - Sugawara等人（2016、2017）基于相似的考虑，提出将阅读理解所需能力分为10种，后来又扩充到13种，并通过人工标注的方法在已有的一些数据集上检测了这种分类体系下数据的分布情况，结果同样显示分布较不均匀。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

序号	分类	序号	分类	序号	分类	序号	分类
1	单支撑事实	6	是否问题	11	简单共指	16	简单增添
2	两支撑事实	7	计数	12	关联	17	位置推理
3	三支撑事实	8	列表/集合	13	复杂共指	18	尺寸推理
4	两因素关系	9	简单否定	14	时间推理	19	路径寻找
5	三因素关系	10	不确定知识	15	简单扣减	20	角色动机

表 10 Weston 等人（2016）对机器阅读理解任务的分类

序号	分类	MCTesst	SQuAD	序号	分类	MCTest	SQuAD
1	目标跟踪	6%	3%	8	省略	4%	3%
2	数学推理	4%	0%	9	桥接[①]	26%	42%
3	共指消解	49%	13%	10	解释	8%	13%
4	逻辑推理	2%	0%	11	元知识	1%	0%
5	类比	0%	0%	12	从句关系	40%	28%
6	因果关系	6%	0%	13	标点符号	1%	24%
7	时空关系	9%	2%		不明	1%	3%

表 11 Sugawara 等人（2016、2017）对机器阅读理解所需能力的分类，以及通过人工标注的方法得到的 MCTest 和 SQuAD 两个数据集在这种分类体系下数据的分布情况，允许一个题目需要综合利用多种能力。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 内容效度：测验内容对测量目标的符合程度
  - 对机器阅读理解能力进行完备的描述有助于限定评测内容的范围，明确子类的结构及各子类应占的比例。
  - 第二步就是要验证和提升数据集里的题目与欲测的机器阅读理解能力相符合的程度，确定测试题所需的知识和技能在评测的考查范围内，检查上一步构造的各子类所辖的题目比例，这一步可以采用专家检查、众包工人交叉检查等方式进行。
  - 最后就是评价评测整体与欲测的机器阅读理解能力相符合的程度，综合各方面因素对评测的内容效度做出评判。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 效标关联效度：测量结果对某种效标的符合程度
  - 效度标准（Validity Critetion）简称效标，是独立于当前测验、具有较高效度、能够用于参考的外部标准。
  - 效标关联效度建立在一种理论假设之上：同一批被试在测量目标相同或相似的两种不同测量工具上的表现分布应该是高度相似的。
  - 对于机器阅读理解评测而言，在讨论某一数据集的效标关联效度时，可以以其他数据集为比较标准。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 效标关联效度：测量结果对某种效标的符合程度
  - Liu等人（2021）假设SQuAD作为机器阅读理解领域的经典数据集具有较高质量，搜集了20个面向SQuAD设计的机器模型，讨论32个其他数据集和SQuAD的一致性，也即这20个模型在每个数据集上测试结果的分布情况是否和在SQuAD上一致。结果显示，许多人工构建的和完形填空式的数据集和SQuAD有很强一致性，而已有的机器自动构建的数据集和SQuAD的一致性较弱。之后Liu等人总结经验成功构造出两个和SQuAD有较强一致性的、小型的、机器自动构建的、完形填空式数据集。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 效标关联效度：测量结果对某种效标的符合程度

We examine this question by conducting a retrospective study of previously-proposed SQuAD modeling approaches. In particular, we ask whether modeling approaches originally developed for and evaluated on SQuAD are ranked similarly by 32 existing and synthesized benchmarks.

We say that two benchmarks have high *concurrence* if they rank a set of modeling approaches similarly, i.e., if approaches that yield performance improvements on one benchmark also produce performance improvements on the other. To measure concurrence between two benchmarks, we first evaluate 20 different modeling approaches on each benchmark—all evaluation is in-domain, using each benchmark's original *i.i.d.* train-test split.

[2]Though datasets unfit for producing systems on their own can still have utility for building accurate systems when used in conjunction with other more realistic resources (e.g., via data augmentation).

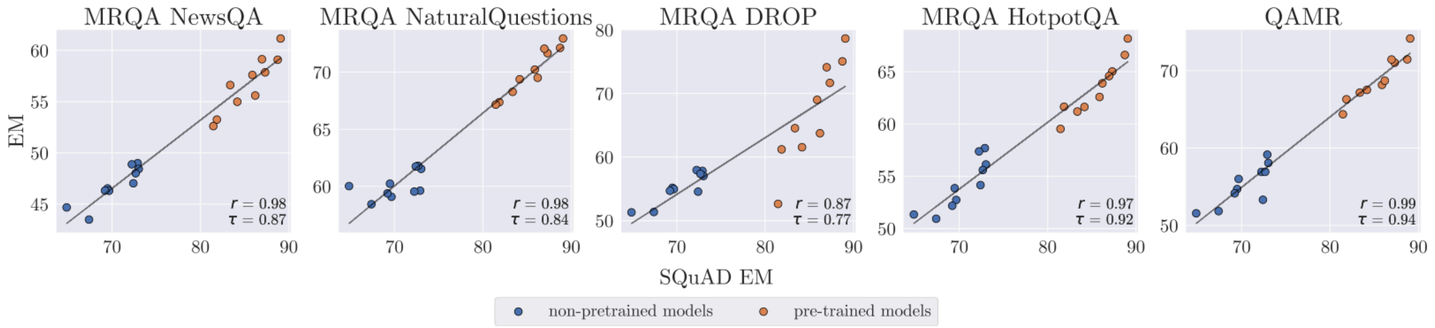


Figure 2: Many human-constructed benchmarks have high concurrence with SQuAD, suggesting that the specific nuances that go into building human-constructed benchmarks may not drastically affect concurrence—natural language collected from humans may be largely sufficient for high concurrence with SQuAD. MRQA DROP has lower overall concurrence, but we see higher concurrence within the non-pretrained and pre-trained subgroups.

- 从效度分析的角度看，Liu等人正是在假设SQuAD数据集有较高效度的前提下，将SQuAD作为效标，对其他数据集和所构建的新数据集的效度进行验证，说明了效标关联效度在机器阅读理解评测中的可用性。

4 机器阅读理解评测的效度

4.2 机器阅读理解评测效度的检测手段

- 构念效度：测验内容对某种构念的检测程度
  - 构念，是心理学概念，指的是智力、动机、兴趣、情绪这样抽象的、假设性的概念.
  - 定义法、效标参照法，相关法、因素分析法、实验操作法、内部一致性分析法……
  - 对于机器阅读理解评测而言，自然语言处理领域对模型鲁棒性的测试方法（Jia等，2017）（Gan等，2019）可以用来测试和机器阅读理解能力有关的构念。

5 余论

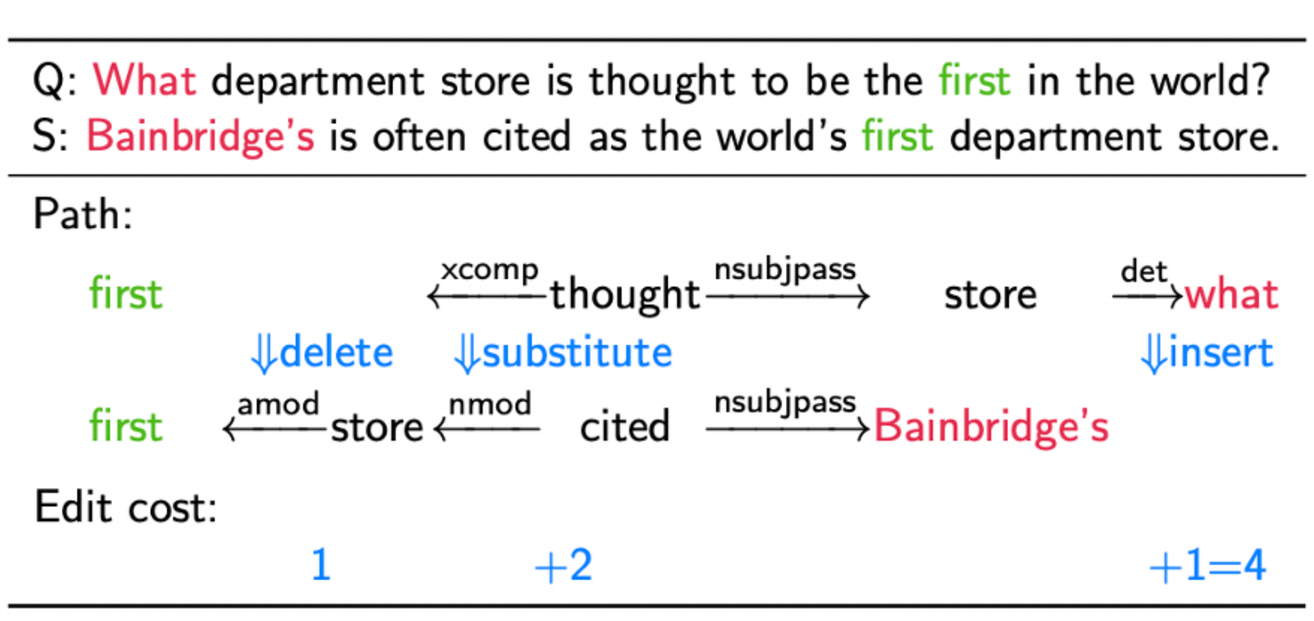
5.1 机器阅读理解评测的难度

- 难度：测试题的难易程度
  - 难度是题目本身和测量客体相互作用的结果，需要在测验结果上计算得到。
  - 常用的计算方法有得分率法和极端分组法，都可以在机器阅读理解评测上应用。
    - 得分率法：某道题的得分率等于所有机器模型在该题上的平均得分与该题满分的比值；
    - 极端分组法：某道题的难度值等于对机器模型按表现分组后的高分组、低分组分别在该题上的得分率的平均值。

5 余论

5.1 机器阅读理解评测的难度

- 难度：测试题的难易程度



- 机器阅读理解数据集自身也发展出了一些难度评价方法，比如SQuAD的句法分歧度

5 余论

5.1 机器阅读理解评测的难度

- 难度：测试题的难易程度
  - Sugawara等人（2017）的研究显示，以往有关语句可读性（Readability）研究中的许多指标，诸如词语平均字符长度、词语平均音节长度等词汇特征，句子平均词语数、句中并列短语数等句法特征，Coleman–Liau指数等传统特征，其实并不能影响机器阅读理解数据集中问题的难度，问题的难度和单个问题所需的机器阅读理解能力种类数更相关。

序号	分类	MCTesst	SQuAD	序号	分类	MCTest	SQuAD
1	目标跟踪	6%	3%	8	省略	4%	3%
2	数学推理	4%	0%	9	桥接①	26%	42%
3	共指消解	49%	13%	10	解释	8%	13%
4	逻辑推理	2%	0%	11	元知识	1%	0%
5	类比	0%	0%	12	从句关系	40%	28%
6	因果关系	6%	0%	13	标点符号	1%	24%
7	时空关系	9%	2%		不明	1%	3%

表 11 Sugawara 等人（2016、2017）对机器阅读理解所需能力的分类，以及通过人工标注的方法得到的 MCTest 和 SQuAD 两个数据集在这种分类体系下数据的分布情况，允许一个题目需要综合利用多种能力。

5 余论

5.1 机器阅读理解评测的难度

- 难度：测试题的难易程度

that Sugawara et al. (2017) proposed for the fine-grained analysis of RC capability. Their study also presented an important observation of the relation between the difficulty of an RC task and prerequisite skills: the more skills that are required to answer a question, the more difficult is the question. Based on this observation, in this work, we assume that the number of skills required to answer a question is a reasonable indication of the difficulty of the question. This is because each skill corresponds to one of the functions of an NLP system, which has to be capable of that functionality.

Our second class defines metrics for "text ease of processing," namely the difficulty of reading the text. We regard it as readability of the text in terms of syntactic and lexical complexity. From

Metrics	r	p	Metrics	r	p
NumChar	0.068	0.095	CoOrd	0.166	0.000
NumSyll	0.057	0.161	Coleman	0.140	0.001
MLS	0.416	0.000	DC/C	0.188	0.000
AWL	0.114	0.005	CN/C	0.131	0.001
ModVar	0.025	0.545	AdvVar	0.026	0.515
F–K	0.343	0.000	Words	0.355	0.000

Table 6: Pearson's correlation coefficients (r) with the p-values (p) for the readability metrics and number of required prerequisite skills for all questions in the RC datasets.

- Ave. no. of characters per word (*NumChar*)
- Ave. no. of syllables per word (*NumSyll*)
- Ave. sentence length in words (*MLS*)
- Proportion of words in AWL (*AWL*)
- Modifier variation (*ModVar*)
- No. of coordinate phrases per sentence (*CoOrd*)
- Coleman–Liau index (*Coleman*)
- Dependent clause-to-clause ratio (*DC/C*)
- Complex nominals per clause (*CN/C*)
- Adverb variation (*AdvVar*)

Table 1: Readability metrics. *AWL* refers to the Academic Word List.[4]

5 余论

5.2 经典测验理论的局限性

- 张厚粲等（2012）将经典测验理论的局限性总结为三点：采用的指标依赖于受试者样本；能力的估计依赖于项目样本；各种参数估计都只能事后进行。
  - 对某一具体的机器阅读理解评测任务所作的信度、效度、难度分析依赖于施测的机器模型能力的高低。
  - 对机器阅读理解能力的评价只能以一项评测的结果为依据，比较只能限于同一项评测上的重复施测，很难进行跨评测任务的比较。
  - 只有在数据集构建完、评测任务设计完以后，才能对机器阅读理解评测任务进行信度、效度分析，无法提前对评测任务的信度、效度进行有效控制，分析前的工作具有一定盲目性。

6 总结与展望

不足

- 缺少对现有机器阅读理解数据集信度、效度进行实验测量的实证研究；
- 没有结合数据集标注等机器阅读理解评测特有的程序和内容，探讨是否能提出机器阅读理解评测特有的信度、效度评估方法；
- 没有将信度、效度研究扩展到更一般的机器能力评测上。

- 这些是下一步研究需要克服的困难和前进的方向。

参考文献

专著和论文

- 戴海琦. 心理测量学. 高等教育出版社. 2010年5月第1版.
- 张厚粲, 龚耀先. 心理测量学. 杭州: 浙江教育出版社. 2012年5月第1版.
- Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M. Rush, Bart van Merriënboer, Armand Joulin, Tomas Mikolov. Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks. arXiv preprint arXiv:1502.05698. 2015.
- Saku Sugawara, Akiko Aizawa. An Analysis of Prerequisite Skills for Reading Comprehension. Proceedings of the Workshop on Uphill Battles in Language Processing: Scaling Early Achievements to Robust Methods. Association for Computational Linguistics. Pages 1-5. 2016.
- Robin Jia, Percy Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics. Pages 2021-2031. 2017.
- Saku Sugawara, Yusuke Kido, Hikaru Yokono, Akiko Aizawa. Evaluation Metrics for Machine Reading Comprehension: Prerequisite Skills and Readability. Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers). Association for Computational Linguistics. Pages 806–817. 2017.
- Nelson F. Liu, Tony Lee, Robin Jia, Percy Liang. Can Small and Synthetic Benchmarks Drive Modeling Innovation? A Retrospective Study of Question Answering Modeling Approaches. arXiv preprint arXiv:2102.01065. 2021.
- 李楠. 基于评测任务调研的机器阅读理解能力研究. 北京大学中国语言文学系本科生学年论文. 2020.
- 李楠. 中文机器阅读理解数据集发展的现状、问题和对策. 北京大学本科生科研训练项目结题论文. 2021.
- 数据集论文略