可信机器学习论文报告

学号：10185102253　姓名：黄宇辰

论文主题

对于语言模型在有语法错误时的鲁棒性研究

On the Robustness of Language Encoders against Grammatical Errors

论文来源

作者信息：

Fan Yin[1], Quanyu Long[2], Tao Meng[3], and Kai-Wei Chang[3]
[1]Peking University
[2]Shanghai Jiao Tong University
[3]University of California, Los Angeles
1600012975@pku.edu.cn;
oscar.long@sjtu.edu.cn;
tmeng@cs.ucla.edu;
kwchang@cs.ucla.edu

会议出处：ACL2020

论文链接：https://www.aclweb.org/anthology/2020.acl-main.310/

一级方向：对抗防御与鲁棒性

二级方向：NLP

摘要阅读

这篇论文的作者们进行了一项关于预训练语言模型遇到自然语法错误时的行文。

预训练语言模型这里作者特殊强调了ELMo，BERT 和 RoBERTa 3种预训练语言模型。

作者们在研究时，从一些非母语使用者那里收集语法错误，之后在一些没有错误的干净文本上模拟这些错误作为对抗性攻击。在使用这种方法之后，论文的作者们发现，所有的实验模型的性能都受到了影响，但是影响的程度并不相同。

为了进一步解释语言模型在遇到错误语法时候这些行为，作者们进一步设计了一个语言可接受性的任务去了解这些语言模型识别有语法错误的句子和找到语法错误位置的能力。他们发现在一个拥有在对于正确句子预测上训练出来的分类器的基于上下文语言模型可以去定位语法错误。

作者们还设计了一个完形填空去测试BERT预训练模型，在这个测试中，他们发现了BERT模型在找到错误与上下文特定的某种标记之间的关系特征。

这篇论文的研究结果有助于明了语言模型与语法错误之间的鲁棒性和行为。

引言阅读

过去预训练语言模型在NLP任务上取得了巨大成功。但是过去人们在预训练语言模型的构建与使用上，通常都假设测试集和样本集都是干净的，即没有语法错误。但是显然这是不可能的，尤其对于非母语使用者来说，总是会有各种各样的语法错误。语法错误是自然语言处理任务中的重要噪声，对于这种噪声的处理，反映了语言模型的鲁棒性。如果在对非母语使用者的文本预料进行处理时，语言模型同样具有相当好的效果，那么这是非常棒的。我认为这也是这篇论文研究的目的与重点。

近年来，有许多研究者在对于语法错误模型的行为的评价上使用了许多不同的研究方法。主要有如下三类：

1- 依靠专家手动对于特定的语言现象进行最小编辑化标注

2- 依靠某些已有特征标签进行评价或者创造出可行的评价指标

3- 在处理诸如情感分类、机器翻译的自然语言任务时，构建模型时模拟噪声

以上已有的方法或多或少有它自己的缺点：大量依赖人类手动标注；需要非常专业的语言学知识；只关注特定的语言现象；主要在指定的语料库上进行相关实验。

相比之下，这篇论文的研究者们，实现了自动模拟自然语言数据和自然语言中可能出现的各种语法错误，并分析语言模型处理自然语言任务时噪声的影响。研究者们的成果对于研究语言模型在应对语法错误时的鲁棒性具有相当实际的意义。

更加具体的解释他们的方案，可以这么去理解。

他们首先提出了一个有效的方法去模拟各种各样的语法错误。这种方法应用了基于在NUCLE上观察到的真实错误形成的黑盒对抗攻击算法，它可以将干净的语料库转换成有许多语法错误的未清洗语料库，这样就方便了在底层任务上调试语言模型。他们在四种语言理解任务上和一种序列标记任务上进行了评估，并证明了他们方案的适用性。

下一步，论文作者们通过语言接受能力任务去探究各个预训练模型，并量化各个模型识别语法错误的能力。他们为八种语法错误类型分别建立了单独的数据集。之后，他们确定一个模型，为每一层都添加一个简单分类器去预测文本的正确性并定位错误位置。这个研究任务中，假定如果这一层的简单分类器在预测某种类型的错误上表现很好，那么这个模型的这一层就很有可能拥有对于该类型的错误的识别能力。

*八种语法错误类型如下

Error type	Error Description	Confusion Set
ArtOrDet	Article/determiner errors	{ a, an, the, ø}
Prep	Preposition errors	{ on, in, at, from, for, under, over, with, into, during, until, against, among, throughout, to, by, about, like, before, across, behind, but, out, up, after, since, down, off, of, ø}
Trans	Link words/phrase errors	{and, but, so, however, as, that, thus, also, because, therefore, if, although, which, where, moreover, besides, of, ø}
Nn	Noun number errors	{SG, PL}
SVA	Subject-verb agreement errors	{3SG, not 3SG}
Vform	Verb form errors	{Present, Past, Progressive, Perfect}
Wchoice	Word choice errors	{Ten synonyms from WordNet Synsets}
Worder	Word positions errors	{Adverb w/ Adjective, Participle, Modal}

最后，他们研究了模型如何捕获语法错误和错误文本之间的联系。他们以BERT预训练模型为例，设计了一个无监督完形填空测试来评估其作为MLM的性能。

简而言之，他们这篇论文的贡献有如下3点：

1. 他们提出了一种模拟各种语法错误的新方法。该方法灵活，可用于验证语言模型对语法错误的鲁棒性。
2. 他们对各类语言模型的鲁棒性进行了系统分析，并通过研究具有各种语法错误类型的底层任务模型的性能，加强了之前的工作。
3. 他们证明了：

   (1)现有语言模型对语法错误的鲁棒性是不同的;

   (2)基于上下文的语言模型比标记嵌入的语言模型具有更强的语法错误识别和定位能力;

   (3)BERT预处理模型验证了上下文中错误和特定标记之间的相互作用，特别是相邻的错误标记。

问题回答

Q1: 论文试图解决什么问题?

因为现有的自然语言处理任务大都在清洗后无语法错误的语料库上进行，而一旦语料中出现了语法错误就会使自然语言处理的效果变差。

而自然语言处理目前在翻译方面应用较广，对于非母语使用者来说，出现语法错误是十分常见的。

这篇论文在这个基础上，研究自然语言任务中语言模型对于语法错误的鲁棒性，同时，还试图解决过去对于语法错误模型的行为的评价方法的缺点。

Q2: 论文中提到的解决方案之关键是什么?

过去对于语法错误模型的行为的评价方法主要有：1）依靠专家手动对于特定的语言现象进行标注；2）依靠某些已有特征标签进行评价或者创造出可行的评价指标；3）构建模型时手动模拟噪声。

这些过去的评价方法有许多缺点：大量依赖人类手动标注；需要非常专业的语言学知识；只关注特定的语言现象；主要在指定的语料库上进行相关实验。

所以论文中提到了这样的方法：应用了基于在NUCLE上观察到的真实错误形成的黑盒对抗攻击算法，它可以将干净的语料库转换成有许多语法错误的未清洗语料库。

Q3: 论文中的实验是如何设计的?

论文中，在语言模型的每个层级上都加上简单的分类器。对于某一层级来说，如果分类器对某一语言错误特征表现出来的分类效果很好，那么这一层级对该类型的语言错误特征就具有比较好的识别效果。

Q4: 用于定量评估的数据集是什么？代码有没有开源?

数据集使用的是GLUE数据集，代码在Github上开源

代码链接：https://github.com/uclanlp/ProbeGrammarRobustness

Q5： 论文中的实验及结果有没有很好地支持需要验证的科学假设?

实验结果如下：

	Infersent	ELMo	BERT	RoBERTa
MRPC	75.42	80.30	86.48	89.88
MNLI-m	68.62	74.91	83.77	87.70
MNLI-mm	69.12	75.50	84.80	87.40
QNLI	77.39	78.23	90.58	92.50
SST-2	83.14	90.37	92.08	94.72
NER	-	91.21	95.20	95.45

Table 2: Original performance of the target models on language understanding and sequential tagging tasks.

Model	Alg.	MRPC	MNLI (m/mm)	QNLI	SST-2	NER
Infersent	dist.	6.51 (14.53)	8.30 (13.98) / 8.80 (14.23)	4.76 (12.53)	5.79 (14.38)	-
	greedy	53.42 (9.02)	36.52 (10.35) / 40.71 (10.06)	44.92 (7.61)	43.44 (8.02)	-
	beam	54.39 (9.08)	36.66 (10.37) / 40.87 (10.06)	45.16 (7.62)	43.86 (8.03)	-
	genetic	79.15 (8.60)	-	-	59.86 (8.39)	-
BiLSTM	dist.	9.99 (14.53)	7.76 (13.98) / 7.83 (14.23)	5.34 (12.53)	4.64 (14.38)	3.29 (13.75)
+ ELMo	greedy	60.84 (8.19)	29.58 (10.28) / 32.92 (9.89)	39.12 (7.25)	37.55 (8.24)	17.81 (7.67)
+ Attn	beam	61.49 (8.29)	29.74 (10.29) / 33.12 (9.91)	40.38 (7.33)	38.32 (8.32)	18.33 (7.85)
	genetic	81.14 (7.41)	-	-	59.25 (8.25)	39.78 (8.19)
BERT	dist.	3.69(14.53)	6.59 (13.98) / 6.95 (14.23)	2.33 (12.53)	4.73 (14.38)	3.07 (13.75)
	greedy	31.25 (7.95)	28.76 (10.28) / 32.04 (10.01)	25.43 (7.38)	33.54 (7.96)	17.12 (7.51)
	beam	31.81 (8.01)	29.03 (10.30) / 32.44 (10.04)	26.42 (7.48)	34.28 (8.01)	18.27 (7.74)
	genetic	59.01 (8.84)	-	-	58.53 (7.83)	38.83(7.64)
RoBERTa	dist.	3.04 (14.53)	5.66 (13.98) / 5.88(14.23)	1.92 (12.53)	3.53 (14.38)	2.52 (13.75)
	greedy	20.45 (8.11)	20.65 (10.43) / 21.47 (10.02)	19.82 (7.18)	31.06 (8.20)	15.84 (8.12)
	beam	20.73(8.14)	20.89 (10.44) / 21.91 (10.06)	20.52 (7.29)	31.91 (8.27)	16.51 (7.47)
	genetic	38.93 (9.17)	-	-	56.41 (8.39)	35.11(7.55)

Table 3: Results of evaluating the robustness of models on downstream tasks. Each column represents a dataset and each row represents a victim model with the attack algorithm (dist. means *probabilistic transformation*). In each cell, we show the mean attack success rate (in percentage) and the mean percentage of modified words (number in the bracket) over the dataset.

	BERT			RoBERTa		
	MRPC	MNLI	SST	MRPC	MNLI	SST
Prep	16	178	36	15	103	43
Art/Det	5	270	20	7	228	28
Wchoice	93	1129	233	64	772	195
Vform	8	231	26	9	314	37
SVA	57	538	83	31	388	83
Nn	14	128	13	3	84	13
Worder	0	62	28	0	43	28
Trans	5	70	25	5	31	25

Table 4: Numbers of times each error type is chosen in successful attacks. We find that `Wchoice` and `SVA` are more harmful.

以上数据均表明，实验与实验结果支持并验证科学假设。作者们还从实验结果中得到了更多的研究内容。

Q6： 这篇论文到底有什么贡献?

1. 他们提出了一种模拟各种语法错误的新方法。该方法灵活，可用于验证语言模型对语法错误的鲁棒性。
2. 他们对各类语言模型的鲁棒性进行了系统分析，并通过研究具有各种语法错误类型的底层任务模型的性能，加强了之前的工作。
3. 他们证明了：

   (1)现有语言模型对语法错误的鲁棒性是不同的;

   (2)基于上下文的语言模型比标记嵌入的语言模型具有更强的语法错误识别和定位能力;

   (3)BERT预处理模型验证了上下文中错误和特定标记之间的相互作用，特别是相邻的错误标记。

Q7： 下一步呢？有什么工作可以继续深入?

我认为基于这篇文章，我们认识到了语言模型在不同层级对于不同语法错误的识别能力是不同的。所以我们可以考虑在不同层级进行模型的改良，从而针对性地提高语言模型各个层级对于某种语法错误的识别能力，使语言模型在自然语言处理任务中的鲁棒性得到提高。