T3：树形自动编码器规整化对抗式文本生成的目标攻击。

**王博新**1**，裴恒志**1**，潘博远**2**，陈倩**3**，王硕航**4**，李博**1 1伊利诺伊大学香槟分校2浙江大学。

3腾讯4微软Dynamics 365人工智能研究boxinw2，hpei4，lbo @illinois.edu，[panby@zju.edu.cn，](mailto:panby@zju.edu.cn)[qianchen@tencent.com，](mailto:qianchen@tencent.com)[shuohang.wang@microsoft.com。](mailto:shuohang.wang@microsoft.com)

*{}*

# 摘要

针对自然语言处理系统的对抗性攻击，对输入进行看似无害的修改，会对目标模型造成任意错误。虽然引起了极大的关注，但这种对抗性攻击可以被用来估计NLP模型的破坏性。与连续数据域(*例如*，图像)中的对抗性示例生成相比，生成保留原始意义的*对抗性文本*是具有挑战性的，因为文本空间是离散的和不可区分的。为了处理这些挑战，我们提出了一个*目标可控的*对抗性攻击框架T3，它适用于一系列的NLP任务。特别是，我们提出了一个基于树的自动编码器，将离散的文本数据嵌入到一个连续的表示空间中，在此基础上，我们优化了ad- versarial扰动。基于树的自动编码器被应用于规范化生成文本的语法正确性，并在句子(T3(SENT))或词(T3(WORD))级别上对其进行处理。我们考虑了两个最具代表性的NLP任务：情感分析（sentiment anal- ysis）和问题回答（QA）。广泛的实验结果和人类研究表明，T3生成的对抗性文本可以成功地操纵NLP模型，在不误导人类的情况下输出*目标*错误答案。此外，我们还表明，生成的对抗性文本具有较高的转移能力，这使得黑盒攻击在实践中得以实现。我们的工作揭示了一种有效且通用的方式来检验NLP模型的鲁棒性。我们的代码公开在<https://github.com/AI-secure/T3/>。

# 介紹

最近的研究表明，深度神经网络（DNNs）很容易受到精心设计的对抗性例子的影响（[Goodfellow等](#_bookmark22)，[2015](#_bookmark22)；[Papernot等](#_bookmark50)，[2016](#_bookmark50)；[Eykholt等](#_bookmark21)，[2017](#_bookmark21)。

**问题：1989年是谁结束了这个系列？**谁在1989年结束了这个系列？

**段。**英国广播公司戏剧部的连续剧部门制作了26季节目，在英国广播公司1台播出。收视率下降，公众对该剧的认知度下降，以及不那么受欢迎的传输时段，1989年BBC 1的控制人Jonathan Powell暂停了该剧的制作。

该系列将返回。*唐纳德-特朗普在1988年结束了一项计划。*

**QA预测。**Jonathan Powell *\_* Donald Trump

**Yelp评论。***我一直期待着看到鸡和鸡-ens走来走去*。如果你觉得拉斯维加斯的白色垃圾越来越多，不要靠近这里。这个地方就像一个斯坦贝克的小说来生活。我一直以为会有驴子和鸡走来走去。呜--猪--苏--eeee这个地方太可怕了！！！！！。

**情绪预测。**最负面*\_*最正面

表1：T3为QA模型和情感分类器生成的两个对抗性例子。在原始段落中加入*对抗性的情感*，可以将正确的预测引向对抗者确定的目标错误答案。

[Moosavi-Dezfooli 等人](#_bookmark41) ，[2016](#_bookmark41)）。)这些例子有助于探索神经网络的脆弱性和可解释性。与非目标攻击相比，*目标可控*攻击（或称定向攻击）更具危险性和挑战性，因为它们会误导系统（例如，自动驾驶汽车）采取目标行动，这对基于DNN的应用程序的鲁棒性提出了安全担忧。虽然在连续数据域（包括图像、音频和视频）中提出了很多成功的攻击方法，但如何在离散文本域中有效地生成对抗性例子仍然是一个具有挑战性的问题。

与计算机视觉中的对抗性攻击不同，计算机视觉中的对抗性攻击是在输入图像中添加难以察觉的噪声，编辑哪怕是原段落中的一个字，也可能会极大地改变其含义，也会愚弄人类。所以在本文中，我们专注于生成一个对抗性句子，并将其添加到输入段落中。生成对抗性文本有几个挑战。1）很难衡量

6134

*Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6134-6150, November 16-20, 2020.Oc 2020计算语言学协会。

与原始文本相比，对抗式文本的有效性和自然性；2）基于梯度的对抗式攻击方法不能直接应用于离散结构化数据；3）与原始句子的原位对抗式修改相比，对抗式句子生成更具挑战性，因为生成器需要同时考虑句子语义和句法连贯性。到目前为止，现有的文本对抗式攻击要么无力地利用启发式解决方案，如遗传算法（[Jin等](#_bookmark31)，[2019](#_bookmark31)）来搜索词级替换，要么仅限于攻击speciﬁc NLP任务（[Jia和Liang](#_bookmark29)，[2017](#_bookmark29)；[Lei等](#_bookmark33)，[2018](#_bookmark33)）。

此外，有效的*目标可控*攻击，可以控制模型输出预期的错误答案，已被证明是NLP模型的difﬁcult。[Wallace等人](#_bookmark47)（[2019](#_bookmark47)）创建了通用触发器，以诱导QA模型输出目标an- swers，但目标攻击成功率很低。其他工作（[Cheng等](#_bookmark16)，[2018](#_bookmark16)；[Jin等](#_bookmark31)，[2019](#_bookmark31)；[Zhang等](#_bookmark54)，[2019](#_bookmark54)）对原始段落进行词级的原位修改，以实现定向攻击，这可能会改变原始输入的含义。因此，如何生成既不改变原始输入含义，又能实现高定向攻击成功率的对抗性句子，似乎是一个有趣且具有挑战性的问题。

在本文中，我们通过提出一个对抗性评估框架T3来解决这些挑战，以生成针对一般NLP任务的对抗性文本，并评估当前NLP模型的鲁棒性。具体来说，T3的核心组件是一个新型的基于树的自动编码器，它在一个大型语料库上预先训练，以捕获和维护语义和句法结构。树型编码器将离散的文本转换为连续的语义信息，解决了离散输入的难题。在此基础上，我们可以利用基于优化的方法在连续嵌入空间上搜索对抗性的扰动，这比遗传算法等启发式方法更有效，因为遗传算法的搜索空间按指数增长。基于树层次结构的不同层次，可以在叶层次和根层次上添加对抗性扰动，以施加词层次(T3(WORD))或句层次(T3(SENT))的扰动。最后，基于树形的去编码器将通过一组树形语法规则将对抗性嵌入映射回对抗性文本。

其中保留了原始输入的语义内容和同义结构。为了保证攻击的成功率，可以采用迭代过程。

综上所述，我们的主要贡献在于。(1)与以往的文本对抗式攻击研究不同，我们通过协整式对抗式文本生成实现了针对性攻击，能够操纵模型输出针对性的错误答案。

(2)我们提出了一种新型的基于树的文本自动编码器，在保留语义的同时，规整了对抗性文本的语法结构。它还解决了攻击离散文本的挑战，将句子嵌入到连续的潜伏空间中，在此基础上应用基于优化的ad- versarial perturbation来指导 adversarial句子的生成；（3）我们进行了大量的实验，成功实现了对不同情感分类器和QA模型的定向攻击，攻击成功率和可转移性高于最先进的基线方法。人体研究表明，T3生成的adversar- ial文本对攻击神经模型是有效的，同时几乎不影响人类的判断。

# 相关工作

大量关于*对抗性例子*的工作都集中在对连续输入空间的扰动上，虽然在生成离散空间的对抗性扰动方面取得了一些进展，但仍有一些挑战没有解决。尽管在离散空间中生成- ing对抗性扰动方面已经取得了一些进展，但仍有一些挑战没有解决。例如，([Zhaoetal.](#_bookmark56) , [2017](#_bookmark56))利用生成式对抗性网络(GAN)来生成自然对抗性文本。然而，这种方法不能明确地控制生成实例的质量。大多数现有的方法（[Ren等](#_bookmark42)，[2019](#_bookmark42)；[Zhang等](#_bookmark55)，[2019](#_bookmark55)；[Jia](#_bookmark31)[和Liang](#_bookmark29)，[2017](#_bookmark29)；[Li等](#_bookmark34)，[2018](#_bookmark34)；[Jin等](#_bookmark31)，[2019](#_bookmark31)）应用启发式策略来合成advers- sarial文本。1）首先识别影响预测的特征（如char- acters，单词和句子），2）遵循不同的搜索策略，用构建的pertur- bation候选人（如错别字，同义词，反义词，频繁的单词）扰乱这些特征。例如，([Liang et al.](#_bookmark36) , [2017](#_bookmark36))采用损失梯度*L*来选择重要的字符和短语进行扰动，而([Samanta andMehta](#_bookmark43), [2017](#_bookmark43))则将错字、同义词和impor- tant副词/形容词作为插入和替换的候选词。一旦获得了infﬂuential特征，应用扰动的策略是

*V*

一般包括*插入*、*删除*和*替换*。这种文本对抗式攻击方法无法保证生成的文本的语法正确性。例如，([Liang et al.](#_bookmark36) , [2017](#_bookmark36))生成的文本几乎是随机的char- acters流。为了生成语法正确的perturba- tion，[Jia和Liang](#_bookmark29)采用了另一种启发式策略，即在段落中添加*人工*构建的合法分心句子来引入假信息。这些启发式方法总体上不具有可扩展性，不能实现定向攻击，即敌方文本可以引导到选定的敌方目标（如分类中的敌方标签）。最近的工作开始使用梯度([Michel等人](#_bookmark39)，[2019](#_bookmark39)；[Ebrahimi等人](#_bookmark19)，[2017](#_bookmark19))来指导寻找适用于任意句子的通用触发器([Wallace等人](#_bookmark47)，[2019](#_bookmark47))来愚弄学习者，尽管所报道的攻击成功率相当低，或者当ap- plied到其他NLP任务时，它们受到了无效率的影响。相比之下，我们提出的T3框架能够有效地生成协同战术正确的对抗性文本，在多个任务上的不同模式下实现高目标攻击成功率。

# 架構

## 前言

在深入探讨之前，我们先重温一下T3框架所支持的at- tack场景和攻击能力。

**攻击方案。**与之前的对抗性文本生成作品（[Lei等](#_bookmark33)，[2018](#_bookmark33)；[Cheng等](#_bookmark16)，[2018](#_bookmark16)；[Papernot 等](#_bookmark49)，[2016](#_bookmark49)；[Miyato 等](#_bookmark40)，[2016](#_bookmark40)；[Alzantot 等](#_bookmark12)，[2018](#_bookmark12)）直接原地修改关键词，可能会有改变语义或编辑地道答案的风险不同，我们生成的是*连词对抗*（[Jia 和 Liang](#_bookmark29)，[2017](#_bookmark29)）（*abbr.*，concat attack）。Con- cat attack不改变原段落或问题中的任何单词，而是在原段落中附加一个新的对抗句来欺骗模型。一个有效的对抗性句子需要确保附加的文字与原段*相一致*，换句话说就是它不应该与段图中的任何陈述事实相矛盾，尤其是正确答案。

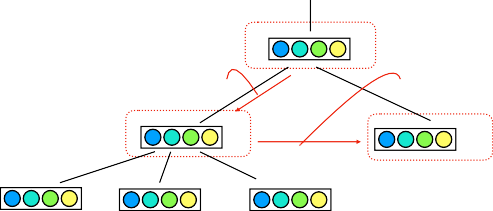
**攻击能力。**T3本质上是一个基于opti- mization的框架，通过优化目标的设定，对对抗性文本进行识别，从而实现**定向攻击**。对于情感分类任务，T3可以进行针对性的攻击，使得

我们设计了两种定向攻击：定位定向攻击和回答定向攻击，将原本积极的评论归为最消极的评论，反之亦然。特别是在QA任务中，我们设计并实现了两种定向攻击：*位置定向攻击*和*答案定向攻击*。成功的位置定向攻击是指模型可以被愚弄，将答案放在段落中特定的目标位置，但在攻击过程中，目标跨度上的内容是优化的。因此，在攻击之前无法确定答案。相反，一个成功的答案定向攻击是一个较强的定向攻击，它指的是无论问题是什么样子，模型总是输出预先确定的定向答案的情况。在表[1](#_bookmark0)中，我们将目标答案设置为"唐纳德-特朗普"，并成功改变了模型的预测。更多的答案目标攻击和位置目标攻击的例子可以在附录*§*[C](#_bookmark67)中找到。

虽然我们的框架被设计为白盒攻击，但我们的实验结果表明，对抗式文本可以转移到其他黑盒模型上，且攻击成功率很高。最后，由于T3是一个单一的对抗式文本生成框架，其输出是离散的标记，因此适用于不同的下游NLP任务。在本文中，我们以情感分类和QA为例进行对抗性评估来说明这一点。

## 树形自动编码器

在本小节中，我们描述了T3的关键组件：一个基于树的自动编码器。与标准的顺序生成方法相比，以非单调顺序（例如，沿着解析树）生成句子最近是一个有趣的话题（[Welleck等人](#_bookmark48)，[2019](#_bookmark48)）。我们的动机来自于这样一个事实，即沿着解析树生成句子可以内在地捕获和保持同步战术信息（[Eriguchi等人](#_bookmark20) ，[2017](#_bookmark20)；[Aharoni和Goldberg](#_bookmark11)，[2017](#_bookmark11)；[Iyyer等人](#_bookmark28)，[2018](#_bookmark28)），并显示出比顺序递归模式更好的性能（[Li等人](#_bookmark35)，[2015](#_bookmark35)；[Iyyer等人](#_bookmark27)，[2014](#_bookmark27)）。因此，我们设计了一种新型的基于树的自动编码器来生成对抗性文本，可以同时保留原始句子的语义和句法结构。此外，语言的离散性促使我们利用自动编码器将离散文本映射到一个高维的连续空间中，在这个空间上，可以通过基于梯度的ap- proaches来计算对抗性的per- turbation，从而实现针对性的攻击。



**ROOT**

***<根>***

LSTM Cell

**谎言**

***<nsubj>***

***<nmod>***

LSTM Cell

LSTM Cell

**猫咪**

**地板**

***<det> <amod >***

**…**

**a**

**昏昏欲睡**

**褐色**

图1：树状解码器。depen- dency树中的每个节点都是一个LSTM单元。黑线指的是父节点和子节点之间的依赖关系。红色箭头指的是解码的方向。在每个步骤中，解码器输出一个标记，显示在节点的右边。

形式上，让*X*是文本的域，*S*是*X*中元素上依赖性解析树的域，一个基于树的自动编码器由一个en- coder :*X S Z，将*文本***x*** *X*连同其依赖性解析树***s*** *S*编码成一个高维的潜伏表示***z*** *Z*和一个解码器 。*Z S X*从给定的上下文向量***z***和预期的依赖性解析树***s***生成cor-响应的文本***x***，给定一个依赖性树***s***，并形成一个反编码器。因此，我们有以下重建损失来训练我们基于树的自动编码器。

*εζ*

*ζx→*

*e*

*e*

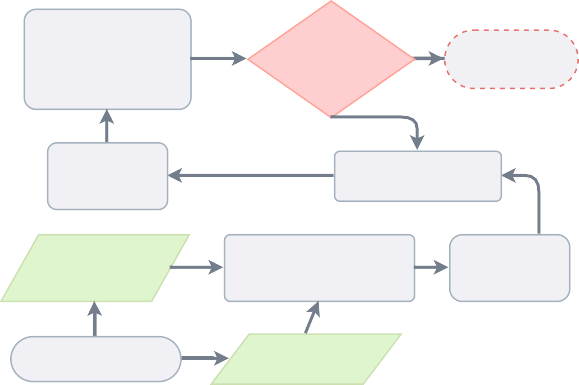
*εx→e*

*/*recon = *\_****Ez****~X*[log *pg*(***x****|****s****, ε*(***x****,* ***s***)] (1)

**编码器。**我们采用Child-Sum Tree-LSTM([Taiet al.](#_bookmark45) , [2015](#_bookmark45))作为我们的树形编码器。具体来说，在编码阶段，每个子节点的状态嵌入是其隐藏的树LSTM状态与依赖关系嵌入的并联。par- ent状态嵌入是通过对其子节点的状态嵌入进行求和提取，并通过树-LSTM单元向前馈送。该过程沿着CoreNLP解析器提取的依赖关系树([Manning et al.](#_bookmark38) , [2014](#_bookmark38))从底部(叶节点，即词)到顶部(根节点)进行。

**解码器。**由于目前还没有基于树的au-toencoder，我们设计了一种新型的树型解码器(如图[1](#_bookmark2)所示)。在解码阶段，我们从根节点开始，沿着相同的depen- dency树按级别顺序遍历。 下一个节点*j*的隐藏状态*hj*来自于(i)当前树节点的隐藏状态*hi*，(ii)当前节点预测词嵌入***wi***，(iii)当前节点*i*和下一个节点*j*之间基于依赖树的依赖关系em- bedding *dij*。根据LSTM单元*hj*的隐藏状态，生成下一个节点对应的词*yj*，通过

图2：对抗式文本生成的流水线。



NLP模型

(QA系统/分类员)

预测

是针对性的回答？

是的，是的

对抗性

文字

对抗性案文

树形解码器

扰动

嵌入

没有

扰动

嵌入

目标label Convert into 树

or answer statement by rules 编码器

开始

种子句

一个线性层，从隐藏的presenta- tion *hj*映射到代表树的词汇的概率分布的logits。

*hj* = LSTM([*hi*; *wi*; *dij*]) (2)

*yj* = one-hot(argmax (***W*** *. hj* + ***b***)) (3)

此外，树形结构允许我们在不同的树形层次上修改树形节点嵌入，以便在词级或句子级上产生可控的扰动，因此，我们对根级和叶级的T3(SENT)和T3(WORD)两种类型的攻击进行了研究，如图3和图4所示。因此，我们在根级和叶级T3(SENT)和T3(WORD)上分别例举了以下两种类型的攻击，如图[3](#_bookmark5)和图[4](#_bookmark6)所示。

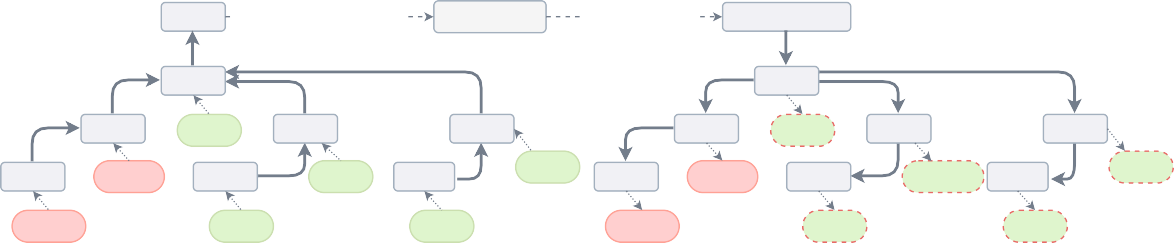
## 对抗式文本生成的流水线

在这里，我们说明如何使用我们的基于树的au- toencoder来执行对抗性文本生成和攻击NLP模型，如图[2所示](#_bookmark3)。

**第一步：选择对抗性种子。**敌意种子是我们的树状au- toencoder的输入句。在树节点嵌入上添加扰动后，解码后的对抗性句子将被添加到原始段落中进行协整攻击。对于情感分类器来说，对抗性种子可以是段落中的任意句子。例如，表[1](#_bookmark0)中Yelp Review例子的对抗性种子是来自段落*"I kept expecting to see donkeys and chickens walking around'的*随机sen- tence*。*

相反，当对QA模型进行答案定向攻击时，我们需要在合理的语境下将我们的目标答案加入到我们的对抗性种子中。基于一组关于对抗性种子如何与攻击efﬁcacy相关的启发式实验（附录[A.4](#_bookmark60)），我们选择使用ques- tion词来制作对抗性种子，因为当模型在匹配上下文之间的语义相似性时，它可以获得更高的关注度得分。

**Tree-Encoder 树形解码器**



ROOT

+

扰动

Adv ROOT

nsubj

根

VB

穿刺

nsubj

根

VB

穿刺

复式

兴趣

复式

兴趣

NN NNPendingNNPendsCD

NNP 特朗普

DT

拘留系列

IN

1989

案子

NNP 特朗普

DT

节目

IN

1988

案子

唐纳德

的

在

唐纳德

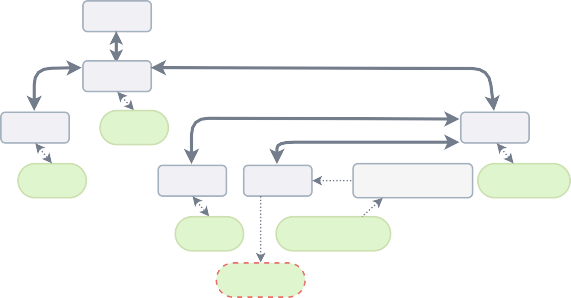
a

关于

**对立性种子**：[唐纳德-特朗普]在1989年结束了这个系列。对立**句**：[唐纳德-特朗普]在1988年结束一个节目。[唐纳德-特朗普]在1988年结束了一个节目。

图3：T3(SENT)如何生成对抗句的一个例子。在ROOT嵌入上加入扰动，并进行优化，以保证定向攻击的成功，同时将扰动的幅度降到最低。

**树形自动编码器**



nsubj

ROOT

根

VBD

兴趣

拘留

PRP

曾

复式

NN

I

DT

NN

扰动情况

+

风湿

惊骇

**对立面种子**：我有一个紧急状况。

**对立的句子**。我有一个令人震惊的情况。

图4：T3(WORD)在叶节点em- bedding上增加扰动。箭头表示编码/解码的方向。

和问题。具体来说，我们将一个问题句转换为一个有意义的陈述句，并分配一个有针对性的假答案。假答案可以根据扰动模型的预测答案来制作（位置定向攻击[3.1](#_bookmark1)），也可以由对手手动选择（答案定向攻击）。例如，表[1](#_bookmark0)所示的答案定向攻击示例通过一组粗粒度规则（附录[A.4](#_bookmark60)）将ques- tion*"Who ended the series in 1989?"*转换为一个声明性语句*"someone ended the series in 1989."。*然后，我们的目标错误答案被分配到生成对立种子*"Donald Trump ended the series in 1989"。*以下步骤将确保解码后的对立句与原段不矛盾。

*§*

**第二步：将离散文本嵌入到con- tinuous embedding中。**T3(SENT)和T3(WORD)的一个区别是在哪个树级上我们嵌入离散句。对于T3(SENT)，我们使用Tree-LSTM ***z*** = hroot的树根节点嵌入来表示离散句子（图[3](#_bookmark5)中的"ROOT"节点）。至于T3(WORD)。

我们将Tree-LSTM *hi*的所有叶子节点嵌入（对应于每个词）***z*** = [h1*，*h2*，......，****hn***]来嵌入离散句。

**第三步：通过优化扰动嵌入。**在嵌入向量*z*上找到最优扰动*z\*，*相当于求解优化问题，可以实现目标攻击目标，同时使per- turbation的幅度最小化。

min*||****z****\*||p* + *cf* (***z*** + ***z****\**)*,*(4)

其中*f*是目标攻击的目标函数，*c*是扰动幅度和攻击目标之间的常量平衡。具体来说，我们设计的目标函数*f*类似于[Carlini和Wagner](#_bookmark14)（[2016](#_bookmark14)）的分类任务的目标函数*f*

┐ j、 ┌╱ ,*ζ /i*

*l*=max Z(***z****/，****s***)；***x***：*i*=*t ，(*5)

jj、┌╱╱╱

*f* (***z****/*) = max *l \_ Zζ*(***z****/,* ***s***); ***x*** *t , \_κ ,* (6)

其中***z****/* = ***z*** + ***z****\**为扰动嵌入，模型输入[ (***z****/，****s***)；***x***]为广告句(***z****/，****s***)和原始段落***x***的连词，*t*为目标类，*Z*( )为分类模型在softmax前的logit输出，*l*为目标类以外的类的最大- imum logits，*κ*为调整误分类率的conﬁdence score。通过二元搜索来选择conﬁdence score *κ*，以寻找攻击成功率和意义毅力之间的折衷常数。通过梯度下降法对最优解*z\**进行迭代优化。

*.*

*ζ*

*ζ*

类似于攻击QA模型，我们巧妙地改变目标函数*f，*由于QA模型和分类模型之间的差异b- tween。

*lj* = max ┐*Zj* j、***x***；*ζ*(***z****/，****s***)┌╱*i* : *i /*= *tj*，*。*

*f* (***z****/*) = max *jlj \_ Zj* j、***x***; *ζ*(***z****/，****s***)┌╱*tj ，\_κ*╱*。*

2

*j*=1

其中，Z1( )和Z2( )分别是QA系统的答案起始位置和结束位置的对数，t1和t2分别是目标起始位置和目标结束位置，*lj*是目标位置以外位置的最大对数。与攻击senti- ment classiﬁer在前面加上对抗句不同，我们选择按照[Jia和Liang](#_bookmark29)的设定，将对抗句加到段末，这样可以和他们的结果进行公平的比较。

*..*

**第四步：解码回对立句。**在将嵌入映射到对抗性句子时，我们需要处理三个问题。(1)对抗句可能与原段的陈述事实相矛盾；(2)解码步骤(式[3](#_bookmark4))使用的是argmax算子，没有给出梯度，但步骤3需要进行梯度下降来寻找最优的*z\**；(3)对于an- swer的定向攻击，目标答案可能在解码阶段被扰动和改变。

为了解决问题(1)，我们保证我们的ap- pended对抗句与地真不矛盾，保证对抗句和答案句没有共同词，否则保持迭代步数。如果达到最大步数，则视为优化失败。

对于问题(2)，在优化过程中，我们采用基于softmax的连续近似，温度*τ*不断降低([Hu等](#_bookmark25)，[2017](#_bookmark25))

*yj\** ~ softmax((***W*** *.* ***hj*** + ***b***)*/τ* )*.*(7)

以使优化可分化。在确定了最优扰动*z\**后，我们仍然使用硬argmax来生成对抗性文本。

对于问题(3)，我们在优化步骤中保持目标答案不被修改，对目标答案跨度设置门：*yj* g1 *yj* + g2 *xj，*(*j* = t1*，*t1 + 1*，...，*t2)，其中*yj*是树解码的对抗性标记。我们在位置目标攻击中设置g1=1，g2=0，在答案目标攻击中设置g1=0，g2=1。

*o*

*-o*

# 实验

我们现在介绍T3的实验评估结果。特别是，我们针对两个流行的NLP任务，情感分类和QA。对于这两个模型，我们进行了基于白盒和转移性的黑盒攻击。除了模型的准确性（非目标攻击评估），我们还进行了

报告了T3的目标攻击成功率。我们表明，所提出的T3可以在不同的mod- els上优于其他最先进的基线方法。The details of pretraining tree decoder and experimental setup can be found in Appendix *§*[A](#_bookmark57) and *§*[B](#_bookmark62).

## 情感分类的对抗性评估设置

在这个任务中，情感分析模型将餐厅和商店的用户评论作为输入，并期望预测用户被赋予的星级数（从1星到5星）。

**数据集。**我们选择Yelp数据集（[Challenge](#_bookmark15)）进行情感分析任务。它由2.7M条yelp评论组成，我们按照[Lin等](#_bookmark37)（[2017](#_bookmark37)）的流程，随机选取500K条评论-明星对作为训练集，2000条作为开发集，2000条作为测试集。

**模型。***BERT* ([Devlin et al.](#_bookmark17) , [2019](#_bookmark17))是一个基于trans- former ([Vaswaniet al.](#_bookmark46) , [2017](#_bookmark46))的模型，它在一个大型语料库上进行了无监督的预训练，并被证明对下游NLP任务有效。*Self-Attentive Model* (*SAM)* ([Lin et al.](#_bookmark37) , [2017](#_bookmark37))是一个使用自我关注机制的最先进的文本分类模型。更详细的模型设置在附录中列出。

**评估指标。***目标攻击成功率*(简称target)是以平均有多少个考试成功攻击输出目标标签来衡量，而*非目标攻击成功率*(简称untarget)则是计算被攻击输出与地面真相不同标签的例子的百分比。

**攻击基线。***Seq2sick*([Chengetal.](#_bookmark16) , [2018](#_bookmark16))是一种攻击seq2seq模型的白盒投影梯度方法。在这里，我们通过改变其损失函数对情感分类模型进行seq2sick攻击，这在origi- nal论文中没有评估。*TextFooler* ([Jin et al.](#_bookmark31) , [2019](#_bookmark31)) 是一种简单而强大的黑盒攻击方法，用于执行词级的原地敌意修改。按照相同的设置，Seq2Sick和TextFooler只允许编辑前缀句。

## 问题回答系统的对抗性评价设置

**任务和数据集。**在这个任务中，我们选择SQuAD数据集([Rajpurkar et al.](#_bookmark53) , [2016](#_bookmark53))进行问答任务。SQuAD数据集是一个阅读理解数据集，由107,785个问题组成，这些问题由人群工作者提出，在一系列的

模型Original 白盒Attack 黑盒攻击

AccT3(WORD)Seq2SickT3(WORD)Seq2sick TextFooler

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BERT | 0.703 | 目标非目标 | **0.990**  **0.993** | 0.974  0.988 | **0.499**  **0.686** | 0.218  0.510 | 0.042  0.318 |
| 萨姆 | 0.704 | 目标 | **0.956** | 0.933 | **0.516** | 0.333 | 0.113 |
|  |  | 非目标 | **0.967** | 0.952 | **0.669** | 0.583 | 0.395 |

表2：从针对性和非针对性攻击成功率方面对情感分类器进行对抗性评价。

模型Origin 白盒Attack 黑盒攻击

Pos-T3() WORDAns-T3() WORDPos-T3() WORDAns-T3(WORD)AddSent)。

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| BERT | EM F1 | 81.2  88.6 | **29.3**  **33.2** | 43.2  47.3 | **32.3** / 52.8  **36.4** / 57.6 | 45.2 / 51.7  49.0 / 55.9 | 46.8  52.6 |
| BiDAF | 电磁 | 60.0 | **15.0** | 21.0 | **18.9** / 29.2 | 20.5 / 28.9 | 25.3 |
|  | F1 | 70.6 | **17.6** | 23.6 | **22.5** / 34.5 | 24.1 / 34.2 | 32.0 |

表3：QA模型的对抗性评价。Pos-T3和Ans-T3分别指位置定向攻击和答案定向攻击。基于转移性的黑盒攻击使用相同架构（前者得分）和不同架构（后者得分）的白盒模型生成的对抗性文本。

维基百科文章，其中每个ques- tion的答案必须是对应阅读段落中的一段文字。为了将我们的方法与其他对抗性评估作品（[Jia和Liang](#_bookmark29)，[2017](#_bookmark29)）在QA任务上进行比较，我们在与[Jia和Liang](#_bookmark29)（[2017](#_bookmark29)）相同的测试集上评估我们的对抗性攻击，该测试集由1000个来自SQuAD开发集的随机采样示例组成。

**模型。**我们采用与[De- vlin等人](#_bookmark17)([2019](#_bookmark17))相同的策略，改编了*BERT*模型在SQuAD v1.1上运行，并在开发集上重现了结果。*BiDAF*([Seo et al.](#_bookmark44) , [2016](#_bookmark44))是一个多阶段的分层过程，它在不同的粒度层次上表示上下文，并使用双向关注浏机制来获得一个查询感知的上下文表示。

**评估指标。**对于非目标攻击评估- uation，我们使用SQuAD数据集的ofﬁcial脚本（[Rajpurkar等人](#_bookmark53)，[2016](#_bookmark53)）来衡量对抗性精确匹配率和F1分数。较低的EM和F1分数意味着较好的攻击suc- cess率。对于目标攻击评估，我们使用目标精确匹配率和目标F1得分，计算有多少模型输出与目标假答案（*例如*，表[1](#_bookmark0)中的假答案"Don- ald Trump"）相匹配。更高的目标EM和F1意味着更高的目标攻击成功率。

**攻击基线。***AddSent*（[Jia和Liang](#_bookmark29)，[2017](#_bookmark29)）在给定文本中附加一个人工构建的合法干扰句，从而引入虚假信息，只能进行无目标攻击。*通用对抗性触发器*（[Wallace](#_bookmark47)

[等人](#_bookmark47) ，[2019](#_bookmark47)）是输入不可知的标记序列，当连接到数据集的任何输入时，会触发模型产生一个规格化的预测。

## 对抗性评价

* + 1. **T3(WORD)**

**攻击情绪分类器。**我们执行基线攻击和我们的T3攻击在concat at- tack方案下的白盒和黑盒设置。我们的目标目标为情绪分类- tion是相反的情绪。具体来说，我们将目标攻击目标设置为5星的评论原初低于3星和1星以上的评论。我们将我们的结果与强势词级攻击者Seq2sick进行比较，如表[2](#_bookmark7)所示。可以看到我们的T3(WORD)在白盒设置下，在BERT模型上的表现优于基线，达到了近100%的攻击成功率。

我们还进行了基于可转移性的黑盒攻击。具体来说，基于转移性的黑盒攻击使用从白盒BERT模型生成的对抗性文本来攻击黑盒SAM，反之亦然。我们将我们的黑盒攻击成功率与黑盒基线TextFooler和基于transferability的黑盒Seq2Sick进行比较。表[2](#_bookmark7)显示了我们的T3(WORD)模型在所有基线模型中仍然具有最好的黑盒目标攻击和未被攻击的成功率。

**攻击QA模型。**我们对测试模型进行白盒攻击和基于转移性的攻击。如表 [3](#_bookmark8) 所示，T3(WORD)在这两个模型上实现了最好的白盒攻击结果。

|  |  |  |  |
| --- | --- | --- | --- |
| 型号 | T3(SENT) | T3(WORD) | UT |
| BERTtarget EM | 32.1 | **43.4** | 1.4 |
| 目标F1 | 32.4 | **46.5** | 2.1 |
| BiDAFtarget EM | 53.3 | **71.2** | 21.2 |
| 目标F1 | 56.8 | **75.6** | 22.6 |

表4：针对QA的白盒攻击结果。UT是通用触发器基线的简称。

BERT和BiDAF。值得注意的是，虽然BERT的性能优于BiDAF，但BERT的per-formance降幅∆F1BERT比BiDAF的性能降幅∆F1BiDAF=53*.*0大55*.*4，这再次证明了BERT在对抗性评价下是不安全的。我们还发现位置定向攻击比答案定向攻击略强。我们认为这是因为答案靶向攻击的目标答案是ﬁxed的，改变附句的自由度有限，但位置靶向攻击从目标位置跨度上改变假答案的自由度更大。

然后，我们对QA模型上的定向攻击perfor- mance进行评估。结果如表[4](#_bookmark9)所示。表明T3(WORD)对QA的定向攻击能力最好。而我们所有的攻击方法都优于基线。

我们还将白盒攻击产生的对抗性文本转移到执行黑盒攻击。Ta- ble [3](#_bookmark8)显示了对测试模型进行黑盒攻击的结果。我们提出的所有方法在具有相同架构的模型之间传输对抗者时，都优于基线方法（AddSent）。

## 人类评价和T3（SENT）

我们进行了全面的人类主体评估，以评估人类对不同类型生成的对抗性文本的反应。主要结论是，即使这些对抗性例子在攻击机器学习模型方面是有效的，但人类对它们的注意程度却要低得多。

## 评估指标和设置

我们专注于两个指标来评估生成的对抗句的有效性：**对抗性文本质量**和**人类**在原始数据集和对抗性数据集上的**表现**。为了评估ad- versarial文本质量，要求人类参与者选择他们认为质量更好的数据。为了确保人类不会被我们的对抗性例子所误导，我们要求人类参与者在原始数据集和对抗性数据集上执行

在原始数据集和对抗性数据集上进行情感分类和问题回答任务。我们将对抗性数据集和原始数据集分发给533个亚马逊Turkers来进行人工评估。更多的实验设置细节可以在附录*§*[B.4](#_bookmark65)中找到。

## 分析报告

人工评价结果如表[5](#_bookmark10)所示。我们看到，T3(SENT)的总体投票率较高，这意味着从人类的角度来看，它的语言质量比T3(WORD)更好。我们假设原因是T3(SENT)在解码阶段是在依赖性约束下进行解码的，所以它可以更充分地利用基于树的au- toencoder结构。而我们有理由看到，更好的语言质量是以较低的对抗成功率为代价的。如表[5](#_bookmark10)所示，T3(SENT)在SAM上的对抗性目标成功率比T3(WORD)低20%，这说明语言质量和对抗性攻击成功率之间的权衡。

原始数据集和对抗式数据集的人类得分也显示在表[5](#_bookmark10)中。我们可以看到，人类的表现几乎没有受到连词对抗句的影响。具体来说，基于T3的QA和分类- ﬁcation任务的得分都下降了10%左右。这优于最先进的算法([Jia和Liang](#_bookmark29)，[2017](#_bookmark29))，该算法对人类per- formance有14%的性能下降。

我们还分析了人为错误案例。进一步的定量分析（附录[B.5](#_bookmark66)）表明，大多数错误的人类答案并不指向我们生成的假答案，而是可能来自于汇总人类结果时的抽样噪声。

*§*

另外，我们发现在我们附加了对抗句后，对抗句段的平均长度比原段的平均长度多了约12个tokens。我们猜测段子长度的增加对人的表现也有影响。

在附录[A](#_bookmark57)中，我们进行了一些消融stud- ies来探索不同自动编码器的攻击效果。我们还通过改变不同的攻击参数（如附加的对抗句的位置）来研究BERT的注意力，并得出了几个有趣的结论。附录[C](#_bookmark67)显示了更多的对抗性例子。

*§*

*§*

方法情感Classiﬁer QA

Human Human Models Quality 原产地 Human Human Models 质量

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| T3(SENT) |  | 0.82 | 0.363 / 0.190 | 65.67% |  | 81.78 | 49.1 / 29.3 | 69.50% |
| T3(WORD) | 0.95 | 0.82 | 0.007 / 0.033 | 34.33% | 90.99 | 82.90 | 29.3 / 15.0 | 30.50% |

表5：人类对T3(SENT)和T3(WORD)的评价。"Origin Human"为原始数据集上的人类评分。"人类"是人类在对抗性数据集上的得分。

# 讨论和今后的工作

除了一般的对抗式评估框架T3，本文还旨在探讨七个科学问题。1) 既然T3允许在树形层次结构的不同层次上进行灵活操作，那么哪个层次的攻击更有效，哪个层次的攻击能更好地保持语法的正确性？2) 在有限的操作自由度下，是否有可能实现一般NLP任务的定向攻击，如情感分类和QA？3）对于很多NLP任务，是否有可能形成黑盒攻击？4）BERT在实践中是否稳健？5）这些对抗性的例子会影响人类读者的表现吗？

We ﬁnd that:1）词和句子层面的攻击都可以达到很高的攻击成功率，而句子层面的操作整合了全局语法约束，可以生成高质量的对抗性句子。2）可以对一般的NLP任务进行各种针对性的攻击（*例如*，在攻击QA时，我们可以确保目标是一个特定的答案或一个句子中的特定位置）；3）基于转移性的黑盒攻击在NLP任务中是成功的。4）虽然BERT已经达到了最先进的性能，但我们观察到在面对对抗性例子时，性能下降也比其他模型更多，这说明BERT在对抗性设置下不够稳健。

除了上述结论外，我们还总结了一些有趣的结论。(1)虽然T3(WORD)在多个任务中实现了最好的攻击成功率，但我们观察到了一个折中的现象--在操作的自由度和攻击能力之间。例如，T3(SENT)有depen- dency树的约束，对于人类读者来说，变得比T3(WORD)更自然，但对于攻击模型来说，效果不如T3(WORD)。同样，由于tar- geted的答案是ﬁxed的，QA中的答案定向攻击比po- sition定向攻击可以操纵的字数更少，因此攻击性能稍弱。(2)从性能较好的模型中转移广告文案

到较弱的则更成功。例如，将BERT- QA中的对抗性例子转移到BiDAF中，实现的攻击suc- cess率比反过来要好得多。(3)我们还注意到，在相似架构的模型之间，对抗性例子的转移能力比不同架构的模型更好。(4)BERT模型对para-图的两端给予了更高的关注度得分，而倾向于忽略中段的内容，从[A.2](#_bookmark59)消融研究中可以看出，在段落中间添加对抗性句子的效果不如前面或结尾。

为了抵御这些对手，我们在此讨论以下几种可能的方法，并将在今后的工作中深入探讨。(1)**Ad- versarial Training**是一种实用的防御敌方例子的方法。然而，其缺点是我们通常无法提前知道威胁模型是什么，这使得对抗性训练在面对看不见的攻击时效果不佳。(2)**区间边界传播**(IBP)([Dvijotham et al.](#_bookmark18) , [2018](#_bookmark18))作为一种新的技术被提出，从理论上考虑最坏情况下的扰动。最近的工作([Jia等](#_bookmark30)，[2019](#_bookmark30);[Huang等](#_bookmark26)，[2019](#_bookmark26))已经在NLP领域应用IBP来认证模型的鲁棒性。(3)包括GPT2在内的**语言模型**([Radford et al.](#_bookmark52) , [2019](#_bookmark52))也可以作为异常检测器来探测不一致和不自然的对抗性句子。

*§*

# 二、结论

综上所述，我们提出了一种用于对抗性文本生成的通用目标攻击框架，这是我们所知的第一个成功对一般NLP任务进行任意目标攻击的方法。据我们所知，这是第一个成功地对一般NLP任务进行任意定向攻击的方法。我们的结果表明，我们的攻击可以在不欺骗人类的情况下实现高攻击成功率。这些结果揭示了一种有效的方法来检验广泛的NLP模型的鲁棒性，从而为开发新一代更可靠和有效的NLP方法铺平了道路。

# 鸣谢

这项工作得到了美国国家科学基金会拨款No.1910100，亚马逊研究奖，DARPA编号HR00111990074的部分支持。我们感谢匿名的重新查看者的深刻评论。

# 参考文献

Roee Aharoni和Yoav Goldberg。2017.[Towards string-to-tree neural machine translation](https://doi.org/10.18653/v1/P17-2021).In *Pro- ceedings of the 55th Annual Meeting of the Associa- tion for Computational Linguistics (Volume 2: Short Papers)*, pages 132-140, Vancouver, Canada.Asso- ciation for Computational Linguistics.

Moustafa Alzantot, Yash Sharma, Ahmed Elgohary, Bo-Jhang Ho, Mani B. Srivastava, and Kai-Wei Chang.2018.Generating natural language adversar- ial examples.In *EMNLP*.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio.2015.Neural machine translation by jointly learning to align and translate.*CoRR*，abs/1409.0473.

Nicholas Carlini and David A. Wagner.2016.To- wards evaluating the robustness of neural networks.*2017 IEEE Symposium on Security and Privacy (SP)*, pages 39-57.

Yelp数据集挑战赛。数据取自Yelp数据集挑战赛，[https://www.yelp.com/ dataset/challenge](https://www.yelp.com/dataset/challenge)。

郑敏浩，易金凤，张欢，陈品玉，谢周钜.2018.[Seq2Sick: Evaluating the Robustness of Sequence-to-Sequence Models with Adversarial Examples](http://arxiv.org/abs/1803.01128).*arXiv e-prints*, page arXiv:1803.01128.

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

Krishnamurthy Dvijotham、Sven Gowal、Robert Stan- forth、Relja Arandjelovic、Brendan O'Donoghue、Jonathan Uesato和Pushmeet Kohli。2018.Train- ing veriﬁed learners with learned veriﬁers.*ArXiv*, abs/1805.10265.

Javid Ebrahimi, Anyi Rao, Daniel Lowd, and Dejing Dou.2017.Hotﬂip:White-box adversarial exam- ples for text classiﬁcation.In *ACL*.

Akiko Eriguchi, Yoshimasa Tsuruoka, and Kyunghyun Cho.2017.[Learningtoparseandtranslateimproves neural machine translation](https://doi.org/10.18653/v1/P17-2012).In *Proceedings of the 55th Annual Meeting of the Association for Compu- tational Linguistics (Volume 2: Short Papers)*, pages 72-78, Vancouver, Canada.Association for Compu- tational Linguistics.

Kevin Eykholt、Ivan Evtimov、Earlence Fernandes、李博、Amir Rahmati、肖超伟、Atul Prakash、Tadayoshi Kohno和Dawn Xiaodong Song。2017.Robust physical-world attacks on deep learning models.

Ian J. Goodfellow、Jonathon Shlens和Christian Szegedy。2015.Explaining and harnessing adversar- ial examples.*CoRR*，abs/1412.6572。

何鲁恒、李肯顿、迈克-刘易斯和卢克-S-泽特莫耶。2017.Deep semantic role labeling:What works and what's next.In *ACL*.

Sepp Hochreiter和Ju¨rgen Schmidhuber。1997.长短期记忆。*神经计算*，9：1735-1780。

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan R. Salakhutdinov, and Eric P. Xing.2017.Toward con- trolled generation of text.In *ICML*.

Po-Sen Huang、Robert Stanforth、Johannes Welbl、Chris Dyer、Dani Yogatama、Sven Gowal、Krish- namurthy Dvijotham和Pushmeet Kohli。2019.Achieving veriﬁed robustness to symbol substi- tutions via interval bound propagation.*ArXiv*，abs/1909.01492.

Mohit Iyyer, Jordan L. Boyd-Graber, and Hal Daume'.2014.Generating sentences from semantic vector space representations.

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

贾彦宏和梁珀西。2017.[Adversarialexam- ples for evaluating reading comprehension systems](https://doi.org/10.18653/v1/D17-1215).In *Proceedings of the 2017 Conference on Empiri- cal Methods in Natural Language Processing*, pages 2021-2031, Copenhagen, Denmark.Association for Computational Linguistics.

Robin Jia, Aditi Raghunathan, Kerem Go¨ksel, and Percy Liang.2019.Certiﬁed robustness to adversar- ial word substitutions.*ArXiv*, abs/1909.00986.

金迪、金志敬、周天一和Peter Szolovits.2019.[Is BERT Really Robust?A Strong Baseline for Natural Language Attack on Text Clas- siﬁcation and Entailment](http://arxiv.org/abs/1907.11932).*arXiv e-prints*, page arXiv:1907.11932.

Diederik P. Kingma和Jimmy Ba.2014.Adam: A method for stochastic optimization.*CoRR*，abs/1412.6980。

Qi Lei，Lingfei Wu，Pin-Yu Chen，Alexand ros G. Di- makis，Inderjit S. Dhillon，and Michael Witbrock.Dhillon, and Michael Witbrock.2018.[Discrete Adversarial Attacks and Submodu- lar Optimization with Applications to Text Classiﬁ- cation](http://arxiv.org/abs/1812.00151). *arXiv e-prints*, page arXiv:1812.00151.

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*.

Jiwei Li, Thang Luong, Daniel Jurafsky, and Eduard H. Hovy.2015.When are tree structures necessary for deep learning of representations?In *EMNLP*.

梁斌，李洪成，苏苗强，边潘，李锡荣，史文昌。2017. Deep *arXiv preprint arXiv:1704.08006*.

Zhouhan Lin, Minwei Feng, C´ıcero Nogueira dos San- tos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio.2017.A structured self-attentive sentence embedding.*ArXiv*，abs/1703.03130.

Christopher D. Manning、Mihai Surdeanu、John Bauer、Jenny Rose Finkel、Steven Bethard和David Mc- Closky。2014.Stanford corenlp natural language processing toolkit.In *ACL*.

Paul Michel， Xian Li， Graham Neubig， 和Juan Miguel Pino.2019.On evaluation of ad- versarial perturbations for sequence-to-sequence models.In *NAACL-HLT*.

Takeru Miyato, Andrew M. Dai, and Ian Goodfel- low.2016.[Adversarial Training Methods forSemi-SupervisedTextClassiﬁcation](http://arxiv.org/abs/1605.07725).*arXiv e-prints*，page arXiv:1605.07725.

Seyed-Mohsen Moosavi-Dezfooli, Alhussein Fawzi, and Pascal Frossard.2016.Deepfool。A simple and accurate method to fool deep neural networks.*2016*

任书怀、邓义和、何坤、车万祥。2019.[Generating natural language adversarial ex- amples through probability weighted word saliency](https://doi.org/10.18653/v1/P19-1103).In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1085-1097, Florence, Italy.Association for Compu- tational Linguistics.

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

Min Joon Seo、Aniruddha Kembhavi、Ali Farhadi和Hannaneh Hajishirzi。2016.[Bidirectionalat- tention ﬂow for machine comprehension](http://arxiv.org/abs/1611.01603).*CoRR*，abs/1611.01603.

戴开胜，Richard Socher，and Christopher D. Manning.2015.Improved semantic representations from tree-structured long short-term memory net- works.In *ACL*.

Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N. Gomez、Lukasz Kaiser和Illia Polosukhin。2017.注意力是你所需要的。在*NIPS*中。

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh.2019.[Universal adversarialtrig- gers for attacking and analyzing NLP](https://doi.org/10.18653/v1/D19-1221).In *Proceed- ings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th Inter- national Joint Conference on Natural Language Pro- cessing (EMNLP-IJCNLP)*, pages 2153-2162, Hong Kong, China.Association for Computational Lin- guistics.

Sean Welleck, Kiante´ Brantley, Hal Daume´, and

*IEEE计算机视觉和模式识别会议（CVPR）*，第2574-2582页。

Nicolas Papernot、Patrick McDaniel、Ananthram Swami和Richard Harang。2016.[Crafting Ad- versarial Input Sequences for Recurrent Neural Net- works](http://arxiv.org/abs/1604.08275).*arXiv e-prints*，page arXiv:1604.08275.

Nicolas Papernot、Patrick D. McDaniel、Xi Wu、Somesh Jha和Ananthram Swami。2016.Distilla- tion as a defense to adversarial perturbations against deep neural networks.*2016 IEEE Symposium on Se- curity and Privacy (SP)*, pages 582-597.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning.2014.Glove:Glove: Global vectors for word rep- resentation.In *EMNLP*.

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。

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang.2016.[SQuAD: 100,000+ questions for machine comprehension of text](https://doi.org/10.18653/v1/D16-1264).In *Proceedings of the 2016 Conference on Empirical Methods in Natu- ral Language Processing*, pages 2383-2392, Austin, Texas.Association for Computational Linguistics.

赵京贤。2019.非单调的顺序文本生成。In *ICML*.

臧元，杨成浩，齐凡超，刘志远，张萌，刘群，孙茂松。2019.Textual adversarial attack as combinatorial optimiza- tion. *arXiv: Computation and Language*.

张黄钊，周浩，苗宁，李磊.2019.[Generating ﬂuent adversarial examples for natural languages](https://doi.org/10.18653/v1/P19-1559).In *Proceedings of the 57th An- nual Meeting of the Association for Computational Linguistics*, pages 5564-5569, Florence, Italy.Asso- ciation for Computational Linguistics.

Z.Zhao, D. Dua, and S. Singh.2017.[Generating Nat-](http://arxiv.org/abs/1710.11342) [ural Adversarial Examples](http://arxiv.org/abs/1710.11342).*ArXiv e-prints*.

# 消融研究

## 自动编码器选择

作为消融研究，我们比较了标准的基于LSTM的自动编码器与我们基于树的au- toencoder。

表6：针对QA的定位定向攻击能力的消融研究。EM和F1得分越低，说明攻击成功率越高。T3(SENT)和T3(WORD)分别指T3(SENT)和T3(WORD)。Adv(seq2seq)是指使用基于LSTM的seq2seq模型作为文本自动编码器的T3。

OriginT3(SENT)T3(WORD)Adv(seq2seq)

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 电磁 | 60.0 | 29.3 | **15.0** | 51.3 |
| F1 | 70.6 | 34.0 | **17.6** | 57.5 |

**树自动编码器。**在整个实验中，我们使用Stanford TreeLSTM作为树形编码器和我们提出的树形解码器一起作为树形自动编码器。我们在包含500K评论的Yelp数据集上训练树型自动编码器。该模型预计读取一个句子，将该句子映射到一个潜伏空间中，并以无监督的方式从嵌入中重建该句子以及依赖树结构。该模型采用300-d向量作为隐藏树节点嵌入，在自适应学习率和权重衰减的情况下，训练30个纪元。训练后，测试集的平均重构损失为0.63。

**Seq2seq自动编码器。**我们还评估了基于LSTM的标准架构(seq2seq)作为T3流水线中不同的自动编码器。对于seq2seq编码器-解码器，我们使用一个双向LSTM作为编码器（[Hochreiter和Schmidhu- ber](#_bookmark24)，[1997](#_bookmark24)），并使用一个两层LSTM加编码状态上的软衰减机制作为去编码器（[Bahdanau等人](#_bookmark13)，[2015](#_bookmark13)）。在400-d的隐藏单元和0.3的丢弃率下，最终的测试重建损失为1.43。

白盒攻击能力与知名QA模型BiDAF的对比见表[6](#_bookmark58)。我们可以看到基于seq2seq的T3未能取得良好的攻击成功率。此外，由于vanilla seq2seq模型没有考虑语法约束，重构损失较大，所以生成的对抗性文本的质量无法保证。

## BERT注意力的消融研究

为了进一步探讨对抗性句子的位置如何影响攻击成功率，我们con-。

duct的消融实验，通过改变附加的对抗性句子的posi- tion。我们从白盒BERT分类和QA模型中生成对抗性句子。然后我们将这些对抗句注入到原始段落的不同位置，并在另一个具有相同架构但参数不同的黑盒BERT中进行测试。结果如表[7](#_bookmark61)和[8](#_bookmark61)所示。我们看到在大多数时间里，在段落的b- ginning处追加敌手句子实现了最好的攻击性能。另外在段末附加对抗句的性能通常比前面稍弱。这一观察结果表明，BERT模型可能会更多地关注段落的两端，而倾向于忽略中间的内容。

## 攻击设置

我们使用Adam（[Kingma和Ba](#_bookmark32)，[2014](#_bookmark32)）作为opti- mizer，设置学习率为0.6，优化步骤为100。我们按照[Carlini和Wagner](#_bookmark14)([2016](#_bookmark14))的方法，通过二元搜索来寻找对象函数中合适的参数(权重const *c*和conﬁdence score *κ*)。

## 启发式实验选择QA的对抗性种子。

关于如何选择一个好的初始化句子来更有效地攻击QA模型，我们进行了以下启发式实验。根据实验，我们认为在攻击QA模型时，选择一个语义上接近上下文或问题的句子作为初始种子是很重要的，这样我们就可以减少itera- tion步骤的数量，更有效地确定对手愚弄模型。这里我们描述了三种选择初始句子的方法，我们将展示在相同的最大优化步数下，这些方法的有效性。

**随机对抗性种子句。**我们的第一个试验是使用一个随机句子（除了答案句），生成一个与真实答案相似的假答案，并将其追加到后面作为初始种子。

**基于问题的对抗性种子句。**我们也尝试使用问题词来制作一个ini- tial句子，理论上，当模型在匹配上下文和问题之间的特征相似性时，它应该获得更多的关注。为了将问题句转化为有意义的陈述句，我们采用以下步骤。

表7：将白盒生成的adv句子插入到不同的posi- tions进行BERT-classiﬁcation后的黑盒攻击成功率。

表8：BERT-QA将白盒生成的对抗句插入到不同位置后的黑盒攻击成功率。

Method Back Mid 前面

target 0.739 0.**6780.820**

T3(WORD)

untarget 0.817 0.**7700.878**

目标**0.**0.174 2200.217

T3(SENT)

untarget 0.531 0.**5040.532**

Method Back Mid 前面

EM 32.3 39.**131.9**

T3(WORD)

F1 36.4 43.**436.3**

EM 47.0 51.**342.4**

T3(SENT)

F1 52.0 56.**747.0**

在步骤1中，我们使用最先进的语义角色标签（SRL）工具（[He等](#_bookmark23)，[2017](#_bookmark23)）将问题解析为动词和论据。一组规则被deﬁned，以删除con- tain问句词和不重要的形容词等参数。下一步，我们访问模型的原始预测答案并定位答案句。我们再次运行SRL解析，并确定答案属于哪个参数。整个an- swer参数被提取出来，但答案标记被替换为我们的目标答案或GloVe词向量（[Pennington et al.](#_bookmark51) ，[2014](#_bookmark51)）中的近似词（位置定向攻击），该词也被用于QA模型。通过这种方式，我们制作了一个共享答案上下文的假答案，从起点上解决兼容性问题。最后，我们用假的论据替换陈述句的去掉的ar- guments，并选择这个基于问题的句子作为我们的初始句。

**基于答案的对抗性种子句。**我们也可以考虑直接使用模型预测的原始答案句子，并进行一些替换作为初始句。要制作一个假的答案sen- tence比从问题词中制作要容易得多。与创建基于问题的初始句的步骤2类似，我们请求模型的原始预言答案，并筛选出答案句。答案句中的答案跨度直接用GloVe词向量空间中最接近的词代入，以初步避免兼容性问题。

**实验结果。**我们在T3(WORD)上尝试了上述ini- tial句子选择方法，并在BERT-QA上执行位置定向攻击，给定相同的最大优化步骤。实验结果如表[9](#_bookmark64)所示。从表中可以看出，我们发现使用不同的初始化方法将极大地影响攻击成功率。There- fore，初始句子选择方法在本质上是很重要的，有助于减少迭代步骤的数量，快速收敛到可以攻击模型的最优*z\**。

# 实验环境

## 情感分类模型

**BERT。**我们使用12层的BERT基础模型[1](#_bookmark63)，有768个隐藏单元，12个自注意力头和110M参数。我们在500K的复习训练集上对BERT模型进行了调整，用于文本分类，批量大小为32，最大序列长度为512，学习率为2e-5，共3次。对于长度大于512的文本，我们只保留第512个标记。

**自注意力模型（SAM）。**我们选择结构化的自注意力句子嵌入模型（[Linetal.](#_bookmark37) ，[2017](#_bookmark37)）作为测试模型，因为它不仅在senti- ment分析任务上取得了其他基线模型中最先进的结果，而且还提供了一种定量确定模型注意力的方法，并帮助我们进行和分析我们的对抗性攻击。具有10个注意力跳的SAM内部使用了一个300-dim的BiLSTM和输出层前的512个单元全连接层。我们在初始学习率为0.1的情况下，用随机梯度下降优化器在500K复习训练集上训练了29个epochs的SAM。

## 情绪分类攻击基线

**Seq2sick**([Chengetal.](#_bookmark16) , [2018](#_bookmark16))是一个白盒pro-jected梯度方法，结合组套索和梯度正则化，制作对抗性的ex-amples来欺骗seq2seq模型。在这里，我们将损失函数定义为*Ltarget* = max *z*(*k*) *z*(*t*)以

┐, *\_*

*e*

*k Y*

攻击情绪分类模型

在原论文中没有进行评估。在我们的设置中，Seq2Sick只允许编辑附加的句子或标记。

**TextFooler**([Jin et al.](#_bookmark31) , [2019](#_bookmark31))是一种简单但强大的黑盒攻击方法，用于生成广告性文本。在这里，TextFooler也只允许编辑附加的句子。

1 https://github.com/huggingface/pytorch-pretrained-。

BERT

表9：BERT-QA的白盒攻击结果，从精确匹配率和F1分数上看，是由不同的评估脚本决定的。EM和F1得分越低，说明攻击成功率越高。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 型号 | 起源 | 随机 | 位置 目标攻击  以问题为基础 以答案为基础 | 基线  添加发送 |
| BERTEM | 81.2 | 67.9 | **29.**350.6 | 46.8 |
| F1 | 88.6 | 74.4 | **33.**255.2 | 52.6 |

## 质量保证模型

**BiDAF.Bi-Directional Attention Flow(BIDAF)网络。**Bi-Directional Attention Flow(BIDAF)网络([Seoetal.](#_bookmark44) , [2016](#_bookmark44))是一个多阶段的分层过程，它在不同的粒度层次上表示上下文，并使用双向注意流机制来获得一个查询感知的上下文表示。我们在([Seo et al.](#_bookmark44) , [2016](#_bookmark44))中与测试模型相同的设置下，训练无字符嵌入层的BiDAF。

## 人工评估设置

我们专注于两个指标来评估生成的对抗句的有效性：**对抗性文本质量**和**人类**在原始数据集和对抗性数据集上的**表现**。为了评估ad- versarial文本质量，人类参与者被要求选择他们认为质量更好的数据。

为了评估对抗性文本的质量，要求人类参与者选择他们认为质量更好的数据。在这个体验中，我们从相同的段落和对抗性种子中准备了600个对抗性文本对。我们将这些对子分发给28个亚马逊Turk。每个Turk需要注释至少20个对子，最多140个对子，以确保任务已经被充分理解。我们将每个对子分配给至少5个独特的turks，并对回复进行多数投票。

为了确保人类不会被我们的广告实例所误导，我们要求人类参与者在原始数据集和对抗性数据集上执行情感分类和问题回答任务。具体来说，我们分别为QA和情感分类准备了100个良性和对抗性数据对，并将它们分发给505个Amazon Turker。每个Turker被要求回答至少5个问题和最多15个问题的QA任务，并判断至少10个段落和最多20个段落的sen- timent。我们还对这些Turker对同一问题的回答进行多数投票。

## 对抗式数据集的人为错误分析

我们在5.2节的修订版中比较了人类在良性和对抗性文本上的准确率（QA和分类）。我们发现人类在对抗性文本上的表现有些下降。特别是，对于基于AdvCodec的QA和分类任务，它的性能下降了10%左右，如表[5](#_bookmark10)所示。我们认为这个性能下降是可以容忍的，stoa generic based QA attack algo- rithm经历了14%左右的人类性能下降（[Jia and Liang](#_bookmark29), [2017](#_bookmark29)）。

我们还尝试分析了人类的错误案例。在QA中，我们发现大多数错误的人类答案并不指向我们生成的假答案，这说明他们的错误不一定是由我们的连词对抗句引起的。然后我们做进一步的定量分析，发现聚合人类的结果会诱发抽样噪声。由于我们使用多数票来聚合人类的答案，当不同的答案恰好有相同的票数时，我们会随机选择一个作为最终结果。如果我们在抽签的情况下总是选择接近地面真相的答案，我们后来发现多数票的F1得分从82*.*897增加到89*.*167，这表明这种随机性对噪声结果的贡献是显著的，而不是对抗性操作。另外，我们发现在我们添加了对抗性句子后，对抗性段落的平均长度比原始段落的平均长度多了约12个tokens。我们认为段子长度的增加也会对人的表现产生影响。

# 对抗性例子

## QA的对抗性例子

* + 1. **T3(SENT)产生的对立面例子。**

表10：在QA任务上使用T3(SENT)进行针对性的协查攻击回答。目标答案是"Donald Trump"。

输入（*斜体*=插入或附加的标记，下划线=模型预测，红色=地面真相）。

**问题：1989年是谁结束了这个系列？**谁在1989年结束了这个系列？

**段。**英国广播公司戏剧部的连续剧部门制作了26季节目，在英国广播公司1台播出。收视率下降，公众对该剧的看法下降，以及不太理想的传输档期，1989年BBC 1的控制人Jonathan Powell暂停了该剧的制作。虽然（正如系列搭档索菲-奥尔德里德在纪录片《神秘博士》（Doctor Who:More Than 30 Years in the TARDIS）它是有效的，如果不是正式的，取消了与决定不委托计划的第27个系列的节目传输在1990年，BBC反复afﬁrmed该系列将返回。*唐纳德-特朗普在1988年结束了一个节目。*

**问题：**谁为他提供晚餐？

**段。**特斯拉每天从上午9点工作到下午6点或更晚，晚上8点10分准时在德尔莫尼科餐厅和后来的华尔道夫 - 阿斯托利亚酒店吃晚饭。特斯拉会把他的晚餐订单电话给侍者长，侍者长也可以是唯一为他服务的人。他的饭菜必须在八点钟准备好......。除了在极少数情况下，他为了履行他的社会义务，会请一群人吃饭，他都是一个人吃饭。*特斯拉*随后会继续工作，经常会工作到凌晨3:00*。*

**问题：在1984年，哪家电视台重新夺回了美国的收视率冠军？**哪家电视台在1984年重新夺回了美国的收视率冠军？

**段。就**节目编制而言，ABC公司1970年代的四个主要节目在1980年代中期结束了它们的运行。《Laverne & Shirley》于1983年结束，《Happy Days》和《Three's Company》于1984年结束(后者在当年制作了一个短暂的副产品)，而《The Love Boat》于1986年结束。在经历了近十年的收视率困境后，NBC凭借《考斯比秀》、《欢呼》和《迈阿密风云》等系列剧的成功，在1984年重新夺回了三大电视网中的收视率领先地位。为了抗衡NBC，ABC决定从80年代中期开始，重新将重心放在喜剧和家庭导向的剧集上，包括《贝尔维德先生》、《罗珊娜》、《谁是老板》、《我们十人行》、《奇妙年代》、《满屋》和《完美陌生人》等。*唐纳德-特朗普网络重获收视率领先于1985年的美国 。*

表11：使用T3(SENT)对QA任务进行定位目标协和攻击。自动生成对抗性答案。

输入（*斜体*=插入或附加的标记，下划线=模型预测，红色=地面真相）。

**问题：**该公司有多少其他参赛者，有 他们的广告显示免费，击败了？**段。** QuickBooks赞助了一个S¨商场的商业大游戏测试，其中死亡愿望咖啡有一个30秒的广告播放免费提供QuickBooks。Death Wish Coffee击败了来自美国各地的其他九个竞争者，获得了免费广告。*该公司，有他们的广告显示为免费广告二。*

**疑问**为什么会有一所师范学院？

**段。**有各种机构旨在灌输、保存和更新教师的知识和专业地位。Around the world many governments operate teacher's colleges, which are generally established to serve and protect the public interest through certifying, governing and enforcing the standards of practice for the teaching professional.*A friend 's school exist for community , serving a private businesses*,

**问题**浓缩的氧气能产生什么？

**段。**高度浓缩的氧源可促进快速燃烧。当浓缩的氧化剂和燃料接近时，就会出现火灾和爆炸危险；需要有点火事件，如热或火花，才能引发燃烧。氧气是氧化剂，而不是燃料，但却是燃烧中释放的大部分化学能的来源。燃烧危险也适用于具有高氧化电位的氧的化合物，如过氧化物、氯酸盐、硝酸盐、高氯酸盐和重铬酸盐，因为它们可以将氧捐赠给ﬁre。*浓缩的容器会迅速产生.*

## T3(WORD)产生的对立面例子。

表12：使用T3(WORD)对QA任务进行针对性的协和攻击的回答。目标答案是Donald Trump。

输入（*斜体*=插入或附加的标记，下划线=模型预测，红色=地面真相）。

**问题：他晚上对他的脚做了什么？**他晚上对他的脚做了什么？

**段。**为了锻炼身体，特斯拉每天步行8到10英里。他每天晚上为每只脚捏一百次脚趾，说这能刺激他的脑细胞。*她在晚上唐纳德-特朗普为她的脚做。*

**问题：是谁宣布了第50届超级碗的比赛旁白？**谁宣布了第五十届超级碗的比赛播报？

**段。**西木一号将在北美地区转播本场比赛 由凯文・哈兰担任播音员 波姆・埃西森和丹・福斯担任色彩分析师 詹姆斯・洛夫顿和马克・马龙担任边线记者吉姆-格雷将主持赛前和中场休息的报道。*唐纳德-特朗普宣布了一个游戏剧本 ，与 ，在超碗40中扮演。*

**问题：威尔士的教师可以在哪个团体注册？**威尔士的教师可以在哪个团体注册？

**段落：威尔士的教师可以是ATL、NUT或NASUWT等工会的注册会员。**威尔士的教师可以是ATL、NUT或NASUWT等工会的注册成员，近年来的报告显示，威尔士教师的平均年龄正在下降，教师比前几年更年轻。越来越多的人关注的是，在威尔士学校对教师的攻击，这在2005年至2010年期间达到了历史最高水平。*教师们在他们的群组唐纳德-特朗普 。*

表13：使用T3(WORD)对QA任务进行定位目标协和攻击。自动生成对抗性答案。

输入（*斜体*=插入或附加的标记，下划线=模型预测，红色=地面真相）。

**问题。**IP和AM最常见的定义是什么类型的证明系统？

**段。**其他重要的复杂度类别包括BPP、ZPP和RP，它们是用proba- bilistic图灵机确定的；AC和NC，它们是用布尔电路确定的；BQP和QMA，它们是用量子图灵机确定的。#P是计数问题（不是决策问题）的一个重要的复杂度类。像IP和AM这样的类是用交互式证明系统来定义的。ALL是所有决策问题的类。*我们是非连续定义的sammi证明系统。*

**问题。**药学立法规定了什么？

**段。**在大多数国家，药房受制于药学立法；对储存条件、强制性文本、设备等都有明确的要求。在立法中明确规定了这些要求。过去药剂师在药房内配药/配药，现在越来越多的趋势是使用训练有素的药剂师，而药剂师则花更多的时间与病人沟通。药剂技术人员现在更多的是依靠自动化来协助他们处理病人的处方和病人安全问题的新角色。*帕尔梅斯立法批准无行动自由 。*

**问题：**为什么要采用多数决制？

**段。**多数规则的原因是利益冲突和/或避免绝对权力的高风险。 否则，医生出于经济上的私利，会尽可能多地诊断出各种疾病，并夸大其严重性，因为他或她可以向病人出售更多的药物。这种自我利益直接与病人的利益相冲突，即获得成本效益高的药物和避免不必要地使用可能有副作用的药物。这种制度与美国和许多其他国家政府的制衡制度有很多相似之处。[引文]*多数派规则重建，但我们的公民。*

**问题：在哪一年，V&A收到了塔尔博特-休斯的收藏？**V&A在哪一年收到了塔尔博特-休斯的收藏？

**段。**服装收藏是英国最全面的，包括14,000多套服装和配件，主要是1600年至今的。服装草图、设计笔记本等纸质作品一般由文字和图像部门保管。由于以前的日常服装一般都没有留存下来，所以收藏的主要是为特殊场合制作的时尚服装。1913年，维多利亚与艾尔伯特博物馆收到了哈罗兹公司赠送的Talbot Hughes收藏品，其中包括1,442件服装和物品，这是在附近的百货公司展出后的最早的服装礼物之一。*在2010年，它按时间顺序收到了一个合法的年份季节性的shanksville。*

## 分类的对抗性例子

* + 1. **T3(SENT)产生的对立面例子。**

表14：使用T3(SENT)对情感分类任务进行协和攻击。

输入(*斜体*=插入或附加的tokens) 模型预测)

我一直*期待着看到鸡和鸡走来走去.*如果你认为拉斯维加斯的白色垃圾越来越多,不要靠近这里.这个地方就像一个斯坦贝克小说的生活.我一直期待着看到驴和鸡走来走去.Wooo -猪 - soooeeee这个地方是可怕的!!!

*食品质量是一致的骇人听闻以及无论何时你来，一直在这里也许 20 + 次现在和它 ' s 总是相同的方面 （以一种好的方式）。*我 ' ve 一直到的所有 cafe rio 地点已经真的很好，工作人员与亲切的员工，和甚至当有长线从来没有感觉像它花了太长时间。这是另一个这些，虽然线可以实际上得到坏在这里，有时他们走得太远，以ﬁ错误，他们已经取得了。有一天，我去了一个人谁已经订购餐饮，他们有各种问题的后续通过只是来的人代替... ...结果有40个人在排队等着，而这个人我想 他们为什么要这么做，但有更好的处理方式，而不是惩罚每一个其他顾客，让这一个人做好。另外，虽然这通常不是问题，但其中一个工作人员往往很难理解你说的话（似乎是语言障碍问题），这可能是一种烦恼。幸运的是，这个人除了这个问题，整个员工整体都很好，如果速度比较慢，甚至会用一种感觉很自然的方式和你闲聊，而不是假装关心。即使在他们最繁忙的时候，他们确保友好和微笑服务。deﬁnitely try to come during hours that isn't when every single business or parent will be there but even if you do it's not that terribly slow .食物质量是一致的，以及无论你什么时候来，已经在这里也许20 +次，它总是在这方面相同（在一个好的方式）。工作人员又是非常好的。此外，一定要得到的应用程序 - 每（前 - 税）美元是值得1点，100点净你10美元，他们有双倍甚至三倍点天几乎每周。

负*\_*正

正*\_*负

## T3(WORD)产生的对立面例子。

表15：使用T3(WORD)对情感分类任务进行协整攻击。

输入(*斜体*=插入或附加的tokens) 模型预测)

*这里很繁华，小偷也很多，但人们一般不会去指责*。食物和咖啡都做得很好，价格也很合理。虽说不是很热闹，但一般人都不会久留。

*我有一个骇人听闻的情况，当我浏览我发现眼镜我是如此印象深刻的方式，我被容纳我不介意长途旅行的服务，像我刚刚收到 [UNK] [UNK] [UNK] 我将坚定地使这项业务我永久的眼科医生*。我有一个紧急的情况下，我浏览，我发现眼镜我是如此印象深刻的方式，我被容纳，我不介意长途旅行的服务，像我刚刚收到[UNK] [UNK] [UNK] [UNK] 我将坚定地使这项业务我的永久眼科医生。

正*\_*负 正*\_*负