语境化扰动的文本对抗性攻击

# 滇池*生活* 张一哲*◇* 郝*鹏飞* 陈立*群品* 克里斯-布罗凯特*◇* 孙明庭比尔-多兰*◇*

*f*华盛顿大学 *◇*微软研究 杜克大学

{dianqili, mts}@uw.edu, [hapeng@cs.uw.edu,](mailto:hapeng@cs.uw.edu) [liqun.chen@duke.edu](mailto:liqun.chen@duke.edu)

{Yizhe.Zhang, Chris.Brockett, billdol}@microsoft.com

# 摘要

对抗性例子暴露了自然语言处理(NLP)模型的脆弱性，可以用来评估和证明它们的鲁棒性。现有的生成这种例子的技术通常是由局部启发式规则驱动的，这些规则对上下文是可知的，往往导致不自然和不符合语法的输出。本文介绍了CLARE，一个上下文化的AdversaRial例证生成模型，它通过一个掩码--然后--再掩码的过程来产生语气和语法输出。CLARE建立在一个预先训练好的屏蔽语言模型上，并以上下文感知的方式修改输入。我们提出了三种上下文化的扰动，*替换*、*插入*和*合并*，允许生成不同长度的输出。CLARE可以灵活地组合这些扰动，并将其应用于输入中的任何位置，从而能够以较少的代价更有效地攻击受害者模型。广泛的实验和人工评估表明，CLARE在攻击成功率、文本相似性、词性和语法性方面都优于基线。

# 介紹

用于自然语言处理（NLP）任务的对抗性示例生成旨在扰动输入文本以触发机器学习模型的错误，同时保持输出接近原始。Be- sides exposing system vulnerabilities and help improve their robustness and security ([Zhao etal.](#_bookmark84) , [2018](#_bookmark84); [Wallaceet al.](#_bookmark70) , [2019](#_bookmark70); [Cheng et al](#_bookmark30)[.](#_bookmark41) , [2019](#_bookmark30); [Jia et al.](#_bookmark41) , [2019](#_bookmark41), *inter alia*), adversarial examples are also used to analyze and interpret the models' deci- sions ([Jia and Liang](#_bookmark40), [2017](#_bookmark40); [Ribeiro et al.](#_bookmark64) , [2018](#_bookmark64)).

为NLP任务生成对抗性例子可能具有挑战性，部分原因是由于自然语言文本的离散性。最近的努力探索了启发式规则，例如用同义词替换标记（[Samanta和Mehta](#_bookmark65)，[2017](#_bookmark65)。

***原文***

图1：CLARE的说明。通过掩码--再掩码的过程，模型通过三种上下文化的扰动来生成对抗性文本。***替换***，***插入***和***合并***。掩码用""表示。淡化程度对应于inﬁll标记的（递减）优先级。



超级蚂蚁群落打

澳大利亚**{沿海地区。**

一个**{巨大的。**

**,**

**,**

**}**.

**}**100公里

蚁群会威胁到

**{insect, , }**种。

**CLARE:**

**情境化扰动**

***对抗性案文***

超级蚁群袭击澳大利亚 。

***巨大的***100公里

蚁群会威胁到

***当地昆虫***种类。

[Liang等人](#_bookmark50)，[2019](#_bookmark50)；[Alzantot等人](#_bookmark28)，[2018](#_bookmark28)；[Ren等人](#_bookmark62)，[2019](#_bookmark62)；[Jin 等人](#_bookmark42)，[2020](#_bookmark42)，*等等*）。)尽管取得了一些经验上的成功，但基于规则的方法对上下文是不可知的，限制了它们产生natu- ral，ﬂuent和语法输出的能力（[Wang等人](#_bookmark74)，[2019b](#_bookmark74)；[Kurita等人](#_bookmark46)，[2020](#_bookmark46)，*除其他外*）。

本作品介绍了CLARE，一个针对文本的ContextuaLized AdversaRial Example生成模型。CLARE通过掩码--再--插入的过程对输入进行扰动：它首先检测模型的漏洞，并将掩码部署到输入中，以确定缺失的文本，然后使用预先训练的掩码语言模型（例如，RoBERTa；[Liu等人](#_bookmark51)，[2019](#_bookmark51)）插入一个替代方案。CLARE具有三种上下文化的扰动功能。*Replace*、*Insert*和*Merge*，分别替换一个token、插入一个新的token和合并一个bigram（图[1](#_bookmark0)）。 [Alzantot 等人](#_bookmark28) ，[2018](#_bookmark28)；[Ren 等人](#_bookmark62) ，[2019](#_bookmark62)。

[Jin et al.](#_bookmark42) , [2020](#_bookmark42)）。)此外，CLARE在更广泛的攻击策略上进行搜索，因此能够以更少的编辑来更有效地攻击受害者模型。在掩蔽语言模型的基础上，CLARE最大限度地保留了输出的文本相似性、ﬂu- ency和语法性。

我们通过攻击ﬁnetuned BERT模型（[De- vlin等人](#_bookmark31)，[2019](#_bookmark31)），对CLARE在文本分类、自然语言推理和句子转述任务上进行评估。广泛的实验和胡人评估结果表明，CLARE在攻击成功率、tex- tual相似性、ﬂuency和语法性方面优于基线，并在攻击成功率和保存输入输出相似性之间取得了更好的平衡。我们的分析进一步表明，当可用的训练数据有限时，CLARE可以用来提高下行模型的鲁棒性，并提高其准确性。我们发布了

tween **x***/* and **x** to be larger than a threshold: sim(**x***/,* **x**) *> l. A* common choice of sim( , ) is to encode sentences using neural networks, and

计算它们在嵌入空间中的余弦相似度（[Jin等](#_bookmark42)，[2020](#_bookmark42)）。

***● ●***

## 遮蔽和情境化的内涵发展。

在输入序列的某一位置，CLARE可以执行三种扰动动作。*替换*，*插入*和*合并*，我们在这一节介绍。这些动作以不同的策略在给定的位置应用掩码，然后根据未掩码的上下文来填充缺失的文本。

***替换*。***替换：替换*动作将给定位置*i*的标记替换为另一种选择(例如，将"The movie is *fan- tastic*."中的"*fantastic"替换*为"*amazing*")。它首先用掩码替换*xi*，然后是

从候选*集多中*选择一个标记*z，并将其放入*其中。

我们的代码和模型在[https://github.com/](https://github.com/cookielee77/CLARE)

**x**～ = x1 *... ... xi*-1。

[MASK]*x*

*i*+1

*.第十。*

[cookielee77/CLARE](https://github.com/cookielee77/CLARE)。

# CLARE

在一个较高的层次上，CLARE将一系列的con- textualized perturbation动作应用到输入上。每一个动作都可以被看作是一个*局部*掩码--然后再填充的过程：它首先在给定的po- sition周围对输入施加一个掩码，然后用一个预先训练好的掩码语言模型填充它（§[2.1](#_bookmark1)）。为了产生输出，CLARE对动作进行打分和降级排序，然后迭代应用到输入中（§[2.2](#_bookmark5)）。我们先简单回顾一下背景，并铺陈一下必要的符号。

**背景：我们认为，对抗式例子的生成是围绕着受害者模型f而展开的。**对抗式实例生成的中心是一个**受害者**模型*f* ，我们假设它是一个文本分类器。我们专注于黑盒设置，允许访问*f*的输出，但*不允许*访问它的conﬁgurations，如参数。给定一个输入序列 **x** = x1x2 *... ... xn* 和它的标签 *y*

(假设*f*(**x**)=*y*)，一个**对抗性的例子x***/。*

应该是修改**x**，以触发在

受害人模式：*f*（**x***/*）*f*（**x**）。同时：

取代(**x***，i*)=x1 *...... xi*-1 *z xi*+1 ....... *xn.*

为了清楚起见，我们用*xz表示*替换（**x***，i*）。要产生一个对立的例子。

～

* *z*应ﬁt入未屏蔽的语境。
  + **x**应与**x**相似。

～*z*

* + 应该在...中触发错误。

～*xz*

*f*

这些都可以通过选择*z来*实现，如

* + *z*从掩蔽语言模型中获得高概率：pMLM(*z |* **x**) *> k。*

～ ～～• 类似于: ;

*xz* **x** sim(**x***, xz*) *> l*

* + *f*预测金标给定*xz*的概率较低，即*pf* (*y | xz*)较小。

～ ～表示一个预先训练好的屏蔽语言模型

pMLM

(例如，RoBERTa；[Liu等人](#_bookmark51)，[2019](#_bookmark51))。使用更高的*k*，*l*阈值产生更多的ﬂuent和更接近原始的输出。然而，这可能会低估攻击的成功率。我们选择*k*，*l*来权衡这两个方面。[2](#_bookmark3)

候选集的构建可以满足前两个要求。*多* =

『～ ～←

*z/ ∈ ì |* pMLM(*z/ |* **x**)*>k，sim*(x，*xz′ )>*l *。*

*ì*是掩饰语言模型的词汇。

*多*

文本修改应该是最小的，这样**x***/*就会接近**x**，而人类对**x***/*的预测保持不变。[1](#_bookmark2)

这一点是通过要求相似度b--来实现的。

为了满足第三种情况，我们选择的标记，如果ﬁlled in，将导致最"混乱"的*f* 。

*z* = arg min *pf* (*y* *|* **x** ～*z′* )*.*

(1)

*×*

1在计算机视觉应用中，轻微的扰动对

*z′ez*

连续的像素对人类来说几乎无法感知，因此人们很难区分**x**和**x***′*（[Goodfellow等](#_bookmark37)，[2015](#_bookmark37)）。然而，对于文本来说，情况并非如此，因为在文本中，对

离散的代币更容易被人类注意到。

2←和é根据经验分别设为5 10*一*3和0．7。这也减少了计算量：在我们的实验中，*|z|*平均为42，远小于VO-。

腔体大小（*|v|* = 50 ，265）。

*插入*和*合并*操作在掩蔽策略上与*重新放置*不同。更改原生标记*z的*选择与*替换*操作的选择类似。

***插入*。**目的是为输入添加额外信息（例如，将"我建议."改为"我强烈建议."）。它在*xi*后面插入一个掩码，然后ﬁlls它。略微超载的符号。

[M

] *x*

*.. . x*

*,*

**算法1** 通过CLARE进行对抗性攻击

1：**输入。**文本标签对(**x***，y*)；受害者模型*f。*

2：**产出。**一个对抗性的例子

3：**初始化：x**(0)=**x。**

4：*仁*a

5: **for** 1 *< i < |x| do*

6:*a 仁*最高分的动作来自于*{。*

repla ce(**x***, i*), insert(**x***, i*)*,* merge(**x***, i*)*}。*

*a 仁a*

7:

8：**结束**

*{a}*

**x**～ = x1 *... ... xi*

ASK

*i*+*1n*

9: **for** 1 *< t < T do*

插入(**x***，i*)=x1 *...... xi z xi*+1 *....... xn。*

这就增加了1的序列长度。

***合并*:**这就遮住了一个大ram *x x*

*i*

*i*+1

带着

10:*a 仁的*最高得分行动从*一个*

11:*a 仁a {a}。*

12:**x**(*t*) *仁*在**x**(*t*-1)上应用*a。*

13:**如果***f*(**x**(*t*))*/*=*y***则返回x**(*t*)

*单个*掩码，然后将其删除，将序列长度减少1。

14:**结束如果**

15：**结束**

16：**返回**NONE

**x**～ = x1 *... ... xi*-1。

[MASK]*x*

*i*+2

*.第十。*

merge (**x***, i*) = x1 *... ... xi*-1 *z xi*+2 ... ..*. xn.*

*z*可以和其中一个被遮蔽的标记相同（例如，遮蔽掉"New York"，然后填入"York"）。这可以看作是从输入中删除一个标记。

*插入*和*合并*时，*z的*选择与替换操作相同。[3](#_bookmark6)

总之，在输入序列的每个位置*i*，CLAREﬁrst。(*i*)用掩码替换*xi*；(*ii*)或在*xi*后面插入一个掩码；(*iii*)或将*xixi*+1合并成一个掩码。然后，用掩码语言模型和一个tex-的候选tokens组成一组候选tokens。

的相似性函数；选择金标概率最小的标记作为备选标记。这三种操作的组合可以实现任意两个序列之间的转换。

CLARE首先为所有位置平行地构造局部动作，即位置*i*的动作不影响其他位置的动作。然后，为了产生对抗性的例子，CLARE收集局部动作并选择一个顺序来执行它们。

## 顺序应用扰动。

给定一个输入对（**x***，y*），让*n表示***x**的长度，CLARE从*3n*个动作中选择产生输出。每一个位置有3个动作，假设候选标记集不是空的。我们的目标是生成一个对输入有最小修改的对抗性例子。为了实现这一目标，我们反复应用动作，并首先选择那些具有最小修改的动作。

3如果一个扰动的候选令牌集是空的，则不会被考虑。

最小化从*f*输出金标*y*的概率 。

每个动作都有一个分数，衡量它"混淆"*f*的可能性有多大：用*a*(**x**)表示对**x**应用动作*a*的输出。

用*a*(**x**)作为输入，从*f*预测金标的负概率。

╱←

*s*(**x***,y*)(*a*) = *-pf y | a*(**x**)*.*

在每个位置上*只能*应用三个动作中的*一个*，我们选择得分最高的一个。这个约束的目的是避免在同一个位置上出现多个修改，例如，将"纽约"合并为"西雅图"，然后用"波士顿"代替。

行动反复应用到输入中，直到找到一个对抗性的例子或达到行动的极限*T*。每一步都从剩下的行动中选择得分最高的行动。算法[1](#_bookmark4)对上述过程进行了总结。[4](#_bookmark7)

**讨论。**CLARE的一个关键技术是局部掩码-时间-内尔扰动。与现有的上下文诊断替换ap- proaches（[Alzantot等人](#_bookmark28)，[2018年](#_bookmark28)；[Jin等人](#_bookmark42)，[2020年](#_bookmark42)；[Ren等人](#_bookmark62)，[2019](#_bookmark62)年*等*）相比，上下文化inﬁll- ing产生了更多的语言和语法输出。掩盖的语言模型生成对抗性例子也被同时进行的工作所探索。

4*插入*和*合并*操作会改变文本长度。当其中任何一个动作被应用时，我们相应地改变*A*中剩余的受影响动作的文本索引。

BERTAttack([Li et al.](#_bookmark49) , [2020](#_bookmark49))和BAE([Gargand Ramakrishnan](#_bookmark35), [2020](#_bookmark35))。[5](#_bookmark9)

* + - BERTAttack只替换代币，因此只能产生与输入相同长度的输出。这与CLARE模型的*替换*动作类似。BAE将替换和插入令牌纠缠在一起：它*只*在被替换的令牌邻近的位置插入，限制了它的攻击能力。与两者不同的是，CLARE使用了三种不同的扰动（*Replace*、*Insert*和*Merge*），每种扰动都允许对输入的*任何*位置进行有效的攻击，并且可以产生不同长度的输出。正如我们将在实验（§[3.3](#_bookmark14)）中显示的那样，CLARE的性能优于这两种方法。
    - 在选择攻击位置的时候，BERTAttack和BAE都没有考虑到要植入的代币，而CLARE则考虑到了。根据我们的消融研究(§[4.1](#_bookmark18))，这导致更好的对抗性攻击perfor- mance。
    - CLARE展示了使用RoBERTa比使用BERT的优势，BERT是在协约作品中使用的（§[4.1](#_bookmark18)）。

# 实验

我们在文本分类、自然语言推理和句子转述任务上对CLARE进行评估。我们首先描述CLARE的实现细节和基线（§[3.1](#_bookmark8)）。§[3.2](#_bookmark11)介绍了实验数据集和评估指标；§[3.3](#_bookmark14)总结了结果。

## 设置

* + - 我们用 RoBERTa（RoBERTadistill；[Sanh 等人](#_bookmark66) ，[2019](#_bookmark66)）的精馏版本作为上下文化的 inﬁlling 的屏蔽语言模型进行实验。我们还在消融研究（§[4.1）](#_bookmark18)中与基础大小的RoBERTa（RoBERTabase；[Liu等人](#_bookmark51)，[2019](#_bookmark51)）和基础大小的BERT（BERTbase；[Devlin等人](#_bookmark31)，[2019](#_bookmark31)）进行比较。
    - 相似性函数建立在通用句子编码器（USE；[Cer等](#_bookmark29)，[2018](#_bookmark29)）的基础上。
    - 受害者模型是在BERTbase之上的MLP分类器。它将第一个token的上下文化表示作为输入。在训练受害者模型时，我们对BERT进行了精确调整。

**基线。**我们将CLARE与最近最先进的字级黑盒对抗式方法进行比较。

5[Li等人](#_bookmark49)（[2020](#_bookmark49)）和[Garg和Ramakrishnan](#_bookmark35)（[2020](#_bookmark35)）都与本工作的初步报告同时发表。

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **数据集** | **平均长度长度** | **# ＃类** | **火车** | **测试** | **Acc** |
| Yelp | 130 | 2 | 560K | 38K | 95.9% |
| AG新闻 | 46 | 4 | 120K | 7.6K | 95.0% |
| [MNLI6](#_bookmark12) | 23/11 | 3 | 392K | 9.8K | 84.3% |
| QNLI | 11/31 | 2 | 105K | 5.4K | 91.4% |

表1：数据集的一些统计数据。最后一列表示受害者模型在*没有*对抗性攻击*的情况下*对原始测试集的准确性。

攻击模式，包括：

* + - **TextFooler**：[Jin等人](#_bookmark42)（[2020](#_bookmark42)）的最先进模型。它用它们的同义词替换了来自于反语词嵌入（[Mrkšic´等人](#_bookmark55)，[2016](#_bookmark55)）的标记，并使用与我们工作相同的文本相似性函数。
    - TextFooler**+LM**：我们基于[Alzan- tot等人](#_bookmark28)（[2018](#_bookmark28)）和[Cheng等人](#_bookmark30)（[2019](#_bookmark30)）实现的TextFooler的改进变体。这继承了TextFooler的令牌替换，但使用额外的小尺寸GPT-2 lan- guage模型（[Radford等人](#_bookmark60)，[2019](#_bookmark60)）来筛选出那些不在上下文中的候选令牌，并计算出迷惑性。
    - **BERTAttack**：[Li等人](#_bookmark49)([2020](#_bookmark49))提出的一种掩盖-然后-in-ll方法。它贪婪地用BERT的预测替换to- kens。BAE没有列出，因为它的性能与BERTAttack相似（[Garg和Ramakrishnan](#_bookmark35)，[2020](#_bookmark35)）。

我们使用作者提供的上述基线的开源实现。更多细节载于附录[A.1](#_bookmark88)。

## 数据集和评估

**数据集。**我们评估CLARE与以下- ing数据集。

* + - **Yelp Reviews** ([Zhang et al.](#_bookmark82) , [2015](#_bookmark82)): a bi- nary sentiment classiﬁcation dataset based on restaurant reviews.
    - **AG新闻**（[Zhang et al.](#_bookmark82) , [2015](#_bookmark82)）：收集了四类新闻文章。*世界*、*体育*、*商业*和*科技。*
    - **MNLI**（[Williams et al.](#_bookmark75) ，[2018](#_bookmark75)）：一个自然语言推理数据集。每个实例都包含一个前提-假设对，模型应该从一个*包含*、*中立*和*矛盾*的标签集来确定它们之间的关系。它涵盖了各种领域的文本。

6我们只考察匹配集的性能，因为不匹配集更容易被攻击。

**Yelp**(PPL=51.5)**AG新闻**(PPL=62.8)

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **型号**  TextFooler | **A-***率个*  77.0 | *Mod1PPL1*  16.6 163.3 | | *GErr1*  1.23 | **Sim***个*  0.70 | **A-***率个*  56.1 | *Mod1*  23.3 | *PPL1*  331.3 | *GErr1*  1.43 | **Sim***个*  0.69 |
| LM + 34.0 | | 17.4 | 90.0 | 1.21 | 0.73 | 23.1 | 21.9 | 144.6 | 1.07 | 0.74 |
| BERTAttack71.8 | | 10.7 | 90.8 | 0.27 | 0.72 | 63.4 | 7.9 | 90.6 | 0.25 | 0.71 |
| **CLARE79.7** | | **10.3** | **83.5** | **0.25** | **0.78** | **79.1** | **6.1** | **86.0** | **0.17** | **0.76** |
| **MNLI**(PPL=60.9)**QNLI**(PPL=46.0) | | | | | | | | | | |
| **型号** | **A-***率个* | *Mod1PPL1* | | *GErr1* | **Sim***个* | **A-***率个* | *Mod1PPL1* | | *GErr1* | **Sim***个* |
| TextFooler | 59.8 | 13.8 161.5 | | 0.63 | 0.73 | 57.8 | 16.9 164.4 | | 0.62 | 0.72 |
| LM + 32.3 | | 12.4 | 91.9 | 0.50 | 0.77 | 29.2 | 17.3 | 85.0 | 0.42 | 0.75 |
| BERTAttack82.7 | | 8.4 | 86.7 | 0.04 | 0.77 | 76.7 | 13.3 | 86.5 | 0.03 | 0.73 |
| **CLARE88.1** | | **7.5** | **82.7** | **0.02** | **0.82** | **83.8** | **11.8** | **76.7** | **0.01** | **0.78** |

表2：对抗式实例生成在攻击成功率（A-rate）、修改率（Mod）、困惑度（PPL）、增加的语法错误数（GErr）和文本相似度（Sim）方面的表现。每个数据集的原始输入的迷惑性在括号中表示。粗体字表示每个数据集的最佳性能

衡量标准。所有数字都是在1000个测试实例上报告的。*个*（*↓*）代表越高（越低）越好。

* + - **QNLI** ([Wanget al.](#_bookmark72) , [2019a](#_bookmark72))：一个由斯坦福问题回答数据集([Rajpurkar et al.](#_bookmark61) , [2016](#_bookmark61))转换而来的二元分类- ﬁcation数据集。其任务是确定上下文是否包含问题的答案。它主要基于英文维基百科文章。

表[1](#_bookmark10)总结了数据集的一些统计数据。除了上述四个数据集，我们还实验了DBpedia本体数据集（[Zhang等](#_bookmark82)，[2015](#_bookmark82)）、Stanford sentiment treebank（[Socher等](#_bookmark67)，[2013](#_bookmark67)）、Microsoft Research Paraphrase Cor- pus（[Dolan和Brockett](#_bookmark32)，[2005](#_bookmark32)）和GLUE基准的Quoraes- tion Pairs。这些数据集的结果总结在附录[A.2](#_bookmark93)中。

按照之前的实践（[Alzantot等人](#_bookmark28)，[2018](#_bookmark28)），我们在训练数据上ﬁne-tune CLARE，并用1000个随机采样的长度为100的测试语句进行评估。在句子对任务中（如MNLI，QNLI），我们攻击较长的句子，排除两者中出现的标记。

*<*

**评价指标。**我们遵循之前的工作（[Jin等人](#_bookmark42)，[2020](#_bookmark42)；[Morris等人](#_bookmark53)，[2020a](#_bookmark53)），用以下自动指标来评估模型。

* + - **攻击成功率（A-rate）**：能够成功攻击受害者模型的对抗性例子的百分比。
    - **修改率（Mod）**：修改过的令牌的百分比。每一个*替换*或*插入*操作都会有一个标记被修改；如果两个合并的标记中的一个被保留，则合并操作被认为是修改了一个标记（例如，将bigram *ab*合并为*a*），否则就是修改了两个标记。

(如：将bigram *ab*合并为*c*)。

* + - **迷惑性（PPL）**：用于评估对手的*敏捷性的度量*（[Kann 等](#_bookmark44) ，[2018](#_bookmark44)；[Zang等](#_bookmark79) ，[2020](#_bookmark79)）。迷惑性是使用小尺寸的GPT-2与50K大小的词汇计算出来的（[Radford et al.](#_bookmark60) ，[2019](#_bookmark60)）。
    - **语法错误(GErr)**：与原文相比，成功的对抗性例子中增加的语法错误的绝对数。按照（[Zang等](#_bookmark79)，[2020](#_bookmark79)；[Morris等](#_bookmark54)，[2020b](#_bookmark54)），我们通过LanguageTool（[Naber等](#_bookmark56)，[2003](#_bookmark56)）来计算。[7](#_bookmark15)
    - **文本相似度（Sim）**：输入与其对手之间的余弦相似度。Fol- lowing（[Jin等人](#_bookmark42)，[2020](#_bookmark42)；[Morris等人](#_bookmark54)，[2020b](#_bookmark54)），我们使用通用句子编码器（USE；[Cer 等人](#_bookmark29)，[2018](#_bookmark29)）来计算这个。

最后四个指标是对那些成功攻击受害者模型的对抗性例子的平均值。

## 结果

表[2](#_bookmark13)总结了结果。总体而言，CLARE在不同数据集的所有指标上都取得了最佳性能。值得注意的是，CLARE比最强的基线BERTAttack的攻击成功率高5.4%以上，而且对文本的平均修改*更少*。我们将此归功于CLARE在任何po- sition下结合三种不同的扰动而获得的灵活的攻击策略。有趣的是，使用上下文化的嵌入- dings似乎*并不能*保证更好的敏捷性。

7 [https://www.languagetool.org/。](https://www.languagetool.org/)

0.8

CLARE

BERTAttack TextFooler TextFooler+LM

文本相似性

0.7

50

150

CLARE

BERTAttack TextFooler TextFooler+LM

Metric CLARE中性文字傻瓜

Similarity 56.1*士*2*●*528.1 15.8 *士*2*●*1 流利性和语法性 42.5 士2*●548.*6 8.9 士*1●*5 标签 68.0Consistency 士2*●470.*1- 士2*●*5

0.6

250

350

困惑

表3：AG新闻数据集上的人类评价表现以百分比表示，表示95%的一致性水平的一致性间值。

0.5

*±*

20% 40% 60% 80% 100%

攻击成功率

20% 40% 60% 80% 100%

攻击成功率

图2：**左**：攻击成功率和文本模拟性权衡曲线（*都是越高越好*）。**右：攻击成功率（越高越好）和困惑度（越低越好）权衡曲线**。攻击成功率（*越高越好*）和困惑度（*越低越好*）权衡曲线。两条曲线下的面积越大，说明两个指标之间的权衡越好。

尽管对文本的修改较少，但BERTAt- tack在四个数据集中的三个数据集上实现了与语言模型增强的TextFooler相似的迷惑性，而CLARE的表现一直优于两者。在语法错误方面，语境化模型(CLARE和BERTAttack)比其他模型次要好，其中CLARE per-形成最好。在相似性方面，CLARE优于所有基线超过0.02，比BERTAttack对TextFooler变体的改进差距更大。我们在附录[A.2](#_bookmark93)中的其他数据集上也观察到了类似的趋势。

图 [2 比较了](#_bookmark16)攻击成功率和文本相似性之间的权衡曲线。我们调整了构建候选令牌集的阈值，并绘制了文本相似性与攻击成功率的对比图。CLARE取得了最好的平衡，在成功率上表现出明显的优势，而相似度下降最小。我们观察到攻击成功率和困惑度交易的相似趋势。

**人工评估。**我们进一步对AG新闻数据集进行人工评估。我们随机抽取了300个实例，CLARE和TextFooler都成功攻击了这些实例。对于每一个输入，我们将两个模型的对抗性例子进行配对，并将它们与原始输入和黄金标签一起呈现给众评委。我们询问他们，他们更喜欢哪个中性选项，在（1）具有与原始输入更接近的意义（相似性），以及（2）更ﬂuent和语法（ﬂuency和grammatical- ity）方面。此外，我们要求评委对对抗性例子进行注释，并将他们的注释与黄金标签（标签一致性）进行比较。我们为每个评价对象收集5个回答。

上升的方面。更多细节见附录 [A.3](#_bookmark94)。如表 [3](#_bookmark17) 所示，CLARE 比 TextFooler 有明显的优势：在相似性方面，56%的回答喜欢 CLARE，而 16%喜欢 TextFooler。词汇量和语法性的趋势相似（42% vs. 9%）。这一观察结果与自动指标的结果是一致的。在标签一致性方面，CLARE的表现略逊于TextFooler，在95%的置信度inter- val（CI）下，CLARE的表现为68%（66%*，*70%），而在95%的CI下，CLARE的表现为70%。

(68%*,* 73%).我们将其归结为内在的过度。

圈的一些类别的AG新闻数据集，如*科技和商业，*证明原始输入的标签一致性为71%。

在本节的最后，表[4](#_bookmark20)比较了TextFooler和CLARE生成的广告实例。更多的例子列在附录[A.4](#_bookmark96)中。

# 分析报告

本节首先进行消融研究（§[4.1](#_bookmark18)）。然后我们在§[4.3](#_bookmark24)中探讨CLARE用于提高下游模型的鲁棒性和准确性的潜力。在§[4.2](#_bookmark23)中，我们实证观察到CLARE倾向于攻击名词和名词短语。

## 消融研究

我们消解CLARE的每个组件来研究其有效性。我们在随机抽取的1000个AG新闻实例上进行评估（§[3.2](#_bookmark11)）。结果总结在表[5](#_bookmark21)中。

我们首先研究了三种per- turbations分别应用时的性能。在三种编辑策略中，使用INSERTONLY获得了最好的性能，REPLACEONLY紧随其后。MERGEONLY的性能低于其他两种，部分原因是攻击仅限于bigram名词短语（§[3.1](#_bookmark8)）。综合所有三种扰动，CLARE以最少的修改达到了最好的性能。

8*Merge* perturbation只能合并名词短语，由NLTK工具箱[(](https://www.nltk.org/)https://www.nltk.org/)牵引。我们认为这有助于产生更多的语法输出。

Sprint Corp.正在与高通公司进行谈判。

**AG**关于使用芯片制造商正在构建的网络。

**Module A率** *个Mod1 PPL1 GErr1* **Sim** *个*

CLARE79.**6.1 186.**0.17 00.76

(Sci&Tech)为Sprint手机提供电视直播服务

客户。

Sprint *Corps*.正在与高通公司谈判。

[MERGEONLY8](#_bookmark19) 47.2 6.2 95.**0.08 30.79**

仅插入68.1 7.2 93.1 0.23 0.74

TextFooler (商业)

CLARE

(业务)

关于*经营*网络的芯片制造商是*con- solidation做viva*电视给Sprint mo- bile电话客户。

斯普林特公司正在与高通公司商谈利用高通公司正在建设的网络向斯普林特移动电话提供有线电视的事宜。

66.7 7.7 85.6 0.10 REPLACEONLY0.72

63.4 7.9 90.6 0.25 BERTAttack0.71

*w/o* sim > é0.82.4 6.9 86.8 0.13 70

*无*扩增MLM > ←95.**70.**6.8 162.8 0.71 61

表5：消融研究结果。"*w/o* sim *> l"* ab-?

客户。

*前提是*：让我试试。她开始啪啪啪

在构建候选集时，确定了文本相似性约束，而"*w/o* pMLM *> k"*则取消了这一约束。

**MNLI**

(中性)

ﬁngers和说这个词急切，但什么都没有

发生了。

屏蔽语言模型概率约束。

*假设*。她变得很沮丧，当

TextFooler（Contra-

咒语没有工作。

*前提*：*授权*我*尝试*。她*触发了拍打小指*，急切地*说着*这个词，但什么也没有*产生*。

**MLM A-rate** *个 Mod1 PPL1* **Sim** *个* **Speed** *个*

RoBERTadistill79.**6.1 86.0 0.76 10.14**

**RoBERTabase79.**6.3 88.9 0.75 30.07

78.4 8.3 95.2 0.71 BERTbase0.06

词法)

*假设*。她变得很沮丧，当

咒语没有工作。

CLARE

(contra- diction)

*前提是*：让我来试试。她开始啪啪啪地急切地说着这个词，但没有发生什么**意外。**

*假设*：咒语失效后她变得很沮丧。

表6.CLARE与不同屏蔽语言模型(MLM)实现的结果。用不同的屏蔽语言模型（MLM）实现CLARE的结果。**速度以**每秒处理的样本数来衡量。

## 语篇标签的扰动（Perturbations by Part-of-speech Tags

表4：不同模型产生的对抗性例子。原始的金色标签显示在（加粗的）数据集名称下面。***替换***、**插入**和**合并**分别用*斜体红色*、**粗体蓝色**和无衬线黄色突出显示。(最好用彩色查看)。

为了考察攻击顺序的有效性，我们将 REPLACEONLY 与 BERTAttack 进行了比较。值得注意的是，REPLACEONLY的表现全面优于BERTAt- tack。这大概是因为BERTAttack在选择攻击位置时没有考虑到要填充的代币。

现在我们来谈谈构建候选令牌集时施加的两个约束。也许并不奇怪，消融文本相似性约束（*w/o* sim *> l）*降低了文本相似性性能，但增加了其他方面。消融掩蔽语言模型会产生更好的成功率，但迷惑性、语法性和文本相似性却差很多。

最后，我们比较了CLARE与不同屏蔽语言模型实现的情况。表[6](#_bookmark22)总结了这些结果。总的来说，精炼的RoBERTa在不损失性能的情况下实现了最快的速度。由于受害者模型是基于BERT的，我们推测使用模型自身的信息来攻击模型是不太经济的。

在本节中，我们将AG新闻数据集中的对抗性at- tacks按语篇（POS）标签进行了分析，发现大部分对抗性at- tacks发生在名词或名词性短语中。我们发现大部分的对抗性标签发生在名词或名词短语上。据推测，在许多主题分类数据集中，前期的词典严重依赖于一些特征性的名词词/短语。如表[7](#_bookmark25)所示，64%的*替换*动作应用于名词。*Insert* ac- tions倾向于将标记插入名词短语bigram中：最常见的两个POS bigram是名词短语。事实上，约有48%的*Insert*动作被应用到名词短语中。这也证明了我们只对名词性短语应用*Merge的*选择。

## 对抗性培训

本节探讨CLARE在改进下游模型的准确性和鲁棒性方面的潜力。继[Tsipras等人](#_bookmark69)（[2018](#_bookmark69)）之后，我们使用CLARE为AG新闻训练实例生成对抗性实例，并将其作为额外的训练数据。我们考虑了两种设置：(1)使用完整的训练数据和完整的对抗性数据进行训练；(2)10%随机采样的训练数据及其对抗性数据，以模拟低资源场景。对于这两种设置，我们比较了一个基于BERT的MLP分类器和一个没有任何预训练嵌入的TextCNN（[Kim](#_bookmark45)，[2014](#_bookmark45)）分类器。

*作为数据增强的对抗性例子，是否可以帮助实现更好的测试准确性？*作为

***替换插入合并***

**被害人Model 事故***个***A-率1Mod***个*

*NOUN：64%（NOUN，NOUN）：12%ADJ-NOUN：31%。*

*ADJ：17%（ADJ，NOUN）：10%NOUN-NOUN：22%。*

BERT (100% data) 95.0 79.1 6.1)

+ 100% adversarial -0.2 -18.0 +5.1

*VERB：7%（NOUN，VERB）：9%DT-NOUN：12%。*

**背景：**...... 阿米特-约兰，政府的*网络安全。*

负责人，昨天突然辞职，一年后......。

**替换**：网络安全 *{安全、监视、密码学、情报、加密...}。*

*仁*

**插入**：网络安全主管*{技术、国防、智能、计划、项目...}。*

*仁*

|  |  |  |  |
| --- | --- | --- | --- |
| TextCNN(100%数据) | 91.2 | 92.7 | 5.0 |
| ＋100%对抗性 | -0.5 | -10.3 | +0.8 |
| BERT（10%数据） | 92.5 | 96.1 | 5.4 |
| ＋10%对抗性 | +0.0 | -12.3 | +7.6 |

TextCNN（10% data) 83.6 99.0 5.6

+ 10% adversarial +2.0 -3.5 +0.3

**合并**：网络安全主管 *{主任、顾问、行政人员}。*

*仁*

*火箭手、科学家、秘书......}。*

表7：**顶部**：每个扰动类型的Top-3 POS标签（或POS标签大图）及其百分比。(*a，b*)：在*a*和*b*之间插入一个标记。*a-b*：将*a*和*b*合并成一个标记。**底部**：一个AG新闻样本，其中CLARE一个AG新闻样本，其中CLARE扰动令牌"*网络安全*"。TextFooler无法攻击这个token，因为它不在其词汇表中。

如表 [8](#_bookmark26) 所示，当有完整的训练数据时，对抗式训练分别略微*降低了* 0.2% 和 0.5% 的测试精度。这与之前的观察结果一致（[Jia et al.](#_bookmark41) , [2019](#_bookmark41)）。有趣的是，在对抗性训练的低数据场景下，基于BERT的分类器没有精度下降，而TextCNN实现了2.0%的绝对改进。这说明容量较小的模型可以从银色数据中获益更多。*对抗性训练是否有助于模型去抵御对抗性攻击？*为了评估这一点，我们使用CLARE来攻击用和不用对抗性例子训练的分类器。[9](#_bookmark27) 较高的成功率和较少的修改表明受害者分类器更容易受到对抗性攻击。如表[8](#_bookmark26)所示，在4个案例中的3个案例中，对抗性训练有助于降低攻击成功率超过10.3%，并使所需的修改次数增加超过0.8。唯一的例外是用10%数据训练的TextCNN模型。一个可能的原因是，它的训练

数据少，因此概括性较差。

这些结果表明，CLARE可以用来提高下游模型的鲁棒性，而精度下降可以忽略不计。

# 相关工作

**文本对抗性攻击。**越来越多的人致力于生成更好的文本对抗性例子，各种

9在初步实验中，我们发现使用其他模型来攻击用CLARE生成的对抗性例子训练的受害者模型，比使用CLARE本身更加困难。

表8：AG新闻测试集的对抗性训练结果。"Acc"表示准确性。

攻击模型。基于字符的模型（[Liang等人](#_bookmark50)，[2019](#_bookmark50)；[Ebrahimi等人](#_bookmark33)，[2018](#_bookmark33)；[Li等人](#_bookmark48)，[2018](#_bookmark48)；[Gao等人](#_bookmark34)，[2018](#_bookmark34)，*间*）使用拼写错误来攻击受害者系统；然而，这些攻击通常可以通过拼写检查器进行防御（[Pruthi等人](#_bookmark58)，[2019](#_bookmark58)；[Zhou等人](#_bookmark86)，[2019b](#_bookmark86)；[Jones等人](#_bookmark43)，[2020](#_bookmark43)）。许多句子级模型（[Iyyer等人](#_bookmark39) ，[2018](#_bookmark39)；[Wang等人](#_bookmark73) ，[2020](#_bookmark73)；[Zou等人](#_bookmark87) ，[2020](#_bookmark87)*等*）已经被开发出来，以引入更复杂的令牌/短语扰动。然而，这些，一般- ally有difﬁculty保持与原始输入的语义相似性（[张等人](#_bookmark81)，[2020a](#_bookmark81)）。最近的词级模型探索了同义词替换规则，以增强语义意义的保存（[Alzantot等人](#_bookmark28)，[2018](#_bookmark28)；[Jin等人](#_bookmark42)，[2020](#_bookmark42)；[Ren等人](#_bookmark62)，[2019](#_bookmark62)；[Zhang等人，2019；Zang等人](#_bookmark79)，[2020](#_bookmark79)*等*）。我们的工作不同的是，CLARE使用了三种语境化的扰动，产生了更多的ﬂuent和语法输出。

**用 BERT 生成文本。**在各种自然语言任务中，使用掩蔽语言模型的生成已经得到了广泛的研究，从词汇替换（[Wu等人](#_bookmark77) ，[2019a](#_bookmark77)；[Zhou等人](#_bookmark85) ，[2019a](#_bookmark85)；[Qiang等人](#_bookmark59) ，[2020](#_bookmark59)；[Wu等人。](#_bookmark78)，[2019b](#_bookmark78)，*等*）到非自动生成（[Gu等](#_bookmark38)，[2018](#_bookmark38)；[Lee等](#_bookmark47)，[2018](#_bookmark47)；[Ghazvininejad等](#_bookmark36)，[2019](#_bookmark36)；[Wang和Cho](#_bookmark71)，[2019](#_bookmark71)；[Ma等](#_bookmark52)，[2019](#_bookmark52)；[Sun等](#_bookmark68)，[2019](#_bookmark68)；[Ren等](#_bookmark63)，[2020](#_bookmark63)；[Zhang等](#_bookmark83)，[2020b](#_bookmark83)，*等*）。

# 结论

我们提出了CLARE，一个针对文本的语境化广告语例生成模型。它使用来自预先训练的屏蔽语言模型的上下文知识，可以生成自然的、流畅的和符合语法的广告语例。语境化的例子生成模式有 *Replace*、*Insert* 和 *Merge* 三种。

CLARE可以产生不同长度的输出，并实现比基线更高的攻击成功率和更少的编辑。人工评估显示，CLARE在文本相似性、广度和语法性方面具有明显优势。我们在 <https://github.com/cookielee77/CLARE>发布我们的代码和模型。

# 鸣谢

我们要感谢审稿人的建设性意见。我们感谢NVIDIA公司为本研究捐赠的GPU。我们也感谢吴同双、王国银和任树槐的讨论和反馈。

# 参考文献

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani Srivastava, and Kai-Wei Chang.2018.Generating natural language adversarial ex- amples.In *Proc.of EMNLP*.

Daniel Cer, Yinfei Yang, Sheng-yi Kong, Nan Hua, Nicole Limtiaco, Rhomni St John, Noah Constant, Mario Guajardo-Cespedes, Steve Yuan, Chris Tar, et al. 2018.Universal sentence encoder. *arXiv preprint arXiv:1803.11175*.

程勇、蒋璐和Wolfgang Macherey.2019.Robust neural machine translation with doubly ad- versarial inputs.In *Proc.of ACL*.

Jacob Devlin，Ming-Wei Chang，Kenton Lee，and Kristina Toutanova.2019.Bert: Pre-training of deep bidirectional transformers for language understand- ing.In *Proc.of NAACL*.

William B Dolan和Chris Brockett。2005.Automati- cally constructing a corpus of sentential paraphrases.In *Proceedings of the Third International Workshop on Paraphrasing*.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou.2018.Hotﬂip:White-box adversarial exam- ples for text classiﬁcation.In *Proc.of ACL*.Anyi Rao, Daniel Lowd, and Dejing Dou.

Ji Gao, Jack Lanchantin, Mary Lou Soffa, and Yan-jun Qi.2018.Black-box generation of adversarial text sequences to evade deep learning classiﬁers.In *IEEE Security and Privacy Workshops（SPW）*.

Siddhant Garg和Goutham Ramakrishnan。2020.Bae。Bert-based adversarial examples for text clas- siﬁcation.In *Proc.of EMNLP*.

Marjan Ghazvininejad, Omer Levy, Yinhan Liu, and Luke Zettlemoyer.2019.Mask-predict:Parallel de- coding of conditional masked language models.In *Proc.of EMNLP*.

Ian J Goodfellow，Jonathon Shlens，和Christian Szegedy。2015.Explaining and harnessing adversar- ial examples.In *Proc.of ICLR*.

Jiatao Gu，James Bradbury，Caiming Xiong，Victor OK Li，and Richard Socher.2018.Non-autoregressive neural machine translation.In *Proc.of ICLR*.

Mohit Iyyer、John Wieting、Kevin Gimpel和Luke Zettlemoyer。2018.Adversarial example generation with syntactically controlled paraphrase networks.In *Proc.of NAACL*.

贾彦宏和梁珀西。2017.Adversarial exam- ples for evaluating reading comprehension systems.In *Proc.of EMNLP*.

Robin Jia, Aditi Raghunathan, Kerem Göksel, and Percy Liang.2019.Certiﬁed robustness to adversar- ial word substitutions.In *Proc.of EMNLP*.

金迪、金志敬、周天一和Peter Szolovits。2020.Is bert really robust? natural lan- guage attack on text classiﬁcation and entailment.In *Proc.of AAAI*.

Erik Jones、Robin Jia、Aditi Raghunathan和Percy Liang.2020.Robust encodings:Robust encodings: A framework for combating adversarial typos.In *Proc. of ACL*.

Katharina Kann, Sascha Rothe, and Katja Filippova.2018.Sentence-level ﬂuency evaluation:Refer- ences help, but can be spared!In *Proc. of CNLP*, pages 313-323.

Yoon Kim.2014.Convolutional neural networks for sentence classiﬁcation.In *Proc. of EMNLP*.

Keita Kurita, Paul Michel, and Graham Neubig.2020.Weight poisoning attacks on pre-trained models.*arXiv preprint arXiv:2004.06660*.

Jason Lee, Elman Mansimov, and Kyunghyun Cho.2018.Deterministic nonautoregressive neural se- quence modeling by iterative reﬁnement.In *Proc.of EMNLP*.

李金凤，纪守玲，杜天宇，李博，王婷.2018.Textbugger:Generating adversarial text against real-world applications. *ArXiv preprint arXiv:1812.05271*.

李林阳，马若天，郭启鹏，薛向阳，邱西鹏。2020.Bert-attack．Adversarial at - tack against bert using bert.In *Proc.of EMNLP*.

梁斌，李洪成，苏苗强，边潘，李锡荣，史文昌。2019.Deep text clas- siﬁcation can be fooled.In *Proc.of IJCAI*.

Yinhan Liu、Myle Ott、Naman Goyal、Jingfei Du、Man- dar Joshi、Danqi Chen、Omer Levy、Mike Lewis、Luke Zettlemoyer和Veselin Stoyanov。2019.罗伯塔。A robustly optimized bert pretraining ap- proach.*arXiv preprint arXiv:1907.11692*.

马学哲、周春婷、李宪、Graham Neu- big和Eduard Hovy.2019.Flowseq: Non-autoregressive conditional sequence generation with generative ﬂow.In *Proc.of EMNLP*.

John Morris, Eli Liﬂand, Jin Yong Yoo, Jake Grigsby, Di Jin, and Yanjun Qi.2020a.文本攻击。A frame-work for adversarial attacks, data augmentation, and adversarial training in NLP.In *Proc.of EMNLP: Sys- tem Demonstrations*.

John X Morris, Eli Liﬂand, Jack Lanchantin, Yangfeng Ji, and Yanjun Qi.2020b.Reevaluating adversar- ial examples in natural language.*arXiv preprint arXiv:2004.14174*.

Nikola Mrkšic´、Diarmuid Ó Séaghdha、Blaise Thom- son、Milica Gasic、Lina M Rojas Barahona、Pei-Hao Su、David Vandyke、Tsung-Hsien Wen和Steve Young。2016.Counter-ﬁtting word vectors to lin- guistic constraints.In *Proc.of NAACL*.

Daniel Naber等人，2003年。*A rule-based style and gram- mar checker*.Citeseer.

Adam Paszke、Sam Gross、Francisco Massa、Adam Lerer、James Bradbury、Gregory Chanan、Trevor Killeen、林泽明、Natalia Gimelshein、Luca Antiga等人2019年。Pytorch。An imperative style, performance deep learning library.In *Proc.of NeurIPS*.

Danish Pruthi, Bhuwan Dhingra, and Zachary C Lip- ton.2019.Combating adversarial misspellings with robust word recognition.In *Proc.of ACL*.

强继鹏，李云，朱毅，袁云浩。2020.A simple bert-based approach for lexical simpliﬁca- tion.In *Proc.of AAAI*.

Alec Radford、Jeffrey Wu、Rewon Child、David Luan、Dario Amodei和Ilya Sutskever。2019.语言模型是无监督的多任务学习者。*OpenAI博客*，1（8）：9。

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang.2016.Squad:100,000+问题的机器理解文本。In *Proc.of EMNLP*.

任书怀、邓义和、何坤、车万祥。2019.Generating natural language adversarial ex- amples through probability weighted word saliency.In *Proc.of ACL*.

任毅，刘景林，谭旭，赵晟，赵周，刘铁岩。2020.A study of nonautoregressive model for sequence generation.*arXiv preprint arXiv:2004.10454*.

Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin.2018.Semantically equivalent adversarial rules for debugging nlp models.In *Proc.of ACL*.

Suranjana Samanta和Sameep Mehta。2017.Towards crafting text adversarial samples.*arXiv preprint arXiv:1707.02812*.

Victor Sanh、Lysandre Debut、Julien Chaumond和Thomas Wolf。2019.Distilbert，bert的精馏版：更小、更快、更便宜、更轻。In *The 5th Workshop on Energy Efﬁcient Machine Learning and Cognitive Computing, NeurIPS*.

Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts.2013.Recursive deep mod- els for semantic compositionality over a sentiment treebank.In *Proc. of EMNLP*.

孙志清，李卓涵，王浩庆，何迪，林子，邓志宏.2019.Fast structured decoding for sequence models.In *Proc.of NeurIPS*.

Dimitris Tsipras，Shibani Santurkar，Logan Engstrom，Alexander Turner和Aleksander Madry。2018.Ro- bustness may be at odds with accuracy.In *Proc.of ICLR*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.2019.Universal adversarial trig- gers for attacking and analyzing nlp.In *Proc.of EMNLP*.

Alex Wang和Kyunghyun Cho。2019.Bert有一张嘴，它必须说话。Bert as a markov ran- dom ﬁeld language model.In Proceedings for Methods for Optimizing and Evaluat- ing Neural Language Generation.In *Proceedings of the Workshop on Methods for Optimizing and Evaluat- ing Neural Language Generation*.

Alex Wang、Amanpreet Singh、Julian Michael、Fe- lix Hill、Omer Levy和Samuel Bowman。2019a.Glue:A multi-task benchmark and analysis plat- form for natural language understanding.In *Proc.of ICLR*.

王博新，裴恒志，潘博远，陈倩，王硕航，李博。2020.T3:树型自动编码器正则化对抗性文本生成用于定向攻击。In *Proc.of EMNLP*.

王晓森，金浩，何坤.2019b.自然语言对抗式攻击和防御在词级中的应用.*arXiv preprint arXiv:1909.06723*.

Adina Williams、Nikita Nangia和Samuel Bowman。2018.A broad-coverage challenge corpus for sen- tence understanding through inference.In *Proc.of NAACL*.

托马斯-沃尔夫、莱桑德-德布特、维克多-桑赫、朱利安-肖蒙、克莱门特-德朗格、安东尼-莫伊、皮埃尔-里克-西斯塔克、蒂姆-罗尔特、R'emi-卢夫、摩根-芬托-伊兹和杰米-布鲁。2019.Huggingface 的 trans- formers。最先进的自然语言处理- ing.*ArXiv*, abs/1910.03771.

吴星，吕尚文，臧良军，韩继忠，胡松林.2019a.Conditional bert contextual augmentation.In *Proc.of ICCS*.

吴星，张涛，臧良军，韩继忠，胡松林.2019b."mask and inﬁll":Apply- ing masked language model to sentiment transfer.In *Proc.of IJCAI*.

袁藏，齐凡超，杨承浩，刘志远，张萌，刘群，孙茂松。2020.Word-level textual adversarial attacking as combina- torial optimization.In *Proc.of ACL*.

张黄钊，周浩，苗宁，李磊.2019.Generating ﬂuent adversarial examples for natural languages.In *Proc.of ACL*.

Wei Emma Zhang, Quan Z Sheng, Ahoud Alhazmi, and Chenliang Li.2020a.Adversarial attacks on deep-learning models in natural language process- ing:A survey.*ACM Transactions on Intelligent Sys- tems and Technology（TIST）*，11（3）：1-41.

张翔，赵俊波，晏乐村。2015.Character-level convolutional networks for text clas- siﬁcation.In *Proc.of NeurIPS*.

张义哲，王国银，李春元，甘哲，Chris Brockett和Bill Dolan.2020b.Pointer:通过基于插入的生成式预训练进行约束文本生成。 *arXiv preprint arXiv:2005.00558*.

Zhengli Zhao，Dheeru Dua，and Sameer Singh.2018.Generating natural adversarial examples.In *Proc.of ICLR*.

周望春树，葛涛，徐可，魏芙蓉，周明.2019a.Bert-based lexical substitution.In *Proc.of ACL*.

周毅超，蒋洁瑜，张凯伟，王伟.2019b.Learning to discriminate perturba- tions for blocking adversarial attacks in text classi-ﬁcation.In *Proc.of EMNLP*.

邹伟，黄树建，谢军，戴新宇，陈家军。2020.A reinforced generation of adversar- ial samples for neural machine translation.In *Proc.of ACL*.

# 附录

## 额外的实验细节

**模型实现。**所有基于RoBERTa和BERTbase的预训练模型- els和受害者模型都是在Py- Torch（[Paszke等人](#_bookmark57)，[2019）](#_bookmark57)的基础上用Hugging Face变换器[10](#_bookmark89)（[Wolf等人](#_bookmark76)，[2019）](#_bookmark76)实现的。RoBERTadistill，RoBERTabase和uncase BERTbase模型分别有82M，125M和110M参数。我们使用RoBERTadistill作为我们的主要骨干，以达到快速推理的目的。[TextFooler11](#_bookmark90)和BERTAt- [tack12的](#_bookmark91)构建是由作者提供的其开源实现。在TextFooler+LM的实现中，我们使用小尺寸的GPT-2语言模型（[Radford et al.](#_bookmark60)在 adver- sarial 训练中（§[4.3](#_bookmark24)），小 TextCNN 受害者模型（[Kim](#_bookmark45)，[2014](#_bookmark45)）有 128 个嵌入大小和 100 个ﬁlters，3*、*4*、*5 窗口大小，0*.*5 辍学，结果有 7M 个参数。

在实施*w/o* pMLM *> k的*消融研究（§4.1[）](#_bookmark18)期间[，](#_bookmark18)我们随机抽取200个tokens，然后应用相似性约束来构建候选集，因为穷尽词汇库- lary是计算上昂贵的。

**评估度量。**相似性函数模拟建立在通用句子编码器（USE；[Cer etal.](#_bookmark29) ，[2018](#_bookmark29)）的基础上，以测量原始输入与其对手之间在窗口大小为15的per- turbation位置的*局部*相似性。*所有基线*在构建候选词汇时都配备了这个 Sim。评估指标**Sim**使用USE来计算两个文本之间的*全局*相似度。这些程序通常遵循[Jin等人](#_bookmark42)（[2020](#_bookmark42)）。我们主要依靠人工评估（§[3.3](#_bookmark14)）得出结论，与TextFooler相比，CLARE上保留- ing文本相似性的优势明显。

**数据处理。**在处理数据时，我们保留了受害者模型训练和攻击时文本中的所有标点符号。这与TextFooler([Jin et al.](#_bookmark42) , [2020](#_bookmark42))中的预处理设置不同，因为我们实证发现，去除标点符号会使受害者模型变得脆弱。由于GLUE

10 [https://github.com/huggingface/。](https://github.com/huggingface/transformers)

[变压器](https://github.com/huggingface/transformers)

[11 https://github.com/jind11/TextFooler](https://github.com/jind11/TextFooler)12[https://github.com/LinyangLee/。](https://github.com/LinyangLee/BERT-Attack)

[BERT-攻击](https://github.com/LinyangLee/BERT-Attack)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **数据集** | **平均长度长度** | **# ＃类** | **火车** | **测试** | **Acc** |
| SST-2 | 10 | 2 | 67K | 0.9K | 92.3% |
| DBpedia | 55 | 14 | 560K | 70K | 99.3% |
| QQP | 13/13 | 2 | 363K | 40K | 91.4% |
| MRPC | 23/23 | 2 | 3.6K | 1.7K | 81.4% |

表9：数据集的一些统计数据。最后一列表示受害者模型在*没有*对抗性攻击*的情况下*对原始测试集的准确性。

基准([Wanget al.](#_bookmark72) , [2019a](#_bookmark72))没有提供测试集的标签，我们反而使用其dev集作为评估中包含的数据集(MNLI, QNLI, QQP, MRPC, SST-2)的测试集。对于句子对任务（如MNLI、QNLI、QQP、MRPC），我们攻击较长的一个排除两个句子中出现的to- kens。这是因为推理任务通常要求包含的数据具有相同的关键词，例如数字、名称实体等。所有实验都是在一个Nvidia GTX 1080Ti GPU上进行的。

## 其他结果

我们在本节中包括DBpedia本体数据集（**DBpedia**；[Zhang et al.](#_bookmark82) , [2015](#_bookmark82)，Stanford sentiment treebank (**SST-2**; [Socher et al.](#_bookmark67) , [2013](#_bookmark67))，Microsoft Research Paraphrase Corpus (**MRPC**; [Dolan and Brockett](#_bookmark32), [2005](#_bookmark32))，以及来自GLUE基准的Quora Question Pairs (**QQP**)的结果。表[9](#_bookmark92)总结了这些数据集的来统计。表[10](#_bookmark95)总结了这些数据集上不同模型的结果。与所有的基线相比，CLARE在攻击成功率、迷惑性、语法性和相似性上取得了最好的表现。这与我们在

§[3.3](#_bookmark14).

## 人类评估详情

对于**AG新闻**数据集上的每一个人类评估，我们从测试集中随机抽取300个句子，结合CLARE和TextFooler的相应对抗性考试-ples（我们只考虑句子可以被两个模型攻击）。为了使任务不那么抽象，我们将两种模型的对抗性例子配对，并将它们与原始输入及其金标一起呈现给参与者。我们问他们在以下方面更喜欢哪一个：(1)与原始输入有更相似的意义(相似性)，(2)更有韵律和语法性(韵律性和语法性)。我们还为他们提供了一个neu- tral选项，当参与者考虑到这两个

SST-2 (PPL = 99.5) DBpedia (PPL = 37.3))

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **型号** | **A-***率个* | *Mod1* | *PPL1GErr1* | | **Sim***个* | **A-***率个* | *Mod1PPL1* | *GErr1* | **Sim***个* |
| TextFooler | 89.8 | 14.9 | 227.7 0.53 | | 0.69 | 56.2 | 24.9 182.5 | 1.88 | 0.68 |
| LM + 51.7 | | 18.3 | 137.5 | 0.50 | 0.69 | 20.1 | 22.4 84.0 | 1.22 | 0.70 |
| BERTAttack87.8 | | 8.1 | 142.9 | 0.03 | 0.67 | 60.7 | 9.1 57.8 | 0.20 | 0.69 |
| **CLARE97.8** | | **7.5** | **137.4** | **0.01** | **0.75** | **65.8** | **7.0 53.3** | **-0.03** | **0.73** |
| QQP（PPL=56.2) MRPC（PPL=42.9）。 | | | | | | | | | |
| **型号** | **A-***率个* | *Mod1PPL1GErr1* | | | **Sim***个* | **A-***率个* | *Mod1PPL1GErr1* | | **Sim***个* |
| TextFooler | 16.2 | 12.7 145.2 0.61 | | | 0.74 | 24.5 | 10.6 118.8 0.35 | | 0.75 |
| LM + 7.8 | | 12.9 | 78.8 | 0.21 | 0.77 | 12.9 | 9.5 71.0 | 0.29 | 0.79 |
| BERTAttack24.2 | | 11.3 | 78.0 | 0.25 | 0.71 | 29.7 | 13.5 74.6 | 0.05 | 0.79 |
| **CLARE27.7** | | **10.2** | **74.8** | **0.14** | **0.76** | **34.8** | **9.1 69.5** | **0.02** | **0.83** |

表10：对抗式示例生成在攻击成功率（A-rate）、修改率（Mod）、迷惑性（PPL）、增加的语法错误数（GErr）和文本相似性（Sim）方面的表现。每个数据集的原始输入的迷惑性在括号中表示。粗体表示每个指标上的最佳表现。

无法区分。此外，我们要求part- ipants对对抗性的例子进行注释，并将他们的注释与黄金标签（标签一致性）进行比较。标签一致性越高，说明该模型在保留人类预言的同时，能更好地导致受害者模型出错。

每一对系统输出被随机预发到5个人群来源的法官，他们使用图[3](#_bookmark99)所示的形式表示他们的相似性，流利性和克成熟度的偏好。标签任务如图[4](#_bookmark100)所示。为了最大限度地减少垃圾信息的影响，我们采用了众包服务提供的排名靠前的30%的美国工人。我们还提供了详细的任务描述和例子来指导评委。我们通过使用10K配对的引导复制，使用R Boot统计包实现，基于95%的conﬁdence in- tervals计算*p*值。

## 定性样品

我们将CLARE和TextFooler在**AG News**、**DBpeida**、**Yelp**、**MNLI**和**QNLI**数据集上生成的对抗性例子列入表[11](#_bookmark97)和表[12](#_bookmark98)。

**AG**

(业务)

TextFooler (Sci&Tech)

CLARE

(科学与技术)

**AG**

(体育)

TextFooler(世界)

CLARE

(世界)

**Yelp**

(正)

TextFooler (Negative)

CLARE

(负数)

**Yelp**

(正)

TextFooler (Negative)

CLARE

(负数)

**MNLI**

(中性)

TextFooler

TECH BUZZ : 雅虎和Adobe联手推出新的网络服务。雅虎公司和Adobe系统公司已联手开发对方的客户，并将网络搜索功能放入Adobe流行的Acrobat Reader软件中，以加强在线搜索和服务的战斗。

TECH BUZZ : 雅虎和Adobe合作开发*浪漫的网络空间实用工具*。雅虎公司和Adobe系统公司已联手开发对方的客户，并将网络搜索功能放入Adobe流行的Acrobat Reader软件中，以加强在线*定位*和服务的战斗。

TECH BUZZ : 雅虎和Adobe联手推出新的Web *Explorer*。雅虎公司和Adobe系统公司已联手开发对方的客户，并将网络搜索功能放入Adobe流行的Acrobat Reader软件中，以加强在线搜索和服务的战斗。

Padres Blank Dodgers 3 - 0.洛杉矶--亚当-伊顿在七局中允许五支安打，为他的职业生涯--高的第10次胜利，布赖恩-吉尔斯连续第二场比赛的本垒打，和圣地亚哥教士击败洛杉矶道奇队3 - 0周四晚上。国联西区-领先道奇队的领先优势被削减到2 1/2场比赛，超过旧金山-他们自7月31日以来最小的......

*道奇*队以3-0击败*洋基队。*在这里，我想说的是，我是一个很好的人，我是一个很好的人，我是一个很好的人，我是一个很好的人，我是一个很好*的人，我是一个很好的人*，我是一个很好的人，我是一个*很好的人，我是一个很好的人，我是一个很*好的人，我是*一个很好的*人，我是一个很好的人，我是一个很好的人。国联*西部人*-*杰出的道奇*"的领先优势被削减到2 1 / 2游戏超过*圣*-他们的*微小作为janvier* 31 ... ...

教士队3 - 0击败道奇队。密尔沃基**下一个**--亚当-伊顿（Adam Eaton）在七局中允许五支安打，取得了他职业生涯--最高的第10场胜利，布莱恩-吉尔斯（Brian Giles）连续第二场比赛打出本垒打，圣地亚哥教士队周四晚上以3 - 0击败洛杉矶道奇队。国联西区-领先道奇队的领先优势被削减到2 1/2场比赛，超过旧金山-他们自7月31日以来最小的......

这家连锁店的食物一直都很好。我们在市中心（我们过年的地方）的服务员是新来的，但这完全没有影响我们的服务。她对我们的需求很及时，很细心。

这家连锁店的食物一直都是*必然的ok*。我们在市中心（我们过年的地方）的服务员是新来的，但这完全没有影响我们的服务。她*很早*，而且很细心地满足我们的需求。

这家连锁店的食物一直都是**看着**一贯的好。我们在市中心（我们过年的地方）的服务员是新来的，但这完全没有*提升*我们的服务。她对我们的需求很及时，很细心。

梭子鱼汤其实很有味道，不只是牛肉和面条的热水味道。我一般都是做外卖的，吃饭的时候点的菜很快就出来了，这应该是意料之中的pho，泡面、切牛肉、倒汤都不难。

辣酱汤其实是很美味的，并不像*屠宰*和*沙拉*的*水*一样，只是*美味*。我*多次*做外卖，晚餐时点*的*菜很快就出来了，这应该是预料到的，泡面、*分割*牛肉和倒汤都不*费劲*。

辣酱汤其实很有味道，并不是像热水加牛肉和面条那样**平淡无奇**。我一般都是做外卖，吃饭的时候点的菜出得**特别**快，这应该是意料之中的pho，泡面、切牛肉、倒汤都不难。

*前提*：底比斯一直保持着权力，直到第12王朝，当它的第五任国王，Amenemhet Iwho在1980年至1951年b.c.之间建立了一个首都孟菲斯附近。

*假设*。孟菲斯附近的首都只维持了半个世纪，然后其居民就放弃了它，前往下一个首都。

*前提*：底比斯*apprehended倒权力*，直到第12个*家庭*，当它的*最早的*国王，Amenemhet岩谁在1980年1951年*c*.c.之间建立了一个首都孟菲斯附近。

矛盾）*假说*。孟菲斯附近的首都只维持了半个世纪，然后居民们就放弃了它，去了下一个首都。

*预设*：底比斯一直保持着权力，直到第12王朝，当它的第一任国王Amenemhet Iho在位的时候，他的权力在12王朝之间。

CLARE 1980年 公元前1951年**此后**在孟菲斯附近建都。

矛盾）*假说*。孟菲斯附近的首都只维持了半个世纪，然后居民们就放弃了它，去了下一个首都。

**MNLI**

(实体)

TextFooler (Neutral)

CLARE

(中性)

*前提*：希望华尔街在被迫行动之前，能自愿采取措施解决这些问题。

*假设*：华尔街面临的问题，需要解决。

*前提*：希望华尔街在被迫行动之前，能自愿采取措施*处理*这些问题。

*假设*：华尔街面临的问题，需要解决。

*前提*：希望华尔街在被迫行动之前，能自愿采取措施*消除*这些问题。

*假设*：华尔街面临的问题，需要解决。

表11：不同模型产生的对抗性例子。原始数据集的金色标签显示在（加粗）数据集名称下方。***替换***、**插入**和**合并**分别用*斜体红色*、**粗体蓝色**和无衬线黄色突出显示。(最好用彩色查看)。

**QNLI**

(实体)

TextFooler (Not- Entailment)

CLARE

(非实体)

**QNLI**

(实体)

TextFooler (Not- Entailment)

CLARE

(非实体)

**DBpedia**

*前提*：是谁推翻了塔夫特谷的判决？

*假设*。新的自由党政府的第一件事就是推翻塔夫谷案的判决。

*前提*：是谁推翻了塔夫特谷的判决？

*假设*：新的自由党政府的第一件事就是*投资*塔夫谷的判决。

*前提*：是谁推翻了塔夫特谷的判决？

*假设*：新自由党*宪法*的第一项法案就是推翻塔夫谷案的判决。

*前提*：软件测试人员要注意什么 ?

*假设*。黑盒测试把软件当作一个黑盒，在不了解内部实现的情况下检查功能，不看源代码。

*前提*：软件测试人员要注意什么 ?

*假设*。*黑盒*测试管理软件作为一个黑盒，*调查功能，除非*任何知识的内部*fulﬁl*，*除非*看到*井喷*代码。

*前提*：软件测试人员要注意什么 ?

*假设*。黑盒测试把软件当作一个黑盒，在不*了解*内部实现的情况下检查功能，不看源代码。

本田Crossroad。本田Crossroad指的是本田生产的两种规格类型的SUV。其中一个是

(运输)重塑路虎发现系列I SUV，而另一款是2008年推出的完全不同的车辆。

文字傻瓜 (专辑)

CLARE

(公司)

**DBpedia**

(公司)

TextFooler(建筑)

CLARE

(建筑物)

**MRPC**

(不是译文)

TextFooler (Paraphrase)

CLARE

(译注)

**MRPC**

(译注)

TextFooler(不是转述)

CLARE

(不是译文)

*铃木交界处*。*铃木*十字路口指的是由*五十铃摆出的三款准确打字*的*prius*。其中一款是重新改装的路虎*识别*系列I *LEXUS*，而另一款则是2008年推出的完全不同的车型。

本田Crossroad。本田Crossroad指的是本田公司生产的两款规格的SUV*厂家*。其中一个是重装的路虎发现系列I SUV，而另一个是2008年推出的完全不同的车辆。

Yellow Rat Bastard。Yellow Rat Bastard是Henry Ishay拥有的纽约市服装零售连锁店中的Flagship机构。它专门经营嘻哈和另类风格的服装和鞋子。

*Yellowish Rats Schmuck* 。*Yellowish Rats Dickwad*是纽约市一家*连锁*零售店的*旗舰店。*

由*Henrik* Ishay拥有的制服店。它*专门从事*嘻哈和另类风格的*洗衣*和*运动鞋*。

黄鼠霸王。Yellow Rat Bastard **Mall**是Henry Ishay拥有的纽约市服装零售连锁店中的Flagship机构。它专卖嘻哈和另类风格的服装和鞋子。

*前提*：2003年美洲市场将下降2.1%，达到306亿美元，然后增长15.7%，达到。

2004年为354亿美元。

*假设*。预计今年美国芯片市场将下降2.1%，2004年将增长15.7%。

*前提*：2003年美洲市场将下降2.1%，达到306亿美元，然后增长15.7%，达到。

2004年为354亿美元。

*假设*。美国芯片市场今年将下降2.1%，2004年将增长15.7%。

*前提*：2003年美洲市场将下降2.1%，达到306亿美元，然后增长15.7%，达到。

2004年为354亿美元。

*假设*：美国芯片市场今年将下降2.1%，2004年将增长15.7%。美国芯片市场预计今年将下降2.1%，2004**年**将增长15.7%。

*前提*：证券交易委员会在波士顿对该少年提起民事欺诈诉讼。

*假设*。证券交易委员会在周四提起了相关的民事案件。

*前提*：证券交易委员会在波士顿对该少年提起民事欺诈诉讼。

*假设*。证券交易委员会在*昨天*提起了*一起连接*民事案件。

*前提*：证券交易委员会在波士顿对该少年提起民事欺诈诉讼。

*假设*。证券交易委员会周四提起了*马萨诸塞州的民事诉讼*。

表12：不同模型产生的对抗性例子。原始数据集的金色标签显示在（粗体）数据集名称下方。***替换***、**插入**和**合并**分别用*斜体红色*、**粗体蓝色**和无衬线黄色突出显示。(最好用彩色查看)。



图3：从文本相似性和词性及语法性方面对人类评价进行配对比较。

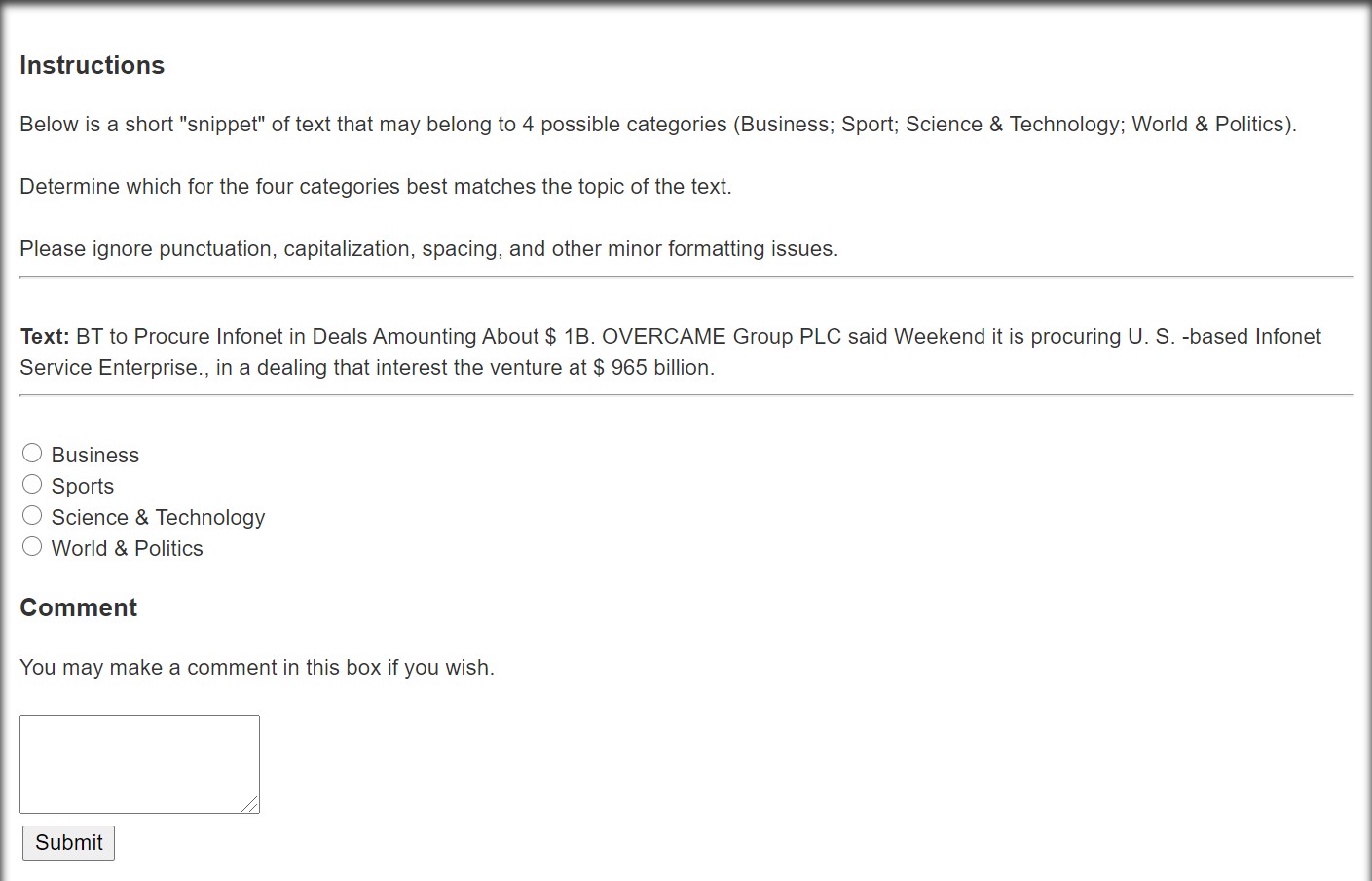


图4：标签一致性任务对人的评价。