NLI数据健全性检查:评估数据损坏对模型性能的影响

# Aarne Talman\*、Marianna Apidianaki\*、Stergios Chatzikyriakidis、Jo rg Tiedemann\*

\*赫尔辛基大学数字人文系

name .姓氏@赫尔辛基. fi

*{ }*

*地下室人工智能*

*哥德堡大学哲学、语言学和科学理论系*

*{name .姓氏}@gu.se*

# 摘要

**前提假设**

预训练的神经语言模型在自然语言干扰(NLI)任务中表现良好。但是他们是否真的理解了加工过的序列的含义还不清楚。我们提出了一个新的诊断测试套件，它允许评估数据集是否构成了一个好的测试平台，用于评估模型的意义理解能力。我们特别将受控腐败转化应用于广泛使用的

*他不超过五英尺四英寸，但随身携带*

*伟大的尊严。*

*两株植物在漫长的旅途中死亡，第三株植物找到了去牙买加的路，具体情况仍是个谜。*

*几天后，货车将向北开往图森，但现在广场上的活动混杂着*

中立的

蕴涵矛盾

*这个人有6英尺高。*

*第三种植物与前两种不同。*

*他们在图森南部。*

基准(MNLI和ANLI)，其中-

涉及删除整个词类，并经常导致不敏感的句子对。如果损坏数据的模型精度仍然很高，那么数据集可能包含指导预测的统计偏差和假象。相反，模型精度的大幅下降表明原始数据集对模型的推理能力提出了适当的挑战。因此，我们提议的控制可以作为为NLI任务开发高质量数据的碰撞测试。

# 介绍

评估模型的自然语言推理(NLI)和理解(NLU)能力带来了许多挑战，其中之一就是用于评估的数据的质量和组成。流行的NLI数据集(Bowman等人，2015年；Marelli等人，2014)包含注释事实和统计不规则性，即使模型没有获得执行这种预测所需的知识，模型也可以在训练和指导预测期间容易地掌握这些事实和统计不规则性。值得注意的是，最近的工作表明，重大修改，如单词洗牌不会损害伯特的(德夫林等人，2019年)NLU的能力

*市场日和嘉年华。*

表1:已删除名词的受损MNLI训练数据集中的句子对。

很多，主要是由于单个单词对预测的影响(Pham等人，2020)。相反，数据中的微小调整或扰动，如用互斥的共形异义词和反义词替换单词(Glockner等人，2018年)或改变两个句子的顺序(Wang等人，2019年b)，已被证明会损害模型的性能。

在这种情况下，我们的目标是推荐一套新的诊断测试，用于评估NLU基准的质量。特别是，我们进行了一系列受控实验，将一组数据损坏转换应用于广泛使用的MNLI (Williams等人，2018年)和ANLI (Nie等人，2020年)数据集，并探讨了它们对微调BERT和ROBERTa(刘等人，2019年)模型性能的影响。所获得的结果提供了可以揭示数据集质量的证据:假设转换严重影响NLI句子的质量，甚至使它们不可理解(参见表1中的例子)，在cor上微调的模型的性能下降

应该是中断的数据集。相反，高性能将表明数据集中存在指导模型预测的双基地和其他假象。这种情况表明数据集质量较低，也就是说，我们不能依靠它来得出关于模型NLI能力的安全结论。

将额外的证据引入到关于有问题的NLI评估设置以及它们代表测试模型的真实推理能力有多差的辩论中，我们提出的诊断允许通过评估数据集有多人工制品和无偏差来评估数据集的质量，从而评估它们在评估NLI模型的语言推理能力时可以信任的范围。我们认为这一步对于评估现有基准的质量和相应地解释模型结果，以及指导处理推理和推理的新数据集的开发非常重要。我们提供我们的代码和数据，以促进这些诊断测试的采用，并促进它们在新数据集上的应用。1

# 相关著作

评估基准的一个众所周知的问题是，所提出的任务通常可以通过简单的启发式算法来解决(休伊特和梁，2019)。这主要是由于数据集内存在语言双音，使得预测变得容易(Lai和Hockenmaier，2014；Poliak等人，2018年)。值得注意的是，表示原始SNLI数据集中的矛盾的90%的假设(Bowman等人，2015)包含动词sleep及其变体(sleep，sleep)，其用于标记与前提中描述的活动的对比(例如，我的姐姐正在玩我的姐姐正在睡觉)；而SICK中的矛盾(Marelli等，2014)往往以明确否定为标志。正如Guru- rangan等人(2018)所发现的，后一个问题也存在于SNLI和MNLI中，在SNLI和MNLI中，否定是矛盾的高度指示，而类属名词(例如动物、某物)是蕴涵的。这些语法或词汇线索很容易被模型在训练中掌握，并帮助它们正确预测两个句子之间的关系，但这并不意味着模型能够执行这种类型的推理。值得注意的是，

*→*

规律，甚至对于仅假设的NLI模型(即只对假设进行微调而不进入前提的模型)也有可能做出正确的预测(Poliak等人，2018)。

最近的工作表明，最先进的NLU模型对词序不是很敏感，然而，词序是序列最重要的特征之一(Pham等人，2020)。具体来说，基于BERT的分类器在GLUE任务上微调的性能(王等人，2018)在随机洗牌输入单词后保持相对高的性能。这主要是通过每个单词的贡献来解释的，这些单词在其上下文被打乱后保持不变。情感分析中关键词的情感或NLI句子对之间的词级相似性等表面线索，使得基于BERT的模型即使在标记以随机顺序排列时也能做出正确的决定，这表明许多GLUE任务并没有真正挑战他们理解句子的意思。

相反，当简单的试探法不足以解决NLI任务时，NLI系统似乎更容易崩溃。例如，当交换不同基准的测试和训练数据集时(即在一个NLI数据集上训练并在另一个数据集上测试)，就会发生这种情况(Talman和Chatzikyriakidis，2019)。王等人(2019b)报告了前提和假设互换时的性能问题。其思想是，在交换的情况下，矛盾或中性对的标签应该保持不变，而在隐含对的情况下，在交换之后应该提出不同的标签。这是可以预期的，因为蕴涵是一种方向关系，而矛盾是对称的。2王等人(2019b)测试了与该诊断相关的各种模型，并观察到当矛盾对和中性对交换时，性能显著下降(即预测的标签变化)。当在交换的蕴涵对上测试这些模型时，这些模型的行为似乎更合理，在交换的蕴涵对上，除了一个模型之外，所有模型都正确地预测了不同的标签。根据这些结果，作者提出交换法作为NLI模型的健全性检验。

现有数据集的低质量和NLI系统令人印象深刻的高性能，

以这些基准衡量，引发了

由于这些注释假象和统计ir-

2更明确地说，对于矛盾，想法是当

1 https://github . com/赫尔辛基-NLP/

数据健全性检查

*A → B(即B与A相矛盾)，然后，通过对位，*

*B → A也成立(A反驳B)。*

一个新的研究方向，目标是提出新的更具挑战性和无伪影的数据集。例如，ANLI数据集的精确构建目标是消除伪像(聂等人，2020年)。作者声称，与以前的基准相比，这个数据集更不容易出现伪像，这是因为在仅基于ANLI假设的数据集上微调的模型的预测精度更低。尽管似乎仍有改进的空间(准确度约为0.5，即远高于可能性)，但报道的发现是有希望的。具体而言，性能低于仅基于假设的SNLI/MNLI数据集，表明该数据集包含较少的可指导预测的假象。因此，ANLI是进一步检验我们假设的自然候选者，因为它声称能弥补早期NLI数据集的许多缺点。

从先前设计可靠的语言探测任务的工作(Hewitt和Liang，2019)中吸取的经验教训，以及上面讨论的模型的过度拟合问题，证明了我们在本文中提出的系统健全性检查的重要性。我们的专用控制任务特别允许确定数据集是否触发模型的推理能力，或者允许它们依赖统计偏差和注释假象进行预测。我们使用模型所做的预测质量作为评估数据质量的代理，这些模型经过了微调并在损坏的数据上进行了测试。

# 数据集

## 多体裁NLI语料库

我们在多体裁自然语言推理系统上进行了实验。MNLI包含433k个标记为“必然”、“矛盾”和“中性”的人类书写句子对。该语料库包括来自10种不同文体的书面英语和口语的句子对，3使得有可能以各种各样的方式来近似

3在线文本类型:面对面和电话对话(面对面、电话)；来自公共领域政府网站(government)的内容；来自印第安纳州跨文化交流中心的信件

现代标准美国英语被使用，并提供了一个评估跨体裁主要改编的环境。所有十个流派都出现在测试和开发集中，但只有五个流派包含在训练集中。MNLI开发和测试集分为“匹配的”和“不匹配的”:前者仅包括来自训练数据中发现的相同体裁的句子，而后者包括来自训练数据中不存在的剩余体裁的句子。对于我们的实验，我们使用开发集作为我们的评估数据，因为带注释的测试集是不公开的。

## 对抗性NLI语料库

对抗性的基准(ANLI)(聂等人，2020年)是专门为解决以前的数据集的一些缺点而设计的。ANLI包含三个数据集(轮次)，R1、R2和R3。每一个数据集都是使用人和模型在回路中的方法收集的，它们的难度和复杂性逐渐增加。向注释者展示了一个上下文(前提)和一个目标标签，并要求他们提出一个假设，该假设会导致模型对标签进行误分类。对于R1来说，注释者被要求欺骗的模型是伯特-拉格尔，而对于R2和R3来说，则是罗伯塔。对于R3，上下文是从更广泛的来源中选择的。4语料库还包括由注释者提供的标签解释。每一轮(R1- R3)都包含训练、开发和测试数据。

ANLI是一个相对较小的数据集。R1只有16，946个训练例子，1，000个开发和1，000个测试例子。R2稍微大一点，它包含45，460个训练样本和与R1相同数量的开发和测试样本。最后，R3包含100，459个训练示例和稍大的开发和测试集(每个1，200个)。

## 系统性NLI数据腐败

我们通过对原始数据集应用一组受控转换来创建MNLI训练和评估数据的修改版本。我们分别称这两个集合为MNLI腐败训练和腐败测试。我们特别要-

慈善筹款演讲(书信)；民众

美国遭受恐怖袭击国家委员会的报告(9/11)；关于纺织工业和儿童发展的非小说作品(OUP)；流行文化文章(SLATE)；旅行指南(TRAVEL)；针对非专业人士的语言学短文(逐字逐句)；小说。

R1和R2的语境是由维基百科上的句子组成的。在R3中，上下文是从维基百科、新闻(普通爬行)、小说、儿童图书测试(CBT)、正式口语文本和从维基百科提取的程序文本中检索的。

在使用NLTK库和平均perceptron tagger用通用词性(POS)标签标记文本之后，移动特定词类的单词。5例如，在获得的MNLI-NOUN训练数据集中，原始MNLI训练数据中的所有名词都已被移除。我们还按照相反的过程创建训练数据，即只保留特定类别的单词，去掉其他单词。例如，名词+动词数据集只包含原始MNLI句子中的名词和动词。

我们通过从匹配的开发数据集中移除特定单词类的单词，或者保留这些单词并移除其余的单词，类似地创建了腐败测试集。Ap- pendix中的表2包含了在应用每个转换后获得的训练和评估数据集的统计数据。最后，我们将原始的MNLI和损坏的训练数据集结合起来。MNLI-ALLDROP包含以下训练集:MNLI(原文)，-NUM，-CONJ，- ADV，-PRON，-ADJ，-DET，-动词，-名词。

我们使用ANLI作为高质量数据集的一个例子，并通过在ANLI测试集上应用所有的-POS转换来创建ANLI-CORPORATE-TEST。附录中的表3包含了关于不同的反腐败测试数据集的统计数据。为了测试破坏ANLI实验中使用的训练数据的效果(聂等人，2020)，我们还创建了一个训练集，该训练集由SNLI、MNLI、PHERE和ANLI训练数据组成，并删除了所有出现的名词(ANLI-COMPORT-TRAIN)。

我们在中断的MNLI数据上测试了BERT的性能，在损坏的ANLI数据上测试了RoBERTa的性能，并将结果与使用原始数据集获得的结果进行了比较。我们预计，与对原始数据进行微调的相同模型相比，对损坏数据进行微调的模型(重要信息缺失，句子通常没有意义)表现不佳。对这些问题数据进行微调的模型的高性能将表明模型利用了数据中存在的线索(偏差和假象)，而不是执行推理操作。相反，低模型性能将表明他们无法使用这些损坏的数据进行推理，并且数据不包含在这种情况下指导预测的假象。

5[https://www.nltk.org/](http://www.nltk.org/)。

# 模型

我们使用谷歌原始的768维BERT模型(BERT-base)的TensorFlow实现6，这是一个通过双向编码器学习表示的转换器模型(Devlin等人，2019)。使用掩蔽语言模型(MLM或完形填空)任务对伯特进行了预训练，在该任务中，一定比例的输入标记被随机掩蔽，模型需要预先预测这些被掩蔽的标记；并且在下一句预测(NSP)任务中，它接收成对的句子(A，B)作为输入，并且学习预测在原始文档中B是否跟随A。(A，B)中的句子B有50%的时间是跟在A后面的实际句子，50%的时间是训练语料中的随机抽取。NSP增加了模型捕捉两个句子之间关系的能力，这是《NLI与问答》中的核心任务。

BERT模型的变体在任务上实现了非常高的性能，超过了hu- man在GLUE上的基线(Wang等人，2018年)，并在具有挑战性的SuperGLUE数据集上达到了接近人类的性能(Wang等人，2019年a)。对于每个实验，我们使用100的批量大小(除非明确说明)在原始MNLI训练数据集或第3节中描述的其转换版本上微调十个时期的BERT。

对于在ANLI基准上的实验，我们应用了RoBERTa-large模型，它是BERT的一个变种，在GLUE和SuperGLUE基准上具有比BERT高得多的性能。7我们使用了聂等人(2020)提供的训练和评估脚本。8我们使用16的批量大小为两个时期微调模型。

# 估价

* 1. **腐败-培训和原始测试**

我们评估了BERT模型在MNLI CORPORT-TRAIN的14个训练集上进行微调时的性能。我们在原始MNLI-上测量模型的预测精度

6h ttps://github . com/Google-research/

伯特

7 Roberta中的修改包括用更大的批次、更多的数据和更长的序列对模型进行更长时间的训练。与BERT相比，预训练方法也进行了修改:删除了下一句预测目标，并引入了动态掩蔽。这导致不同的令牌在不同的训练时期被屏蔽。

8 https://github . com/face book research/

安利

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **数据** | 腐败列车 | ∆ | 腐败测试 | ∆ | 腐败列车 | 和 | 试验 | ∆ |
| MNLI-NUM | 82.37% | -1.37 | 81.71% | -2.03 | 81.87% | -1.87 | | |
| MNLI-CONJ | 83.09% | -0.65 | 82.75% | -0.99 | 83.10% | -0.64 | | |
| MNLI-ADV | 80.21% | -3.53 | 72.41% | -11.33 | 75.69% | -8.05 | | |
| MNLI-PRON | 83.27% | -0.47 | 81.98% | -1.75 | 82.65% | -1.09 | | |
| MNLI-ADJ | 81.67% | -2.07 | 74.61% | -9.13 | 76.44% | -7.30 | | |
| MNLI-DET | 83.15% | -0.59 | 79.29% | -4.44 | 81.32% | -2.42 | | |
| MNLI-动词 | 81.40% | -2.34 | 73.96% | -9.78 | 76.30% | -7.44 | | |
| MNLI-NOUN | 80.72% | -3.02 | 69.80% | -13.94 | 73.38% | -10.35 | | |
| PRON | 79.74% | -4.00 | 68.41% | -15.33 | 72.14% | -11.60 | | |
| 名词+PRON+动词 | 72.55% | -11.19 | 54.59% | -29.15 | 62.18% | -21.56 | | |
| 名词+ADV+动词 | 67.58% | -16.16 | 62.58% | -21.16 | 67.58% | -16.16 | | |
| 名词+动词 | 71.14% | -12.60 | 52.90% | -30.84 | 61.31% | -22.43 | | |
| 名词+动词+ADJ | 75.54% | -8.20 | 61.90% | -21.84 | 68.20% | -15.54 | | |
| 名词+动词+ADV+ADJ | 79.81% | -3.93 | 71.81% | -11.93 | 76.29% | -7.45 | | |

表2:基于BERT模型的预测精度(%)，该模型在CORPORT-TRAIN上进行了微调，并在原始的MNLI匹配评估(dev)集上进行了测试(第2列和第3列)；对原始MNLI数据进行微调，并在CORRUPT-TEST上进行测试；在腐败训练中进行微调，并在腐败测试中进行测试(第6列和第7列)。与在原始MNLI训练集上微调的模型和在MNLI匹配开发集上评估的模型(83.74%)相比，delta显示了准确性的差异。

匹配的开发数据集，作为我们的测试集。表2第一列给出的结果显示，从MNLI训练数据中移除特定词类的所有出现对BERT性能的影响低得惊人，BERT性能仍然很高。不出所料，当删除内容词时，下降幅度最大，副词的影响最大(-3.53)，其次是名词(-3.02)和动词(-2.34)。有趣的是，数据集中名词的数量是副词的4.5倍，这表明后者对NLI预测有更大的影响。然而，全面观察到的精确度的小幅下降非常令人惊讶。可以说，名词重新移动的句子对人类来说意义不大(参见表1)。9在这些有问题的数据上观察到的BERT的高性能可能是由于模型在预训练期间获得的关于间隙填充和下一句预测的知识，它仍然可以利用和结合训练和测试数据中的其他线索进行预测。

* 1. **腐败测试评价**

基于原始数据进行微调的模型。我们对BERT模型的性能进行了评估，该模型根据原始的MNLI训练数据和我们的腐败测试数据进行了微调。表2的中间几列显示了在不同的腐败测试数据集上的实验结果，以及与原始(未定义)数据集上的结果的差异(增量)

9Cf。附录中的表1给出了BERT做出正确预测的MNLI-NOUN测试集中损坏的句子对的例子。

与MNLI匹配的开发集。

我们观察到与先前实验相似的模式。去除内容词(名词、动词和副词)对模型准确性的影响最大，而去除连词和数词对结果的影响很小。在这种情况下，预测精度的下降比在腐败训练中微调并在未修改的数据上测试的模型的评估更重要。然而，在所有名词或动词都被移除的情况下，BERT仍然可以以相当高的准确度预测正确的标签，这一事实令人惊讶，因为这些转换经常导致几乎不可忽略的句子对(参见本文表1和附录表1中的例子)。由于这种无意义句子中的干扰不能依赖于意义，我们对模型性能的解释是，它们利用腐败后句子中剩余的其他线索和情况进行预测。请注意，在此设置中测试的模型是根据原始MNLI数据进行微调的。我们相信，在这个阶段，模型需要关于可能的序列对的知识，包括人工产物和其中的其他线索。

在腐败列车上微调的模型。我们评估了在腐败训练和腐败测试中微调的伯特模型的性能。

这些实验的结果显示在表2的最后两列。就不同词类的相对重要性而言，我们再次观察到类似的模式，其中内容词的影响最大。什么是绝对的

**训练数据匹配的**

MNLI 83.74% 83.76%

**MNLI-ALLDROP 84.09% 84.30%**

表3:在原始MNLI训练集和MNLI-ALLDROP上微调并在原始MNLI评估(dev)集上测试的基于BERT的模型的预测精度(%)的比较。

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **数据** | 腐败测试 | **R1** | ∆ | 腐败测试 | **R2** | ∆ | 腐败测试 | **R3** | ∆ |
| ANLI-CONJ | 70.2% | -3.6 | | 49.0% | 0.1 | | 46.5% | 2.1 | |
| 安立-PRON | 69.6% | -4.2 | | 49.7% | 0.8 | | 45.0% | 0.6 | |
| 安立-DET | 69.5% | -4.3 | | 49.4% | 0.5 | | 45.0% | 0.6 | |
| 安立-ADV | 67.1% | -6.7 | | 49.6% | 0.7 | | 43.8% | -0.6 | |
| 安立-ADJ | 60.2% | -13.6 | | 45.1% | -3.8 | | 45.0% | 0.6 | |
| ANLI-NUM | 58.7% | -15.1 | | 43.8% | -5.1 | | 45.1% | 0.7 | |
| 动词 | 54.6% | -19.2 | | 44.7% | -4.2 | | 39.3% | -5.1 | |
| ANLI-NOUN | 43.7% | -30.1 | | 36.0% | -12.9 | | 32.4% | -12.0 | |

表RoBERTa大型模型在腐败的R1、R2和R3测试集上的预测精度(%)。Delta显示了与Nie等人(2020)在原始测试集上报告的最先进结果相比的准确性差异，:73.8%，:48.9%和R3: 44.4%。

在这种情况下，令人惊讶的是，性能的下降比在原始数据上训练并在CORPORT-TEST上测试的模型观察到的下降要小，这表明在这种情况下，模型更加依赖数据假象。

* 1. **多边核方案评估**

受从训练数据中移除特定词类时观察到的预测准确率小幅下降的影响(参见第5.1节)，我们还在包含不同的腐败训练集和原始多边学习训练集的大数据集上微调了模型。BERT微调代码正在洗牌提供的例子，所以我们在这里的目标是探索看到不同类别的单词缺失的句子对(例如，在不包含名词的句子后面没有动词的句子)是否会混淆模型。

这个实验的结果见表3。他们指出，在训练过程中从句子中去除不同词类的出现可以作为一种正则化技术，从而提高模型的性能。我们观察到，当在原始MNLI匹配的开发数据上评估时，有小幅增加(+0.35)，当在原始MNLI不匹配的开发数据上评估时，增加了0.56。

## 对ANLI的评价

为了证明系统数据一致性可以作为评估基准质量的有用诊断，我们在ANLI测试集上进行了额外的实验(聂等人，2020)。罗伯塔大型模型的结果很好

表4给出了在原始训练集上进行的调整和在腐败测试R1、R2和R3数据上的评估。

不出所料，我们观察到数据集的准确性明显下降，其中包含内容的单词被移除(-名词，-动词)，而当虚词缺失(-康吉，-DET)时，准确性下降相对较小，但仅在R1。然而，R2和R3数据集上的精度在应用了一些损坏变换(ANLI-PRON，-CONJ，-DET)后有所提高，这是一个有趣的发现。一种可能的解释是，与其他NLI数据集相比，ANLI中的感知(尤其是前提)要长得多，因此删除不含内容的单词使模型更容易掌握做出正确预测所需的基本信息。去除名词和动词后准确率的大幅下降支持了我们的假设，即ANLI语料库的质量优于MNLI语料库，这表明该数据集包含的人工产物较少，而模型可以在损坏后进行预测。

我们还将聂等人(2020)报告的RoBERTa-large模型的结果与在ANLI-NOUN训练集上微调模型获得的结果进行了比较。10我们在原始的、和R3测试集上测量了模型的预测精度，并在表5中报告了结果。预测精度的下降明显大于在多边核方案数据上观察到的下降。因此，数据损坏过程揭示了

10这对应于MNLI+SNLI+发烧+ANLI，去掉所有名词。

**训练数据R1 R2 R3**

安立73.8% 48.9% 44.4%

ANLI-NOUN 57.6% 40.3% 41.0%

表5:在ANLI-NOUN数据集上RoBERTa- large的预测精度(%)。与聂等(2020)在原始ANLI数据集上的结果进行比较。ANLI包含MNLI、SNLI、发烧和ANLI。

作为NLU基准的ANLI数据集的质量提高。然而，该模型能够以57.6%的准确率预测ANLI R1的正确标签这一事实突出表明，即使在训练句子没有多大意义的情况下，该模型也能够从预测一对的la- bel时使用的数据中学习一些因素。这些结果进一步证明了仔细运行诊断的重要性，例如我们评估在NLU任务中使用新基准的重要性。

# 讨论

当前最先进的神经网络模型能否真正理解语言，这个问题目前存在很大争议。神经网络模型在NLU任务中的表现优于人类。我们提出的腐蚀转换经常导致句子变得毫无意义。然而，我们观察到BERT在这些实验中表现得出奇的好。这表明，模型能够从数据中获得其他线索，从而做出正确的预测，而不是理解句子的意思和它们之间的语义关系。

我们提出的诊断测试是评估数据集质量的有用工具，它是评估模型语言理解能力的测试平台。在我们的实验中，他们证明了NLI数据集的质量优于其他数据集。我们在一个额外的实验中测试了这一发现，在该实验中，我们将Pham等人(2020)的单词洗牌机制应用于ANLI数据，该机制被证明不会降低基于BERT的模型在GLUE任务上的性能。我们在表6中的结果表明，这一过程极大地损害了ANLI的模型精度，并带来了额外的证据来支持该数据集优于MNLI(它是GLUE基准的一部分)的质量。

我们的测试套件可以看作是评估工作台质量的附加“碰撞测试”

**测试集R1 R2 R3 ANLI 73.8% 48.9% 44.4% ANLI-SHUFFLE-n1 35.5% 33.8% 36.0% ANLI-SHUFFLE-N2 45.4% 39.8% 37.1%**

ANLI-SHUFFLE-n3 49.4% 40.7% 38.4%

表6:单词重排后RoBERTa- large的预测准确性(%)(Pham等人，2020年)。与在原始ANLI数据集上获得的结果进行比较(聂等人，2020年)。安里-

SHUFFLE-n1/n2/n3测试集包含SHUFFLE n克，分别为n = {1，2，3}。

标记解决常识推理的数据集。它与早期数据集的突出问题并导致ANLI的创建的工作是一脉相承的。我们的提议可以成为构建未来NLI数据集的良好方法的一部分。当前NLI数据集存在的问题的多面性使得研究这些问题非常重要；我们拥有的诊断测试越多，数据集就越有希望变得可靠。事实上，一种类型的测试(仅假设、单词洗牌或单词分类丢弃)并不能消除数据集中存在的所有问题，这突出表明需要多种诊断设备来解决不同的现象。

我们建议在开发新的NLI数据集时，将以下诊断集作为最低健全性检查:

* 仅假设基线(Gururangan等人，2018年；Poliak等人，2018年)
* 词序重排(Pham等人，2020年)
* 交换前提和假设(王等，2019b)
* 词类删除(我们建议的诊断)

回到本文的具体发现，我们进行了一组额外的分析，旨在确定提议的修改的观察到的、相对较小的影响是由于什么。我们探究性能下降是否可以用与被移除的单词类相关的标记数量(较小或较大)来解释。从图1中可以看出，我们比较了BERT的准确性和每个集合中从训练数据中移除的标记数，但是这个因素并不能解释获得的结果。例如，只有492，895个发生-

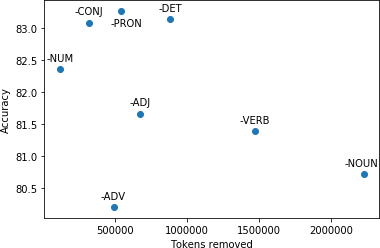
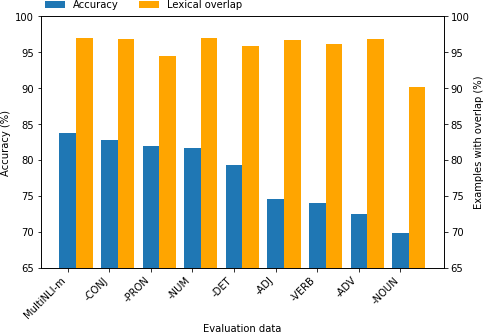
 

图1:基于BERT的模型准确性与移除的令牌的比较。该模型在MNLI训练数据上进行微调，删除特定词类的实例，并在原始MNLI匹配的开发数据上进行评估。

从训练集中移除的副词的百分比，但是与原始结果的差值最高(-3.53分)，而移除886，966个确定者对准确性的影响很小(-0.59分)。这个情节展示了内容词在NLI预测中的重要作用。

周和班萨尔(2020)已经表明前提和假设之间的高度词汇重叠可以指导模型的预测。因此，我们探索了我们的结果在多大程度上可以用腐败测试集中词汇重叠的数量来解释。我们通过计算句子对中前提和假设所共有的标记来衡量词汇重叠。图2中的橙色条显示了原始MNLI和CORRUPT-TEST测试集中前提和假设之间的词汇重叠量(计算在示例总数上的百分比)。蓝色条显示了在每个测试集上评估时，BERT对原始MNLI数据进行微调所获得的预测精度。我们观察到，尽管在一些测试集中(例如，在MNLI-NOUN中)词汇重叠有所减少，但词汇重叠和准确性之间没有明显的相关性，这表明该模型利用了残留在受损句子中的其他线索进行预测。

# 结论

我们提出了一种新的诊断套件，用于评估用于NLI模型训练和评估的数据集的质量。我们表明，数据损坏是估计数据集质量及其影响的有效方法

图2:原始MNLI测试和腐败测试集中模型准确性和复杂重叠的比较。模型根据原始的MNLI训练数据进行微调。

反映模型的真实语言理解能力的潜力。我们在MNLI和ANLI数据集上的结果表明，我们的方法学可以成功地识别高质量或低质量的数据集，即数据集是否触发模型的推理潜力，或者更确切地说，是否允许它们依赖线索和其他统计偏差进行预测。我们提出的测试可用于评估社区使用的现有基准的质量，并相应地解释结果，还可用于指导解决推理任务的新数据集的开发。在后一种情况下，数据一致性将有助于确定数据集构建方法和采用的解释指南是否正确。

最后，尽管比较大量的体系结构会很有意思，但由于空间不足，我们将这种比较留给将来的工作，同时也是为了不混淆读者，因为实验在大量的环境中进行。在本文中，我们还将重点放在MNLI和ANLI数据集上，因为我们主要关注的是尽可能多地覆盖损坏设置。将目前的工作扩展到其他模型和NLU数据集是未来研究的自然的下一步。我们已经使我们的代码可用于促进这一方向的研究。11此外，由于目前的工作对在损坏的数据集上提供的高性能背后的因素存在未解决的问题，我们计划更彻底地研究模型所依赖的线索和假象，并允许它们

11 https://github . com/赫尔辛基-NLP/

数据健全性检查

在这些任务中表现出色。

# 感谢

Marianna Apidianaki和Jo rg Tiede- mann得到了FoTran项目的支持，该项目由欧洲研究理事会(ERC)根据欧洲联盟的地平线2020研究和创新方案(第771113号赠款协议)资助。Stergios

Chatzikyriakidis获得了瑞典研究委员会2014-39年度拨款的支持，该委员会资助了哥德堡大学哲学、语言学和科学理论系的语言理论和概率研究中心(CLASP)。我们感谢审稿人深思熟虑的意见和宝贵的建议。

# 参考

塞缪尔·鲍曼、加博尔·安格利、克里斯托弗·波茨和克里斯托弗·曼宁。2015.https://doi.org/10.18653/v1/D15-1075用于学习自然语言的大型注释语料库。《欧洲自然语言处理会议录》，第632-642页，葡萄牙里斯本。

雅各布·德夫林、张明伟、李健顿和克里斯蒂娜·托塔诺瓦。2019.https://doi.org/10.18653/v1/N19-1423·伯特:用于语言理解的深度双向转换器的预训练。《计算语言学协会北美分会2019年会议记录:人类语言技术》，第1卷(长论文和短论文)，第4171-4186页，明尼苏达州明尼阿波利斯。计算语言学协会。

马克斯·格洛克纳、弗瑞德·施瓦茨和尤夫·戈德堡。2018.用需要简单词汇推理的句子打破nli系统。在美国公民自由联盟议事录中。

苏钦·古兰根、斯瓦巴·斯瓦扬迪普塔、奥马尔·利维、罗伊·施瓦茨、塞缪尔·鲍曼和诺亚·史密斯。2018.自然语言推理数据中的注释工件。《美国国家航空和航天学会会报:高等法律》，第107-112页，路易斯安那州新奥尔良。

约翰·休伊特和珀西·梁。2019.设计和解释带有控制任务的探针。中国香港，第2733-2743页。

爱丽丝·莱和朱莉娅·霍克尼迈尔。2014.伊利诺伊州https://doi.org/10.3115/v1/S14-2055-LH:一种指称和分布方法

语义学。《第八届语义评估国际研讨会论文集》(SemEval 2014)，第329-334页，爱尔兰都柏林。

刘、、奥特、戈亚尔、杜、乔希、陈、利维、、斯托特莱迈尔和斯托扬诺夫。2019.<http://arxiv.org/abs/1907.11692>罗伯塔:稳健优化的伯特预处理方法。

马尔科·马雷利、斯特凡诺·梅尼尼、马尔科·巴罗尼、路易莎·本蒂沃利、拉斐尔·贝尔纳迪和罗伯托·扎姆-帕雷利。2014.组合分布语义模型评估的SICK疗法。《LREC论文集》，第216-223页，冰岛雷克雅未克。

伊欣涅、阿迪娜·威廉姆斯、艾米丽·迪南、莫希特·班萨尔、杰森·韦斯顿和杜维·基拉。2020.https://doi.org/10.18653/v1/2020.acl-main.441·NLI:自然语言理解的新标杆。《计算语言学协会第58届年会论文集》，第4885-4901页，在线。计算语言学协会。

唐明范、忠弼、龙脉和阮。2020.无序:在自然语言理解任务中，句子中单词的顺序有多重要？arXiv预印本arXiv:2012.15180。

亚当·波利亚克、杰森·纳拉多斯基、阿帕拉吉塔·哈尔达尔、雷切尔·鲁丁格和本杰明·范·杜梅。2018.假设只是自然语言推理中的基线。在第七届词汇和计算语义学联合会议论文集中。

Aarne Talman和Stergios Chatzikyriakidis。2019.跨NLI基准测试神经网络模型的泛化能力。在黑盒神经网络程序:分析和解释神经网络神经网络，第85-94页，佛罗伦萨，意大利。

王敬实、亚达·普鲁克萨卡特昆、尼基塔·南加、阿曼普里特·辛格、朱利安·迈克尔、菲利克斯·希尔、奥马尔·利维和塞缪尔·鲍曼。2019a。强力胶:通用语言理解系统的一个更严格的基准。神经信息处理系统进展，第32卷，第3266-3280页。柯伦联合公司。

王敬实、阿曼普里特·辛格、朱利安·迈克尔、菲利克斯·希尔、奥马尔·利维和塞缪尔·鲍曼。2018.https://doi.org/10.18653/v1/W18-5446胶水:自然语言理解的多任务基准和分析平台。在黑盒神经网络程序:分析和解释神经网络神经网络，第353-355页，布鲁塞尔，比利时。

王、和邢鹏。2019b。如果我们简单地交换两个文本片段呢？一种直截了当但有效的方法来测试方法的鲁棒性，以混淆自然语言中的信号

推理任务。《AAAI人工智能会议录》，第33卷，第7136-7143页。

阿迪娜·威廉姆斯、尼基塔·南加和塞缪尔·鲍曼。2018.基于推理的句子理解大范围挑战语料库。在全国公民自由联盟的支持下。

象州和莫希班萨尔。2020.对抗词汇数据集偏差的稳健NLI模型。计算语言学协会第58届年会论文集，第8759-8771页。计算语言学协会。

# 附录

表1包含了来自MNLI-NOUN测试集的句子对的例子，BERT为其预测了正确的标签。表2包含从损坏的MNLI数据集中移除的令牌数量的统计信息。表3包含从损坏的ANLI测试集中移除的令牌数量的统计信息。

**标签前提假设**

发展中的矛盾与适宜。

矛盾像和一样，警告，日本人加入了，但锁定在传统。

*不鼓励和任何人商量。而日本人之间没有。*

矛盾可以肯定的是，并不是所有人都是。每一首都是。

蕴涵，或者在哪里？储蓄的价值。

蕴涵在原著中，是由他设置然后伏击的

*被一个不怀好意的人点名，而当他试图以口若悬河的口气回答时(就是握紧拳头。*

另一个是回顾性的，旨在帮助那些审查者评估已完成的。

*出去抓他。*

*它是用来帮助评估的。*

中立，呃，它让你远离你，因为它更像是-

*对你很重要。*

中立者在美国的一个地方被发现。盒子里的是粉末状的。

在另一种情况下，节拍就是节拍。他们是更好的。

表1:从MNLI-NOUN测试集中随机选择的句子对，BERT为其预测了正确的标签。

**训练数据集测试数据集**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **配置** | **前提** | **假设** | **总数** | **前提** | **假设** | **总数** |
| MNLI-NUM | 119,587 | 44,289 | 163,876 | 3,100 | 1,133 | 4,233 |
| MNLI-CONJ | 320,210 | 76,466 | 396,676 | 7,584 | 1,874 | 9,458 |
| MNLI-ADV | 492,895 | 237,250 | 730,145 | 11,777 | 5,862 | 17,639 |
| MNLI-PRON | 543,968 | 301,293 | 845,261 | 13,060 | 7,466 | 20,526 |
| MNLI-ADJ | 677,095 | 302,652 | 979,747 | 16,162 | 7,562 | 23,724 |
| MNLI-DET | 886,966 | 483,238 | 1,370,204 | 21,198 | 11,723 | 32,921 |
| MNLI-动词 | 1,474,454 | 886,597 | 2,361,051 | 35,813 | 22,101 | 57,914 |
| MNLI-NOUN | 2,228,780 | 1,090,814 | 3,319,594 | 54,700 | 27,182 | 81,882 |
| PRON | 2,772,748 | 1,392,107 | 4,164,855 | 67,760 | 34,648 | 102,408 |
| 名词+PRON+动词 | 4,501,189 | 2,166,146 | 6,667,335 | 109,325 | 53,647 | 162,972 |
| 名词+ADV+动词 | 4,552,262 | 2,230,189 | 6,782,451 | 110,608 | 55,251 | 165,859 |
| 名词+动词 | 5,045,157 | 2,467,439 | 7,512,596 | 122,385 | 61,113 | 183,498 |
| 名词+动词+ADJ | 4,368,062 | 2,164,787 | 6,532,849 | 106,223 | 53,551 | 159,774 |
| 名词+动词+ADV+ADJ | 3,875,167 | 1,927,537 | 5,802,704 | 94,446 | 47,689 | 142,135 |

表2:从MNLI中移除标记形成的数据集。这些数字对应于从前提和假设中移除的令牌数，以及移除的令牌总数。

**R1 R2 R3**

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **测试数据集** | **前提** | **假设** | **总数** | **前提** | **假设** | **总数** | **前提** | **假设** | **总数** |
| ANLI-NOUN | 23,523 | 4,719 | 28,242 | 23,646 | 4,275 | 27,921 | 23,086 | 4,033 | 27,119 |
| 动词 | 6,057 | 1,657 | 7,714 | 6,155 | 1,668 | 7,823 | 11,281 | 2,258 | 13,539 |
| 安立-PRON | 1,567 | 184 | 1,751 | 1,657 | 178 | 1,835 | 4,152 | 446 | 4,598 |
| 安立-ADJ | 2,827 | 514 | 3,341 | 2,783 | 495 | 3,278 | 3,525 | 625 | 4,150 |
| 安立-ADV | 899 | 267 | 1,166 | 917 | 313 | 1,230 | 2,898 | 470 | 3,368 |
| ANLI-NUM | 2,934 | 576 | 3,510 | 2,862 | 515 | 3,377 | 1,737 | 286 | 2,023 |
| ANLI-CONJ | 1,816 | 161 | 1,977 | 1,897 | 122 | 2,019 | 2,073 | 142 | 2,215 |
| 安立-DET | 5,631 | 1,195 | 6,826 | 5,669 | 1,086 | 6,755 | 7167 | 1,406 | 8,573 |

表3:从ANLI测试集中移除标记形成的数据集。这些数字对应于从前提和假设中移除的令牌数，以及三个数据集(轮次)中移除的令牌总数。