面向中文实体识别的Transformers模型句子级非对抗鲁棒性研究

王立帮　王裴岩*　沈思嘉
沈阳航空航天大学计算机学院，辽宁沈阳 110136
wangpy@sau.edu.cn
{wanglibang, shensijia}@stu.sau.edu.cn

摘要

基于Transformers的中文实体识别模型在标准实体识别基准测试中取得了卓越性能，其鲁棒性研究也受到了广泛关注。当前，中文实体识别模型在实际部署中所面临的句子级非对抗鲁棒性问题研究不足，该文针对该问题开展了研究。首先，该文从理论上分析并发现了Transformer中自注意力、相对位置嵌入及绝对位置嵌入对模型鲁棒性的负面影响。之后，提出了实体标签增强和滑动窗口约束的鲁棒性增强方法，并从理论上证明了提出方法能够提升Transformers模型的实体识别鲁棒性。最后，通过在3个中文数据集的实验，研究了4种基于Transformer的实体识别模型的脆弱性，所提出方法使模型的鲁棒性F1值提升最高可达4.95%。

关键词： 实体识别；非对抗鲁棒性；Transformers模型；句子级扰动

On Sentence-level Non-adversarial Robustness of Chinese Named Entity Recognition with Transformers Model

Libang Wang　Peiyan Wang　Sijia Shen
School of Computer Science, Shenyang Aerospace University,
Shenyang, Liaoning 110136, China
wangpy@sau.edu.cn
{wanglibang, shensijia}@stu.sau.edu.cn

Abstract

Chinese entity recognition models based on Transformers have shown exceptional performance in standard benchmarks, with robustness studies in this area gaining extensive attention. However, research on sentence-level non-adversarial robustness issues faced by Chinese entity recognition models in practical deployment is insufficient. This study addresses this gap by first theoretically analyzing the negative impacts of the Transformer model's self-attention, relative position embeddings, and absolute position embeddings on model robustness. It then proposes robustness enhancement methods through entity label augmentation and sliding window constraints, theoretically proving these methods can improve the Transformers model's entity recognition robustness. Experimental results on three Chinese datasets reveal vulnerabilities in four Transformer-based entity recognition models, with the proposed methods increasing model robustness F1 scores by up to 4.95%.

* 通讯作者

基金项目：辽宁省应用基础研究(2022JH2/101300248)，国家自然科学基金资助项目(U1908216)

Keywords: entity recognition , non-adversarial robustness , transformers model , sentence-level perturbation

1 引言

命名实体识别（Named Entity Recognition，NER）指从文本中识别出具有特定意义的实体，并将其分类为预定义类别，如人名、地名、组织和机构等。其是自然语言处理（Natural Language Processing，NLP）领域中的一项基础任务，在信息检索 (Tan et al., 2021)、问答系统 (Min et al., 2021)、对话系统 (Wang et al., 2020)等多个领域中具有广泛应用。

目前，以Transformer为典型代表的神经网络架构模型在NLP领域的多项任务中均取得了卓越性能 (Vaswani et al., 2017; Kenton and Toutanova, 2019)，但此类模型在应对扰动输入时的脆弱性也引发了研究者对其安全性的思考。最近一些研究 (Liang et al., 2018; Wu et al., 2023; Zhu et al., 2023)表明，此类模型容易受到扰动示例影响。通过扰动输入序列来干扰模型，最终使其预测结果发生改变。

NER作为研究和工业应用中最常见的NLP任务之一，近年来随着深度学习兴起，其也取得了长足发展。特别是，Transformer架构的提出为此带来了新动力。借助Transformer强大的长序列建模与并行计算能力，NER模型在识别性能和运行效率上都取得了显著提升 (Zhang and Yang, 2018; Li et al., 2020)。同时，基于深层Transformer构建的预训练模型结合微调或提示范式的广泛应用也为NER研究带来新突破 (Hu et al., 2022)。

NER作为NLP领域中的一个基础任务，其对构建更智能、更精准的NLP系统具有关键作用，因此NER模型的鲁棒性问题也引起了研究者的关注。现有研究 (Srinivasan and Vajjala, 2023)证明NER模型对输入数据的微小变化很敏感。表 1展示了一个扰动实例，在所列出的扰动序列输出中，模型预测输出的“深圳特区”应被标记为地缘政治（GPE）实体。然而，正如扰动后的输出所示，其却被标记为组织（ORG）实体。显然，一些看似无害的微小扰动正在改变模型的预测结果。

类型	实例
原始序列1	香港$_{GPE}$ 回归后，深$_{GPE}$ 港$_{GPE}$ 两地经贸关系得到加强，深$_{GPE}$ 港$_{GPE}$ 西部$_{LOC}$ 通道等三大基础设施衔接项目目前进展顺利。
预测输出	香港$_{GPE}$ 回归后，深港$_{GPE}$ 两地经贸关系得到加强，深港$_{GPE}$ 西部通道等三大基础设施衔接项目目前进展顺利。
原始序列2	深圳$_{GPE}$ 特区$_{LOC}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。
预测输出	深圳特区$_{GPE}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。
扰动序列	原始序列1 + 原始序列2
预测输出	香港$_{GPE}$ 回归后，深港$_{GPE}$ 两地经贸关系得到加强，深港$_{GPE}$ 西部通道等三大基础设施衔接项目目前进展顺利。深圳特区$_{ORG}$ 今年推进国有企业改革成效显著，全市今年工业总产值预计达一千二百五十五亿元，比去年增长百分之十七点五。

Table 1: 对OntoNote4.0中序列进行扰动后BERT-CRF模型预测实例

目前，对于NER模型鲁棒性的研究多集中在英文环境下 (Namysl et al., 2020; Lin et al., 2021; Das and Paik, 2022)，其通过多种方法构造不同扰动样本对模型鲁棒性展开评估，其中多数方法也取得了很高的攻击成功率，最新研究将此推广到了多种语言环境下 (Srinivasan and Vajjala, 2023)。但Wang等人 (2022a)表明，现有模型鲁棒性研究中所构造的扰动样本多是非自然的，这些样本在现实场景中并不会出现，同时这类样本大多是不流畅的。因此，根据这些非自然扰动样本来判断模型鲁棒性是不合理的。

基于上述分析，在中文NER任务中，结合目前中文NER模型在实际部署中所面临的句子级自然扰动现象，本文开展对中文NER模型在句子级自然扰动下鲁棒性的研究。同时本文遵循计算机视觉中Drenkow等人 (2021)对模型鲁棒性的进一步划分，将这类在句子级自然扰动下模型鲁棒性称作模型的句子级非对抗鲁棒性，同时提出两个主要问题：

- 中文NER下基于Transformer构建的模型是否会受到句子级自然扰动样本的影响?
- 如果此类NER模型在该类自然扰动影响下不鲁棒，应如何提高模型句子级非对抗鲁棒性?

针对问题1，本文遵循NER模型在实际应用中所出现的句子级自然扰动场景，在多个中文NER数据集上构建句子级鲁棒性评估数据集以评估模型鲁棒性；针对问题2，本文引入两种约束策略来改善NER模型的句子级非对抗鲁棒性。在这项工作中，本文通过一系列实验对中文NER模型进行鲁棒性评估，最终得到如下发现：

- 与LSTM相比，使用Transformer或预训练语言模型结合微调或提示学习所构建的NER模型更容易受到句子级自然扰动影响。
- Transformer中的位置嵌入和自注意力模块削弱了句子级自然扰动下模型鲁棒性。其中绝对位置嵌入模块对模型鲁棒性的影响独立于自注意力模块，而相对位置嵌入模块则依赖自注意力模块，只有自注意力模块发挥作用时其才会对模型鲁棒性产生影响。
- 在训练阶段增强模型对已知实体的关注度可提高模型的句子级非对抗鲁棒性，但这种策略提升效果有限。
- 在模型中引入滑动窗口注意力机制可以改善句子级自然扰动下模型的鲁棒性。

2 相关工作

2.1 模型微调

BERT (Kenton and Toutanova, 2019)首次证明了在大规模预训练语言模型的基础上，通过微调可以在各种任务上实现卓越的性能。通过在大规模通用语料库上进行无监督预训练，然后在特定任务或领域的有监督微调中逐步调整模型参数，以实现更高级别的自然语言理解和生成，从而为各种文本处理任务提供强大的通用基础。当前，由大语料训练的预训练模型在中文序列标注方面表现出色，使用BERT字符特征的模型在中文NER和中文词性标注等方面的表现也远远优于基于静态嵌入的方法 (Hu et al., 2022)。

2.2 提示学习

提示学习是一种有监督学习方法，旨在减少对庞大标注数据集的依赖。与预训练加微调范式不同，提示学习将下游任务转换为与模型预训练任务更一致的形式。基于提示学习的模型通过应用提示功能将输入序列转化为具有未填充槽的各种提示模板，通过填充这些槽，可以完成各种NLP任务，例如推荐系统 (Wang et al., 2022b)、情感分析 (Li et al., 2023)、事件抽取 (Ma et al., 2022b)等。同时，该方法的发展也使NER下零样本和少样本学习成为可能 (Ma et al., 2022a)。本文主要贡献在于探讨了在受到句子级自然扰动时，基于提示学习的中文NER模型鲁棒性所受影响，扩展了当前的研究。

2.3 模型鲁棒性

鲁棒性是深度学习模型的一个重要属性，其是指模型在面对与训练数据不同的输入时依然能够产生预期输出的能力。在计算机视觉中，Drenkow等人 (2021)对模型鲁棒性做了进一步划分，包括对抗鲁棒性和非对抗鲁棒性，其中非对抗鲁棒性也被称为自然鲁棒性。对抗鲁棒性衡量了模型对经过故意破坏的输入数据进行准确分类的能力，其目标是通过欺骗模型使其以高置信度输出错误的预测结果。这种类型的鲁棒性取决于攻击者对模型和数据属性的了解程度。

NLP下模型对抗鲁棒性研究通常采用多种形式对模型进行干扰，以改变模型的输出结果。Zhu等人 (2023)为评估大语言模型（Large Language Models，LLMs）对对抗性提示的鲁棒性，推出了一个系统基准，称之为PromptBench，旨在评估、理解和分析LLMs的鲁棒性。在多个数据集上的实验结果显示，不同类型的攻击在效果上存在显著差异。其中单词级攻击威胁最大，可致所有数据集的平均性能下降33%，字符级攻击次之，句子级攻击的威胁最低。Wang等人 (2023)对ChatGPT在NLP任务中的对抗鲁棒性和分布外泛化能力进行了深入评估，结果表明，以ChatGPT为代表的LLMs在鲁棒性和分布外泛化能力上较以往模型有显著提升，但与人类水平相比，仍有一定差距。

NER下模型对抗鲁棒性研究涵盖了对抗性攻击的多个粒度，包括字符级、单词级和句子级攻击。Namysl等人 (2020)、Lin等人 (2021)、Das等人 (2022)及Srinivasan等人 (2023)在文本中随机翻转输入序列中的字符或单词、随机单词序列或者添加分散模型注意力的句子来混淆模型，使模型预测出错。尽管对抗性鲁棒性可以在模型面对恶意攻击时保持模型的稳健性，但非对抗性鲁棒性能够在自然引发的数据扰动场景下保持模型性能和预测结果一致性。因此，非对抗鲁棒性可被视为一个理想的模型属性，在模型实际部署中，自然扰动下的数据最为常见。

3 实体识别模型的鲁棒性评估

3.1 问题定义

本文将NER下的句子级鲁棒性问题定义为模型在受到句子级扰动时所出现的预测结果不一致现象。即对任意原始序列$\mathbf{X}^o = \{c_i\}_{i=1}^p$，使用自然扰动样本生成算法$A$可构造其扰动序列$\mathbf{X}^r = A(\mathbf{X}^o)$。该扰动序列中某个字符输入已训练好的模型，模型将以较高置信度输出与原始标签不同的结果，如式(1)所示：

$$c^o \in \mathbf{X}^o,\ c^r \in \mathbf{X}^r;\ \exists c^o = c^r,\ f(c^o) \neq f(c^r) \tag{1}$$

同时，本文将模型在受到句子级自然扰动时的扰动成功率Asr定义为：

$$Asr = \frac{1}{l}\sum_i \mathbb{I}(f(c_i^o) \neq f(c_i^r)) \tag{2}$$

式中，l为序列$\mathbf{X}^o$的长度，$\mathbb{I}(x)$为指示函数，c_i^o和c_i^r分别为原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中索引为i处的字符。

对任意一段输入序列$\mathbf{X} = \{c_i\}_{i=1}^n$，各字符经过词嵌入后的输入表征如式(3)所示：

$$\boldsymbol{x}_1, \boldsymbol{x}_2, \cdots, \boldsymbol{x}_n = Embedding(c_1, c_2, \cdots, c_n) \tag{3}$$

式中，$Embedding()$可为静态词嵌入如Glove (Pennington et al., 2014)等，亦或动态词嵌入如BERT (Kenton and Toutanova, 2019)等。

在句子级自然扰动下，Transformer模型中输入向量$\boldsymbol{x}$在经过自注意力模块后的输出向量$\boldsymbol{t}$被改变，最终干扰了模型的预测结果。假设原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中字符$\boldsymbol{c}$在经过自注意力模块后的输出向量分别为$\boldsymbol{t}^o$和$\boldsymbol{t}^r$，若$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r)$低于某个阈值时会改变模型对于当前字符的预测结果，可列出相应表达式如式(4)所示：

$$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r) = \frac{\boldsymbol{t}^o \cdot \boldsymbol{t}^r}{\|\boldsymbol{t}^o\| \|\boldsymbol{t}^r\|} < 1 - \varepsilon \tag{4}$$

式中，ε用于衡量扰动的大小。

此时，结合式(4)可将式(2)中原始扰动成功率Asr调整为式(5)来衡量Transformer中在扰动样本生成算法A影响下的扰动成功率：

$$Asr = \frac{1}{l}\sum_i \mathbb{I}(\cos(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r) < 1 - \varepsilon) \tag{5}$$

3.2 鲁棒性评估数据集的构建

遵循AdvGLUE (Wang et al., 2021)中设计的扰动样本生成原则，并结合中文NER模型在实际应用场景中最常见的句子级扰动情况，本文使用基于分心的句子级扰动方法，扰动标准数据集中的原始序列来生成鲁棒性评估数据集。具体来说，通过拼接测试集中相邻两段序列$\mathbf{X}^{o_1}$和$\mathbf{X}^{o_2}$来生成句子级扰动序列$\mathbf{X}^r$，即$\mathbf{X}^r = A(\mathbf{X}^{o_1}) = [\mathbf{X}^{o_1}; \mathbf{X}^{o_2}]$，该扰动不会改变原始序列的语义。

3.3 评价指标

针对模型性能和鲁棒性的评估，本文在多种数据集上评估了实体级微平均F1（Micro-F1）分数，以比较不同模型的性能。目前NER任务下模型鲁棒性评估主要报告了两种F1变体的实验结果，以量化模型在标准数据集和鲁棒性评估数据集上的性能，分别是标准F1和扰动F1。

- **标准F1 (Standard-F1，S-F1)**：在标准数据集上评估模型所得的F1分数，其综合了模型的精确性和完整性，这是NER任务中使用最广泛的指标。
- **扰动F1 (Perturbation-F1，P-F1)**：在所构造的鲁棒性评估数据集上评估模型所得的F1分数。P-F1越大表明模型在鲁棒性评估数据集上的预测准确性越高。

在评估NER模型鲁棒性时，上述两种评价指标存在局限。如表 1所展示的模型鲁棒性评估实例，对于原始序列1和2的输出结果和对应的扰动输入序列，若仅计算其S-F1和P-F1，不难看出最终F1值是相等的，也即为此时在这类自然扰动影响下模型是鲁棒的。然而，仔细观察这三段序列各自预测结果，可以看出原始序列2中“深圳特区”的预测标签和扰动序列中“深圳特区”的预测标签是不一致的，即在这类自然扰动影响下模型仍是不鲁棒的。

鉴于此，本文参考Huang等人 (2021)使用标准精度、扰动精度、鲁棒精度来评估语义解析器的性能，在当前任务下，本文额外报告了一种F1变体的实验结果，称之为鲁棒F1。

- **鲁棒F1 (Robust-F1，R-F1)**：综合标准和鲁棒性评估两个数据集下模型的评估结果，以标准数据集下对应字符的预测标签作为鲁棒性评估数据集下该字符的真实标签，来评估模型在扰动数据集下的F1分数。R-F1越大表明模型在标准数据集和鲁棒性评估数据集上的预测结果的差异性越小，也即为模型在这两个数据集上预测结果的一致性越高。

3.4 自注意力和位置嵌入模块影响分析

Transformer作为一种基于自注意力机制的神经网络模型，其能够对序列中的每个字符进行全局建模，并在各字符间建立联系。与循环神经网络模型相比，Transformer具有更好的并行性能和更短的训练时间。Transformer的核心之一是自注意力机制，其作用是为输入序列中各字符间计算一个权重，然后将这些加权后的输入向量作为输出。

如下针对NER任务对Transformer架构作简单介绍，本文仅讨论其中的编码器部分。其由自注意力层和前馈网络层（Feed-Forward Networks，FFN）组成，每个子层后都有残差连接和层归一化，其中FFN是具有非线性变换的多层感知器。Transformer通过H个注意力头对序列$\mathbf{X}$进行自注意力，然后拼接H个头的输出得到最终结果。为简化计算过程，如下计算将忽略自注意力模块中注意力头数的影响，同时假设所有原始序列的长度都为l。在自注意力模块中，可以使用三个不同的参数矩阵$\mathbf{W}_q$、$\mathbf{W}_k$、$\mathbf{W}_v$将$\boldsymbol{x}_i$映射为三个新向量$\boldsymbol{q}_i = \boldsymbol{x}_i\mathbf{W}_q$、$\boldsymbol{k}_i = \boldsymbol{x}_i\mathbf{W}_k$、$\boldsymbol{v}_i = \boldsymbol{x}_i\mathbf{W}_v$。此时，经过自注意模块后输出向量$\boldsymbol{t}_i$计算式如式(6)所示：

$$
\begin{aligned}
\hat{\alpha}_{i,j} &= \frac{\boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}}}{\sqrt{d}} \\
\hat{\boldsymbol{\alpha}}_i &= [\hat{\alpha}_{i,1},\ \hat{\alpha}_{i,2},\ ...,\ \hat{\alpha}_{i,l}] \\
\mathbf{A} &= [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} \\
\alpha_{i,j} &= \mathrm{S}\left(\mathbf{A}_i\right)_j \\
\boldsymbol{t}_i &= \sum_j \alpha_{i,j} \boldsymbol{v}_j
\end{aligned}
\tag{6}
$$

式中，$\mathrm{S}\left(\boldsymbol{x}\right)$为对$\boldsymbol{x}$的Softmax运算。

Transformer中除了自注意力模块，还同时存在位置嵌入模块。常用的位置嵌入方式有两种，分别为绝对位置嵌入和相对位置嵌入。对于使用绝对位置嵌入的神经网络模型，例如Transformer，其输入向量$\boldsymbol{x}$主要由两部分构成，分别为标识嵌入向量$\boldsymbol{x}_e$和绝对位置嵌入向量$\boldsymbol{x}_p$，即$\boldsymbol{x} = \boldsymbol{x}_e + \boldsymbol{x}_p$。对于使用相对位置嵌入的神经网络模型，例如TENER (Yan et al., 2019)，其相对位置嵌入计算如式(7)所示：

$$
\hat{\alpha}_{i,j} = \boldsymbol{q}_i \boldsymbol{k}_j^{\mathrm{T}} + \boldsymbol{q}_i \boldsymbol{r}_{i-j}^{\mathrm{T}} + \boldsymbol{u} \boldsymbol{k}_j^{\mathrm{T}} + \boldsymbol{v} \boldsymbol{r}_{i-j}^{\mathrm{T}} \tag{7}
$$

式中，$\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^{d_k}$是可学习的参数，$\boldsymbol{r}_{i-j} \in \mathbb{R}^{d_k}$是相对位置嵌入。

3.4.1 自注意力对模型鲁棒性的影响

在句子级自然扰动下，Transformer模型的全局自注意力权重分布改变，导致模型不鲁棒。若假设当前模型中未使用位置嵌入模块，则此时输入向量$\boldsymbol{x}$满足$\cos(\boldsymbol{x}^o, \boldsymbol{x}^r) = 1$。在句子级自然扰动影响下，受扰动序列与原始序列相比其对应注意力权重分布被改变，导致输出向量发生改变。此时，可列出经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(8)所示：

$$\begin{aligned}\boldsymbol{t}_i^o &= \begin{cases}\sum\limits_{j\in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,l}]\right)_j \boldsymbol{v}_j, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{v}_j, & i \in C_2\end{cases} \\ \boldsymbol{t}_i^r &= \sum_{j\in C_1\cup C_2} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{v}_j, \ i \in C_1 \cup C_2\end{aligned} \tag{8}$$

式中，$C_1 = C_{\mathbb{N}}[1, l]$，$C_2 = C_{\mathbb{N}}[l+1, 2l]$。

由式(8)可以看出，由于自注意力权重分布改变，导致满足式(5)的输出向量$\boldsymbol{t}$数量增加，提高了扰动成功率，最终影响了模型的鲁棒预测。

3.4.2 绝对位置嵌入对模型鲁棒性的影响

为便于后续计算$\cos(\boldsymbol{t}^o, \boldsymbol{t}^r)$，本文按照扰动序列构建方式将原始序列$\mathbf{X}^{o_1}$和$\mathbf{X}^{o_2}$中各字符对齐到扰动序列$\mathbf{X}^r$中对应位置字符。即在$i < l$时，$\boldsymbol{x}_i^o$赋值为$\mathbf{X}^{o_1}$中索引为$i$处字符的表征向量$\boldsymbol{x}_i^{o_1}$，而在$i > l$时，$\boldsymbol{x}_i^o$赋值为$\mathbf{X}^{o_2}$中索引为$i-l$处字符的表征向量$\boldsymbol{x}_{i-l}^{o_2}$。对于使用绝对位置嵌入的模型，输入向量$\boldsymbol{x}$计算如式(9)所示：

$$\begin{aligned}\boldsymbol{x}_i^o &= \begin{cases}\boldsymbol{x}_{i:e}^{o_1} + \boldsymbol{x}_{i:p}^{o_1}, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2} + \boldsymbol{x}_{i-l:p}^{o_2}, & i \in C_2\end{cases} \\ \boldsymbol{x}_i^r &= \begin{cases}\boldsymbol{x}_i^o, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2} + \boldsymbol{x}_{i:p}^r, & i \in C_2\end{cases}\end{aligned} \tag{9}$$

为排除自注意力对模型鲁棒性的影响，在扰动序列中，式(6)中矩阵$\mathbf{A}$加入额外约束以排除自注意力模块的干扰，计算公式如式(10)所示：

$$\begin{aligned}\mathbf{W} &= [-\infty] \in \mathbb{R}^{l\times l} \\ \mathbf{A} &= [\hat{\boldsymbol{\alpha}}_1, \hat{\boldsymbol{\alpha}}_2, ..., \hat{\boldsymbol{\alpha}}_{2l}]^{\mathrm{T}} + \begin{bmatrix}\mathbf{O} & \mathbf{W} \\ \mathbf{W} & \mathbf{O}\end{bmatrix}\end{aligned} \tag{10}$$

式中，$\mathbf{O}$为零矩阵。

此时，经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(11)所示：

$$\begin{aligned}\boldsymbol{t}_i^o &= \begin{cases}\sum\limits_{j\in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1}, \hat{\alpha}_{i,2}, ..., \hat{\alpha}_{i,l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_2\end{cases} \\ \boldsymbol{t}_i^r &= \begin{cases}\boldsymbol{t}_i^o, & i \in C_1 \\ \sum\limits_{j\in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1}, \hat{\alpha}_{i,l+2}, ..., \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^r \mathbf{W}_v, & i \in C_2\end{cases}\end{aligned} \tag{11}$$

观察式(11)可以看出，当$i \in C_1$时，此时模型的预测结果是鲁棒的，当$i \in C_2$时，模型可能会出现预测结果的改变，此时模型是不鲁棒的，综合上述两种情况如式(12)所示：

$$\cos(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r) = \begin{cases}1, & i \in C_1 \\ p < 1, & i \in C_2\end{cases} \tag{12}$$

因此，和未使用绝对位置嵌入的模型相比，绝对位置嵌入的使用会降低模型在句子级自然扰动下的鲁棒性。同时，绝对位置嵌入对模型鲁棒性的负面影响不受限于自注意力模块。

3.4.3 相对位置嵌入对模型鲁棒性的影响

对于使用相对位置嵌入的模型，式(7)所示给出了相对位置嵌入计算过程。此时，输入向量$\boldsymbol{x}$的计算如式(13)所示：

$$\boldsymbol{x}_i^o = \boldsymbol{x}_i^r = \begin{cases} \boldsymbol{x}_{i:e}^{o_1}, & i \in C_1 \\ \boldsymbol{x}_{i-l:e}^{o_2}, & i \in C_2 \end{cases} \tag{13}$$

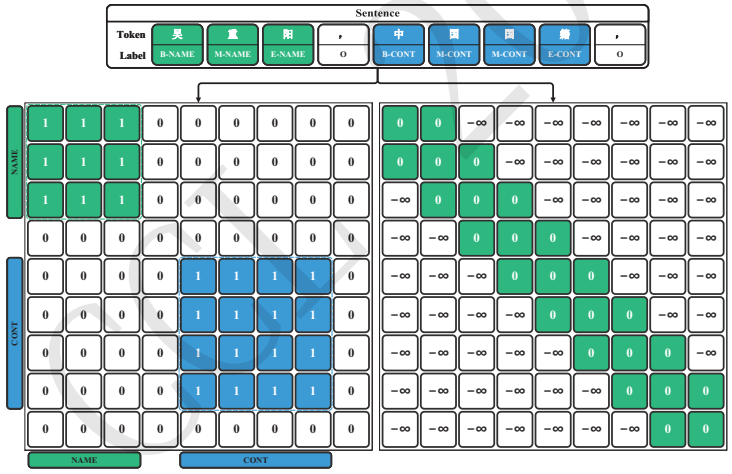
这里为排除自注意力模块的干扰依旧沿用3.4.2节中式(10)所示的约束条件。同时，由于没有绝对位置嵌入向量$\boldsymbol{x}_{i:p}$，原始序列$\mathbf{X}^o$和扰动序列$\mathbf{X}^r$中对应位置处字符的向量表征是相等的。这时经过自注意力模块后的输出向量$\boldsymbol{t}$计算如式(14)所示：

$$\boldsymbol{t}_i^o = \boldsymbol{t}_i^r = \begin{cases} \sum\limits_{j \in C_1} \mathrm{S}\left([\hat{\alpha}_{i,1},\ \hat{\alpha}_{i,2},\ ...,\ \hat{\alpha}_{i,l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_1 \\ \sum\limits_{j \in C_2} \mathrm{S}\left([\hat{\alpha}_{i,l+1},\ \hat{\alpha}_{i,l+2},\ ...,\ \hat{\alpha}_{i,2l}]\right)_j \boldsymbol{x}_j^o \mathbf{W}_v, & i \in C_2 \end{cases} \tag{14}$$

观察式(14)可以看出，在排除自注意力模块影响后，对$\forall i \in C_1 \cup C_2$，$\cos\left(\boldsymbol{t}_i^o, \boldsymbol{t}_i^r\right) \equiv 1$，此时相对位置嵌入并不会影响模型在句子级自然扰动下的鲁棒性。只有同时考虑自注意力模块对模型鲁棒性的影响时，相对位置嵌入对模型鲁棒性的影响才会显现。

4 提高中文实体识别模型鲁棒性方法

基于3.4节对Transformer结构的分析，本文引入两种约束策略来提高模型的句子级非对抗鲁棒性，这两种约束策略对应矩阵的构建方式如图 1所示。



(a) 实体标签增强矩阵　　(b) 滑动窗口约束矩阵

Figure 1: 两种约束策略对应矩阵的构建

4.1 实体标签增强

对于训练集中任意序列$\mathbf{X}$，可得到所有实体的跨度集合$S = \{s_1, s_2, ..., s_m\}$和每个实体的跨度$s$所对应标签$y$。给定起始索引$b_i$和结束索引$e_i$，跨度$s_i$可表示为$(b_i, e_i)$。如图 1所示，此段序列中可能跨度集合与标签分别为$\{(1, 3), (5, 8)\}$和$\{NAME, CONT\}$。如图 1 (a)所示，实体标签增强矩阵$\mathbf{W}_{cl}$为定义在$\mathbb{R}^{l \times l}$上的矩阵，计算方法如式(15)所示，l为序列$\mathbf{X}^o$的长度。

$$\begin{aligned} \mathbf{W}_{cl} &= \beta\left[a_{ij}\right] \in \mathbb{R}^{l \times l} \\ a_{ij} &= \sum_{s_n \in S} \mathbb{1}_{C_{\mathbb{N}}[b_n, e_n]}(i)\, \mathbb{1}_{C_{\mathbb{N}}[b_n, e_n]}(j) \end{aligned} \tag{15}$$

式中，$\mathbb{1}_A(x)$为指示函数，$\beta \in [0,1]$为超参数。若字符c_i和c_j在同一实体跨度内，则a_{ij}为1，否则为0。$\mathbf{W}_{cl}$描述了同实体内字符间的关联关系。

定理1：对受到句子级自然扰动影响的任意一段序列$\mathbf{X}^o$，有扰动序列$\mathbf{X}^r$，实体标签增强矩阵$\mathbf{W}_{cl}$，自注意力权重矩阵$\mathbf{A}$，使用$\mathbf{A}$获得$\boldsymbol{x}$对应输出$\boldsymbol{t}$，基于式(16)增强矩阵$\mathbf{A}$，可获得新输出$\tilde{\boldsymbol{t}}$，则有式(17)成立。证明过程见附录。

$$\mathbf{A} = (\mathbf{W}_{cl} + \mathbf{J}) \odot [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} \tag{16}$$

$$\cos\left(\tilde{\boldsymbol{t}}_i^o,\ \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\ \boldsymbol{t}_i^r\right) \tag{17}$$

式中，$\odot$为哈达玛积，$\mathbf{J} \in \mathbb{R}^{l \times l}$为全1矩阵。

4.2 滑动窗口约束

Transformer的注意力分布是按比例缩放且平滑的，但结合3.4节分析，在NER任务下，若期望模型在句子级自然扰动下更鲁棒，稀疏注意力是更适合的，并非所有字符都需要被关注。对于给定字符，几个上下文字符就足以判断其标签，平滑注意力可能包含一些嘈杂信息。

如图 1 (b)所示，滑动窗口约束矩阵$\mathbf{W}_{cs}$为定义在$\mathbb{R}^{l \times l}$上的矩阵，计算公式如式(18)所示。

$$\begin{aligned} \mathbf{W}_{cs} &= [a_{ij}] \in \mathbb{R}^{l \times l} \\ a_{ij} &= \begin{cases} -\infty, & |i-j| > w \\ 0, & otherwise \end{cases} \end{aligned} \tag{18}$$

式中，$w \in C_{\mathbb{N}}[0, l-1]$为滑动窗口注意力模式中的窗口长度。

定理2：对受到句子级自然扰动影响的任意一段序列$\mathbf{X}^o$，有扰动序列$\mathbf{X}^r$，滑动窗口约束矩阵$\mathbf{W}_{cs}$，自注意力权重矩阵$\mathbf{A}$，使用$\mathbf{A}$获得$\boldsymbol{x}$对应输出$\boldsymbol{t}$，基于式(19)增强，可获得新输出$\tilde{\boldsymbol{t}}$，则有式(17)成立。其中，本文的滑动窗口约束遵循Beltagy等人 (2020)所设计的滑动窗口注意力模式。证明过程见附录。

$$\mathbf{A} = [\hat{\boldsymbol{\alpha}}_1,\ \hat{\boldsymbol{\alpha}}_2,\ ...,\ \hat{\boldsymbol{\alpha}}_l]^{\mathrm{T}} + \mathbf{W}_{cs} \tag{19}$$

5 实验设置与实验结果分析

5.1 实验数据

本文在三个中文NER数据集对模型的句子级非对抗鲁棒性展开评估。

- **Weibo** (Peng and Dredze, 2015)：此数据集源自于社交媒体领域中的新浪微博。
- **Resume** (Zhang and Yang, 2018)：此数据集由上市公司高管的1027份简历经处理而得。
- **OntoNotes4.0** (Pradhan et al., 2013)：此数据集由标注一个大型语料库得到，覆盖了多种文本类型，包括新闻文章、社交媒体文本等，以及不同领域内容，如政治、经济、医学等。在本文中使用其中文部分，并使用了与Che等人 (2013)相同的预处理方法。

5.2 模型

本文依照3.2节构建鲁棒性评估数据集，并在下列模型上评估模型句子级非对抗鲁棒性。

- **BiLSTM-CRF** (Huang et al., 2015)：应用BiLSTM-CRF到NER任务中，使模型有效利用已有和新输入的特征。同时由于CRF层的存在，模型能够利用句子层面的标注信息。
- **Lattice-LSTM** (Zhang and Yang, 2018)：通过将潜在词信息融入到字符信息中并构建有向无环图（晶格结构），利用LSTM门控机制计算字符与词之间的相关性，以提升模型的识别性能。
- **TENER** (Yan et al., 2019)：使用基于相对位置嵌入的Transformer编码器建模句子中的字符和单词级特征，同时删除缩放点积自注意力中比例因子以提升Transformer识别性能。
- **BERT或RoBERTa-CRF**：使用BERT (Kenton and Toutanova, 2019)、RoBERTa (Liu et al., 2019)这类预训练语言模型作为编码器，同时使用CRF作为解码器，在特定领域数据集上对预训练语言模型进行微调。其中，RoBERTa是对BERT进行调整后得到的模型。这一调整涉及对BERT预训练任务的改进，以提高模型稳健性。
- **EntLM** (Ma et al., 2022a)：将NER任务表述为无需任何模板的LM问题，放弃模板构建过程，同时保留预训练模型的单词预测范式，来预测实体位置处与类别相关的标签词。

5.3 模型鲁棒性评估实验

为评估中文NER模型的句子级非对抗鲁棒性，本文设计了一组句子级自然扰动影响实验。通过3.2节所示的句子级扰动样本生成方式在原始序列中引入句子级自然扰动，观察模型对于原始和扰动两段序列中对应命名实体的识别结果，最终实验结果如表 2所示。

模型	Weibo			Resume			OntoNotes4.0		
	S-F1	P-F1	R-F1	S-F1	P-F1	R-F1	S-F1	P-F1	R-F1
BiLSTM-CRF	49.60	50.20	95.41	93.57	93.44	98.96	61.39	60.71	95.50
Lattice-LSTM	57.91	58.30	**98.31**	94.65	94.23	**99.35**	73.79	73.46	**98.96**
TENER	55.51	54.92	92.45	95.19	94.84	98.68	73.02	72.99	95.28
BERT-CRF	67.83	69.41	91.28	95.55	95.46	97.67	81.65	81.32	94.94
RoBERTa-CRF	67.28	68.31	89.67	96.10	95.80	98.00	81.59	81.68	95.23
EntLM	60.65	59.33	86.87	94.24	94.47	96.03	70.92	70.53	84.71

Table 2: 句子级自然扰动下不同数据集中各模型的标准、扰动、鲁棒F1值

表 2实验结果揭示了Transformers模型在处理扰动序列时的脆弱性，从表 2可以观察到：

1. 与LSTM模型相比，Transformers模型在句子级自然扰动下鲁棒性较差。本文认为这是由于Transformer中的自注意力机制允许模型对输入序列中所有位置的字符信息进行加权处理，这使得Transformer易受这类扰动影响。通过变动序列中某些词，使自注意力机制关注到不相关部分，最终干扰模型预测结果。
2. 与Transformer及预训练语言模型加微调相比，预训练语言模型加提示学习所构建模型在应对句子级自然扰动时的鲁棒性较差。
3. 与Transformer相比，预训练语言模型结合微调或提示所构建模型虽可借助预训练语言模型在大规模语料上进行上游学习优势，有助于模型更好地捕捉句子中的语义和语法信息，提高模型对实体边界和类别识别的准确性，但实验结果也显示这类模型在面对句子级自然扰动时的鲁棒性更为脆弱。这一现象与计算机视觉中Zhang等人 (2023)在ImageNet上观测结果相契合，即预训练模型在目标数据集上进行微调的方法不只会使模型获得更强的特征提取能力，也会将预训练模型带有的非鲁棒特征转移到微调后的模型中。

5.4 自注意力及位置嵌入对鲁棒性影响实验

为研究句子级自然扰动下基于Transformer构建的模型鲁棒性较差的主要原因，基于3.4节的分析，发现Transformer易受这类扰动影响的主要原因是由于其中的自注意模块和位置嵌入模块，因此本文进行了如下实验：依次约束Transformer下的自注意力模块和位置嵌入模块，同时观察两种模块对模型鲁棒性的影响，最终实验结果如表 3所示。其中PosE表示位置嵌入模块，Attn表示自注意力模块。

表 3展示了Transformer中自注意力模块和位置嵌入模块单独和共同作用时所对应的实验结果，从中可以观察到：

1. Transformer中位置嵌入模块和自注意力模块的共同作用削弱了句子级自然扰动下模型鲁棒性。但此操作也会使模型预测精度显著降低，尤其对EntLM这类基于预训练语言模型加提示学习范式所构建模型，其预测精度下降幅度最大。
2. 对于使用相对位置嵌入模型，如TENER，单独移除自注意力模块与同时移除位置嵌入模块和自注意力模块都可以显著提高模型鲁棒性，并且这两种方式产生的影响相同。
3. 对于使用绝对位置嵌入模型，如BERT或RoBERTa加CRF、EntLM等，自注意力模块和位置嵌入模块共同影响了模型在句子级自然扰动下的鲁棒性，仅约束其中之一无法有效改善模型在受到此类自然扰动时的鲁棒性。

模型	PosE	Attn	Weibo			Resume			OntoNotes4.0		
			S-F1	P-F1	R-F1	S-F1	P-F1	R-F1	S-F1	P-F1	R-F1
TENER	✓	✓	55.51	54.92	92.45	95.19	94.84	98.68	73.02	72.99	95.28
	✓	×	53.44	53.46	**98.13**	92.72	92.37	**99.66**	68.27	68.21	**99.29**
	×	✓	51.10	50.07	89.15	94.01	92.90	97.60	69.47	68.68	92.63
	×	×	53.44	53.46	**98.13**	92.72	92.37	**99.66**	68.27	68.21	**99.29**
BERT-CRF	✓	✓	67.83	69.41	91.28	95.55	95.46	97.67	81.65	81.32	94.94
	✓	×	23.88	23.82	92.41	82.39	80.70	93.00	53.03	53.02	91.15
	×	✓	38.02	31.13	67.88	79.06	72.38	81.20	59.89	57.02	80.55
	×	×	22.64	22.64	**100.00**	82.84	82.84	**100.00**	54.04	53.69	**99.37**
RoBERTa-CRF	✓	✓	67.28	68.31	89.67	96.10	95.80	98.00	81.59	81.68	95.23
	✓	×	22.37	22.45	95.17	81.74	80.08	92.90	53.39	53.14	92.39
	×	✓	39.14	31.66	61.03	80.42	74.22	81.31	59.11	56.94	79.90
	×	×	22.25	22.25	**100.00**	82.56	82.56	**100.00**	54.45	54.10	**99.41**
EntLM	✓	✓	60.65	59.33	86.87	94.24	94.47	96.03	70.92	70.53	84.71
	✓	×	0.90	0.91	94.50	21.43	19.12	83.93	6.04	6.98	82.52
	×	✓	26.87	23.73	70.23	48.30	37.80	68.32	27.11	23.74	64.11
	×	×	0.89	0.89	**100.00**	14.70	14.70	**100.00**	4.35	4.35	**100.00**

Table 3: 自注意力及位置嵌入对鲁棒性影响实验下不同数据集中各模型标准、扰动、鲁棒F1值

5.5 模型鲁棒性增强实验

针对Transformers模型在句子级自然扰动下所表现出的不鲁棒现象，本文在模型训练与推理阶段引入多种约束策略以提高模型对于此类自然扰动的鲁棒性。不同约束策略下模型的评估结果如表 4所示，其中cl表示使用实体标签增强策略，cs表示使用滑动窗口约束策略。

模型	Weibo			Resume			OntoNotes4.0		
	S-F1	P-F1	R-F1	S-F1	P-F1	R-F1	S-F1	P-F1	R-F1
TENER	55.51	54.92	92.45	95.19	94.84	98.68	73.02	72.99	95.28
+ cl	55.32	54.62	93.87	94.49	94.58	98.81	72.67	72.83	95.73
+ cs	55.73	55.60	**97.40**	94.96	94.47	**99.27**	73.25	73.12	**99.42**
BERT-CRF	67.83	69.41	91.28	95.55	95.46	97.67	81.65	81.32	94.94
+ cl	68.09	67.45	92.22	95.81	96.16	**98.43**	81.12	80.95	95.14
+ cs	66.27	66.12	**93.41**	95.25	94.69	98.22	79.96	79.49	**96.28**
RoBERTa-CRF	67.28	68.31	89.67	96.10	95.80	98.00	81.59	81.68	95.23
+ cl	68.85	67.37	90.57	95.60	95.17	97.96	81.56	81.48	95.21
+ cs	67.29	66.82	**92.96**	95.70	95.85	**99.24**	78.52	78.68	**96.23**
EntLM	60.65	59.33	86.87	94.24	94.47	96.03	70.92	70.53	84.71
+ cl	60.32	59.37	84.52	94.84	93.24	95.52	72.27	72.19	85.57
+ cs	59.64	59.29	**88.57**	93.77	93.49	**98.06**	71.89	71.57	**89.63**

Table 4: 模型鲁棒性增强实验下不同数据集中各模型的标准、扰动、鲁棒F1值

观察表 4实验结果可以得出：

1. 与原始模型相比，本文引入的两种鲁棒性增强策略可提高Transformers模型在句子级自然扰动下的鲁棒性，同时也注意到这两种鲁棒性增强策略也会对模型预测精度产生影响。

2. 与实体标签增强策略相比，滑动窗口约束策略所带来的鲁棒性提升更为明显。同时由于这种约束策略将Transformer中自注意力限制在了某个预先设定的范围，因此该策略会对模型预测精度产生较大的负面影响。
3. 两种约束策略在Transformer下的表现优于预训练语言模型加微调或者提示学习。本文推测，预训练语言模型中多层Transformer块的堆叠可能也会对句子级自然扰动下模型鲁棒性产生影响。未来工作中，将对预训练语言模型中多层Transformer块的堆叠情况进行研究。

6 总结

目前，基于Transformer构建的中文NER系统在实际应用中已取得先进性能，但在实际应用场景下发现，此类系统对输入数据的敏感性较高。当前，NER鲁棒性研究多专注于英语，同时此类研究多专注于对抗性攻击，这类攻击所构造的多数样本在现实场景中并不会出现，因此，本文开展了面向中文NER任务的Transformers模型的非对抗鲁棒性研究。同时由于句子级自然扰动在现有中文NER系统实际应用阶段更为常见，因此本文聚焦于句子级自然扰动下模型鲁棒性。最后，针对模型在句子级自然扰动下的不鲁棒现象，本文引入了两种约束策略以提高模型鲁棒性。

在多个不同规模数据集下实验结果表明：首先，基于Transformer构建的NER模型鲁棒性评估结果均低于LSTM，表明Transformer在面对句子级自然扰动时更为脆弱；其次，同时约束Transformer中的位置嵌入和自注意力模块可有效改善模型在句子级自然扰动下的鲁棒性，但这种策略也会造成模型预测精度的显著降低；最后，本文引入了两种约束策略，实验结果表明其可改善Transformers模型在句子级自然扰动下的鲁棒性，但同时也注意到这两种策略在预训练语言模型加微调或者提示学习范式下所获得鲁棒性受益较低。

参考文献

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Wanxiang Che, Mengqiu Wang, Christopher D Manning, and Ting Liu. 2013. Named entity recognition with bilingual constraints. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 52–62.

Sudeshna Das and Jiaul Paik. 2022. Resilience of named entity recognition models under adversarial attack. In *Proceedings of the First Workshop on Dynamic Adversarial Data Collection*, pages 1–6.

Nathan Drenkow, Numair Sani, Ilya Shpitser, and Mathias Unberath. 2021. A systematic review of robustness in deep learning for computer vision: Mind the gap? *arXiv preprint arXiv:2112.00639*.

Shulin Hu, Huajun Zhang, Xuesong Hu, and Jinfu Du. 2022. Chinese named entity recognition based on bert-crf model. In *2022 IEEE/ACIS 22nd International Conference on Computer and Information Science (ICIS)*, pages 105–108. IEEE.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Shuo Huang, Zhuang Li, Lizhen Qu, and Lei Pan. 2021. On robustness of neural semantic parsers. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3333–3342.

Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, pages 4171–4186.

Xiaonan Li, Hang Yan, Xipeng Qiu, and Xuan-Jing Huang. 2020. Flat: Chinese ner using flat-lattice transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6836–6842.

Zaijing Li, Ting-En Lin, Yuchuan Wu, Meng Liu, Fengxiao Tang, Ming Zhao, and Yongbin Li. 2023. Unisa: Unified generative framework for sentiment analysis. In *Proceedings of the 31st ACM International Conference on Multimedia*, pages 6132–6142.

Bin Liang, Hongcheng Li, Miaoqiang Su, Pan Bian, Xirong Li, and Wenchang Shi. 2018. Deep text classification can be fooled. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence*, pages 4208–4215.

Bill Yuchen Lin, Wenyang Gao, Jun Yan, Ryan Moreno, and Xiang Ren. 2021. Rockner: A simple method to create adversarial examples for evaluating the robustness of named entity recognition models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3728–3737.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Ruotian Ma, Xin Zhou, Tao Gui, Yiding Tan, Linyang Li, Qi Zhang, and Xuan-Jing Huang. 2022a. Template-free prompt tuning for few-shot ner. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5721–5732.

Yubo Ma, Zehao Wang, Yixin Cao, Mukai Li, Meiqi Chen, Kun Wang, and Jing Shao. 2022b. Prompt for extraction? paie: Prompting argument interaction for event argument extraction. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6759–6774.

Sewon Min, Kenton Lee, Ming-Wei Chang, Kristina Toutanova, and Hannaneh Hajishirzi. 2021. Joint passage ranking for diverse multi-answer retrieval. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6997–7008.

Marcin Namysl, Sven Behnke, and Joachim Köhler. 2020. Nat: Noise-aware training for robust neural sequence labeling. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1501–1517.

Nanyun Peng and Mark Dredze. 2015. Named entity recognition for chinese social media with jointly trained embeddings. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 548–554.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Hwee Tou Ng, Anders Björkelund, Olga Uryupina, Yuchen Zhang, and Zhi Zhong. 2013. Towards robust linguistic analysis using ontonotes. In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning*, pages 143–152.

Akshay Srinivasan and Sowmya Vajjala. 2023. A multilingual evaluation of ner robustness to adversarial inputs. In *Proceedings of the 8th Workshop on Representation Learning for NLP (RepL4NLP 2023)*, pages 40–53.

Xingwei Tan, Gabriele Pergola, and Yulan He. 2021. Extracting event temporal relations via hyperbolic geometry. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8065–8077. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Kai Wang, Junfeng Tian, Rui Wang, Xiaojun Quan, and Jianxing Yu. 2020. Multi-domain dialogue acts and response co-generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7125–7134.

Boxin Wang, Chejian Xu, Shuohang Wang, Zhe Gan, Yu Cheng, Jianfeng Gao, Ahmed Hassan Awadallah, and Bo Li. 2021. Adversarial glue: A multi-task benchmark for robustness evaluation of language models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

Jiayi Wang, Rongzhou Bao, Zhuosheng Zhang, and Hai Zhao. 2022a. Distinguishing non-natural from natural adversarial samples for more robust pre-trained language model. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 905–915.

Xiaolei Wang, Kun Zhou, Ji-Rong Wen, and Wayne Xin Zhao. 2022b. Towards unified conversational recommender systems via knowledge-enhanced prompt learning. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 1929–1937.

Jindong Wang, HU Xixu, Wenxin Hou, Hao Chen, Runkai Zheng, Yidong Wang, Linyi Yang, Wei Ye, Haojun Huang, Xiubo Geng, et al. 2023. On the robustness of chatgpt: An adversarial and out-of-distribution perspective. In *ICLR 2023 Workshop on Trustworthy and Reliable Large-Scale Machine Learning Models*.

Hongqiu Wu, Ruixue Ding, Hai Zhao, Pengjun Xie, Fei Huang, and Min Zhang. 2023. Adversarial self-attention for language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13727–13735.

Hang Yan, Bocao Deng, Xiaonan Li, and Xipeng Qiu. 2019. Tener: adapting transformer encoder for named entity recognition. *arXiv preprint arXiv:1911.04474*.

Yue Zhang and Jie Yang. 2018. Chinese ner using lattice lstm. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1554–1564.

Jiaming Zhang, Jitao Sang, Qi Yi, Yunfan Yang, Huiwen Dong, and Jian Yu. 2023. Imagenet pre-training also transfers non-robustness. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 3436–3444.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Neil Zhenqiang Gong, Yue Zhang, et al. 2023. Promptbench: Towards evaluating the robustness of large language models on adversarial prompts. *arXiv preprint arXiv:2306.04528*.

附录.原始扰动

对受到句子级自然扰动影响的任意一段原始序列$\mathbf{X}^o$，令l^o为序列$\mathbf{X}^o$长度，设其对应扰动序列为$\mathbf{X}^r$，令l^r为序列$\mathbf{X}^r$长度，同时有$l^o < l^r$，且满足$\forall i,\, j \in C_{\mathbb{N}}\left[1,\, l^r\right],\, i \neq j,\, \langle \boldsymbol{v}_i,\, \boldsymbol{v}_j \rangle = 0$。

$$
\begin{aligned}
\cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right) &= \cos\left(\frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j}{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}},\, \frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j}{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,j}}}\right) \\
&= \cos\left(\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle + \left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j,\, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} = \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}} \boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}} \boldsymbol{v}_j \right\|^2}}
\end{aligned}
$$

若将约束矩阵$\mathbf{W}_{\mathrm{x}}$加入自注意力模块中，设此时过自注意力模块后输出向量为$\tilde{\boldsymbol{t}}_i^o$和$\tilde{\boldsymbol{t}}_i^r$，要证使用约束矩阵$\mathbf{W}_{\mathrm{x}}$可改善模型在句子级自然扰动下的鲁棒性，即证$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。

附录.定理1证明

$$\begin{aligned}
\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) &= \cos\left(\frac{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}} + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}}, \frac{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,j}} + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}}\right) \\
&= \cos\left(\sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=1}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l_0+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}}
\end{aligned}$$

由于，$\sqrt{1 - \frac{\left\| \sum_{j=l_0+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=b_i}^{e_i} e^{\beta\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}} > \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} = \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。因此，可证：$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。

附录.定理2证明

$$\begin{aligned}
\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) &= \cos\left(\frac{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j}{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}}, \frac{\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j}{\sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}}\right) \\
&= \cos\left(\sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j + \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j\right) \\
&= \frac{\left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle + \left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\langle \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j, \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\rangle}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\| \left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} \\
&= \frac{\left\| \sum_{j=l^o-w}^{l^o} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|} = \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}}
\end{aligned}$$

由于，$\lim_{w\to 0}\sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^o+w} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=l^o-w}^{l^o+w} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} > \sqrt{1 - \frac{\left\| \sum_{j=l^o+1}^{l^r} e^{\hat{\alpha}_{i,j}}\boldsymbol{v}_j \right\|^2}{\left\| \sum_{j=1}^{l^r} e^{\hat{\alpha}_{i,\,j}}\boldsymbol{v}_j \right\|^2}} = \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。因此，可证：$\cos\left(\tilde{\boldsymbol{t}}_i^o,\, \tilde{\boldsymbol{t}}_i^r\right) > \cos\left(\boldsymbol{t}_i^o,\, \boldsymbol{t}_i^r\right)$。