SeqXGPT: Sentence-Level AI-Generated Text Detection

authors： Pengyu Wang, Linyang Li, Ke Ren, Botian Jiang, Botian Jiang, Dong Zhang, Xipeng Qiu

Journal： EMNLP 2023

keywords： Sentence-level, Word-wise, Detection of AIGT

2026-01-24, 杨烨翰

目录

1 背景

研究背景与问题层面动机: 当前的 AI 生成文本 (AI-Generated Text, AIGT) 检测方法仍面临多方面挑战，主要体现在以下几个方面：首先，模型无关 (model-wise) 检测方法，如 DetectGPT 和 Sniffer [2, 4]，通常依赖较长的输入文本（一般超过 100 个 token）以获得稳定的统计特征，因此在短文本或句子级别场景下检测效果受限。其次，基于监督学习的方法，例如对 RoBERTa 等预训练模型进行微调，往往对训练数据分布高度敏感，容易产生过拟合问题，从而影响其在未知模型或跨域场景下的泛化能力。此外，文档级检测方法在处理包含人类与 AI 混合撰写内容的文档，或仅由少量句子构成的文本时，难以进行精细化判断，容易导致较高的假阴性率或假阳性率。综上所述，为有效应对短文本、混合文本以及复杂生成来源等实际应用场景，构建句子级 (sentence-level) 的 AIGT 检测方法具有重要的研究意义和实际价值。因此作者提出 SeqXGPT 来解决上述问题。

方法层面的动机：该工作主要基于以下两点理论依据。首先，已有研究表明，困惑度（Perplexity, PPL）及其相关的概率统计特征在 AIGT 检测任务中具有显著的判别能力 ~ [3]。基于语言模型的 token 级对数概率能够有效反映文本在生成过程中的不确定性差异，从而为区分人类文本与 AI 生成文本提供重要线索。其次，作者观察到，token 级对数概率序列（token-wise log probability list）在结构上类似于语音信号中的连续波形特征。在语音处理领域，卷积神经网络（Convolutional Neural Networks, CNNs）已被广泛应用于从波形信号中提取局部时序模式 [1]。受此启发，本文采用卷积神经网络对 token-wise 对数概率序列进行建模，以提取其局部模式特征，并进一步结合自注意力机制，对序列级上下文信息进行建模。该方法的设计思想主要来源于 DetectGPT 与 Sniffer。已有工作表明，基于平均 token 对数概率的统计特征（DetectGPT），以及基于 token-wise 概率构造的对比特征（Sniffer），均能有效提升 AIGT 检测性能。因此，本文选择直接对 token-wise log probability 序列进行特征提取与建模，以统一并扩展上述两类方法的优势。

The current AIGT methods face two main challenges: 1. model-wise mehtods like DetectGPT, sniffer require a long document as input text (over 100 tokens), making them less effective to detect short text. 2. Supervised methods like fine-tuning RoBERTa prone to overfitting on the training data. 3. The document-level detection methods struggle to accurately evaluate documents containing a mix of AI-generated and human-authored content or consisting of few sentences, which can potentially lead to a higher rate of false negatives or false positives Therefore, it is necessary to **build a sentence-level AIGT detecctors** to deal with the challenges above.

Method-level Motivation: 1. They found perplexity is a significant feature to be used in AIGT detectors. ~ [3] 2. They think the world log probability list are composed like waves in speech processing, where convolution networks are often be used. So they

use conv network to extract this wave feature, followed by a self-attention layer, to process wave-like features. [1] Previous works (etc, Sniffer and Detect GPT)demonstrate that both the average per-token log probability (DetectGPT) and contrastive features derived from token-wise probabilities (Sniffer) can contribute to AIGT detection. They therefore use a token-wise list of log probabilities as the raw feature detection signals.

2 方法

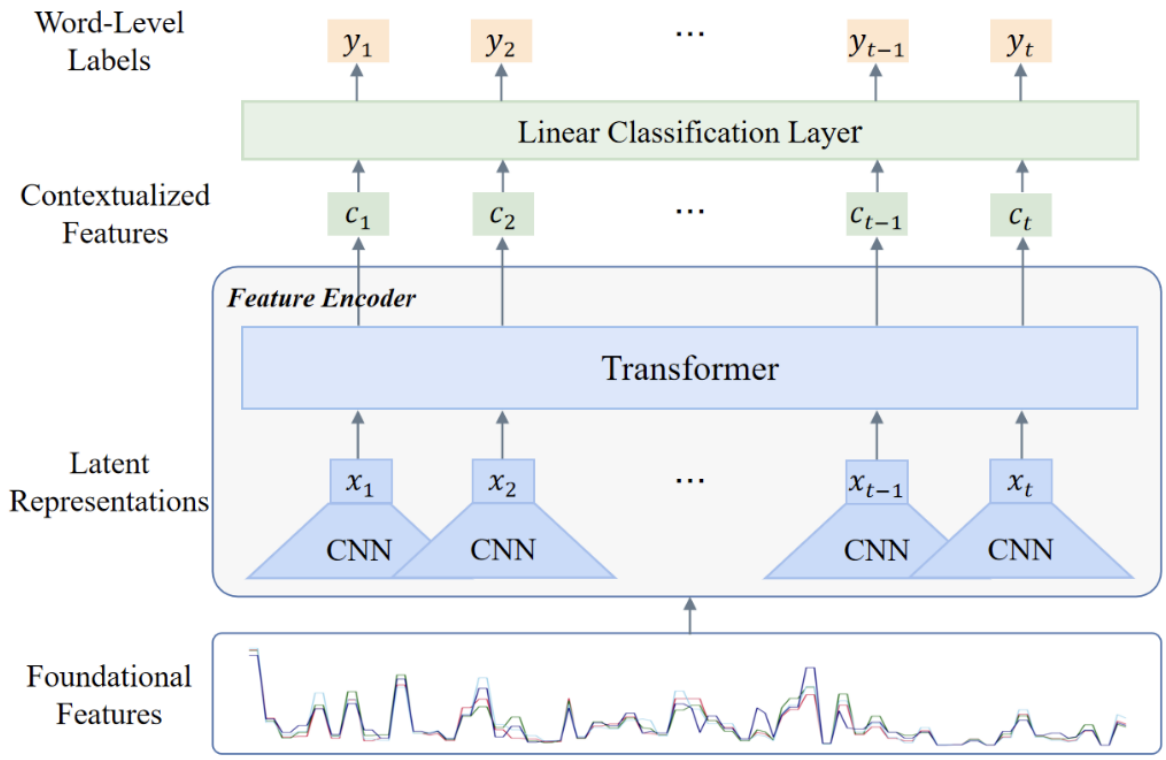

图 1: SeqXGPT Framework

模型架构: 图1 展示了 SeqXGPT 模型-框架架构:token 级概率序列建模为核心，采用“卷积特征提取—上下文建模—层级预测” 的三阶段结构，实现从 word-wise 到 sentence-level 及 context-level 的 AIGT 检测。首先，针对输入文本，利用多个代理语言模型获取对应的 token 级对数概率序列（word-wise probability list）。随后，采用卷积神经网络（CNN）对该概率序列进行一维卷积操作，以捕获其局部模式特征。卷积网络的输出通道配置为 $(64, 128 \times 3, 64)$，由此得到基础特征表示（foundation features）。其次，将来自四个代理模型的 foundation features 在特征维度上进行拼接（concatenation），并输入至 Transformer Encoder 以建模序列级上下文信息。该编码器由两层自注意力结构组成，每层包含 16 个注意力头，前馈网络（FFN）的隐藏维度为 256，从而获得上下文化特征表示（Contextualized Features）。最后，将上下文化特征送入线性分类层（Linear

Classification Layer)，输出对应的 word-wise 预测标签。在此基础上，通过对 word-wise 预测结果进行聚合，进一步得到句子级（sentence-level）以及上下文级（context-level）的检测结果。

句子层级的检测任务: 如图 2所示, 作者将 sentence-level detection task 定义为聚合 world-level 的检测信号，通过统计 world-label 的数量来投票 sentence-label。因此作者在制作数据的时候，同时提示学习，会收集到大量的（human-query，AIresponse）问答对，这个 human-prompt 及作为 human-label，AI response text 即作为 generated-label。同时标注 world 时遵循父子关系，所有句子的 label 属于段落 label，所有句子中的词 label，属于句子 label。因此便地得到了一个自动标注 world-label 的方案。

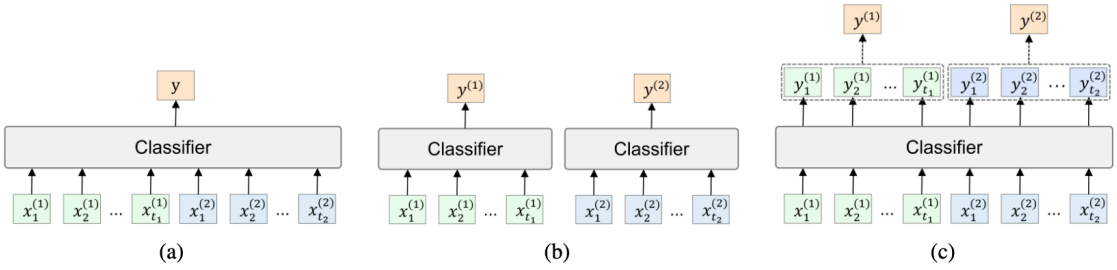

图 2: Strategies for AIGT Detection. (a) Document-Level AIGT Detection: Determine the category based on the entire candidate document. (b) Sentence Classification for Sentence-Level AIGT Detection: Classify each sentence one by one using the same model. (c) Sequence Labeling for Sentence-Level AIGT Detection: Classify the label for each word, then select the most frequently occurring category as the final category for each sentence.

关于方法上的一些小细节: 作者在论文中提到，由于需要不同的代理白盒模型来提取原始的概率特征，但是由于不同模型的分词器不同，可能导致同一个 input text，不同的分词器分出的 sub-token 不同，因此输出的 prob-list 的长度也不一样，因此作者参考 Sniffer 的方法，对齐所有的 token-wise 到 word-wise，这样输出的 prob 可以直接对齐到 word-wise，保证所有模型的输出 prob-list 长度一致，作为后续卷积处理的的基石。作者在 github 上开源代码实现，是直接将四个代理白盒模型的 raw feature 直接 concat，然后送进卷积神经网。

3 实验

本文的主要实验任务是句子级 AI 生成文本检测 (Sentence-Level AIGT Detection)，旨在识别给定文档中的每一句话是由人类撰写的还是由 AI 生成或修改的。作者定义了三种具体的检测设置:

1. **特定模型二分类数据集（Particular-Model Binary Detection Dataset）**：该

数据集用于二分类任务，目标是区分文本是否由指定单一语言模型生成。

2. **混合模型二分类数据集（Mixed-Model Binary Detection Dataset）**：同样用于二分类任务，但包含来自多个语言模型生成的文本，从而评估模型在跨模型场景下的泛化能力。

3. **混合模型多分类数据集（Mixed-Model Multiclass Detection Dataset）**：该数据集用于多分类任务，不仅区分文本是由 AI 生成还是人类撰写，还进一步识别文本的具体来源模型（如 GPT、Claude 等）。

值得注意的是，数据集 (1) 与 (2) 主要用于 **二分类任务**，而数据集 (3) 则用于 **多分类任务**，以验证模型在同时判断文本来源及类别的能力。为此，作者构建了名为 **SeqXGPT-Bench** 的数据集。该数据集基于 SnifferBench [2]，涵盖了新闻、社交媒体、科学文章等多个领域。研究人员通过随机选择人类文档的前 1-3 句作为提示词（Prompt），并利用 GPT-2、GPT-Neo、GPT-J、LLaMA 和 GPT-3.5-turbo 等模型生成后续内容，总计合成 30,000 篇包含人机混合文本的文档。

3.1 评价指标

评价指标：实验采用了分类任务中常用的三种指标来综合衡量模型性能：**精确率 (Precision, P.)**：反映分类的准确性。**召回率 (Recall, R.)**：反映对各类别样本的覆盖能力。**Macro-F1 分数**：作为核心评价指标，综合考虑了精确率和召回率，以评估模型的整体性能。

3.2 实验结果分析与论断支持

实验结果分析与论断支持：表3和表4展示的实验结果有力地支持了作者提出的论断：**超越基准模型**：在句子级检测任务中，SeqXGPT 在多分类任务下的 Macro-F1 达到 95.7%，显著优于传统的 RoBERTa 基准（64.9%）和针对文档级设计的 Sniffer（44.7%）。**Zero-shot 方法的失效**：实验证明，诸如 Log $p(x)$ 和 DetectGPT 等零样本检测方法在句子层面表现极差，原因是句子长度较短，导致困惑度（Perplexity）敏感度高且扰动策略难以奏效。**强大的泛化能力**：在分布外（OOD）测试中，SeqXGPT 的性能远超基于语义学习的 RoBERTa。这验证了作者关于“对数概率特征（Log probability lists）比语义特征更能抵抗过拟合”的观点。

3.3 局限性

局限性：尽管实验结果显著，但仍存在局限，1. 在文档级检测实验中，SeqXGPT 对纯人类撰写文档的识别准确率略低于预期。作者指出，这是因为训练集中人类样本仅来源于每篇文档的前 1-3 句，导致模型对长篇幅人类文本的特征学习不够充分。2. 为了专

	Different AIGT Origins									
	GPT-2					GPT-Neo				
Method	P.(AI)	R.(AI)	P.(H.)	R.(H.)	**Macro-F1**	P.(AI)	R.(AI)	P.(H.)	R.(H.)	**Macro-F1**
log $p(x)$	82.2	74.9	43.1	53.9	63.1	81.2	67.8	34.2	51.7	57.5
DetectGPT	80.9	55.4	32.7	62.4	54.3	82.6	44.2	29.1	71.2	49.4
Sent-RoBERTa	89.3	96.9	88.1	66.5	84.4	89.8	95.6	82.7	66.0	83.0
SeqXGPT	**99.3**	**97.9**	**94.5**	**97.1**	**97.2**	**99.5**	**98.2**	**94.8**	**98.1**	**97.6**

图 3: Results of Particular-Model Binary AIGT Detection on two datasets with AI-generated sentences are from GPT-2 and GPT-Neo, respectively. The Macro-F1 is used to measure the overall performance, while the P. (precision) and R. (recall) are used to measure the performance in a specific category.

	Different AIGT Origins												
	GPT-2		GPT-2-Neo		GPT-J		LLaMA		GPT-3		Human		
Method	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	P.	R.	**Macro-F1**
Sniffer	47.5	56.3	48.4	42.9	39.0	33.5	41.8	16.0	52.8	55.4	51.2	67.2	44.7
Sent-RoBERTa	38.6	48.9	36.9	27.6	34.9	28.7	57.5	33.6	65.5	97.1	89.4	91.6	52.9
Seq-RoBERTa	42.1	81.4	45.3	30.9	61.6	21.6	75.5	82.0	90.3	**98.9**	**94.6**	90.1	64.9
SeqXGPT	**99.2**	**97.9**	**99.3**	**98.2**	**97.6**	**96.8**	**95.8**	**90.8**	**94.1**	93.7	90.7	**95.2**	**95.7**
w/o Transformer	92.4	93.1	92.7	88.9	93.3	62.1	82.1	14.3	22.7	0.2	42.0	95.7	56.9
w/o CNN	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	24.8	100.0	6.6

图 4: Results of Mixed-Model Multiclass AIGC Detection.

注于检测机制的研究，作者在构建数据集时采用了较为简单的续写指令。但在现实场景中，人类使用 LLM 的指令极具多样性，实验未能充分覆盖各种复杂提示策略下的文本生成情况。3. 该方法需要从白盒 LLM（如 GPT-2, LLaMA 等）中提取词级对数概率作为特征。这限制了其在无法获取模型权重的黑盒 API 环境下的直接部署能力。

4 综合评价与个人思考

这篇文章总体来说还是启发很大的，尤其是讲述的把概率特征比做 speech-process 中的波形，然后从 wave2vec 中得到启发通过卷积神经网络来提特征。多代理模型，来获取原始的特征的方法使得模型在溯源 AIGT 的效果得到了很大的提升，这里后续的溯源工作可以基于此，来提取家族模型内在的指纹特征，从而更好的提升家族模型的溯源性能。同时这篇文章将 AIGT 从 document-level 细化到了 sentence-level，AIGT 的检测粒度变得更细了，提出了新的短文本的检测范式。因此我认为这篇工作对于 AIGT detection 领域的共享还是很大的。因为作者 sentence-level detection 的思路是通过聚合 world-level 的检测信号，但很明显作者在处理 world-level detection 的做法很粗糙，尤其是自动 world-level 打标的方法很粗糙。后续可以进行更细粒度的改进。同时因为在实际部署的时候，推理阶段用到的了四个白盒代理模型来实时提取原始特征，部署成本还是很高的，如果在训练阶段，能加一个小模型来学习四个代理白盒的特征提取，推理的时候，只需要这个小模型特征提取模块即可，实际部署成本将会大大降低。

参考文献

[1] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460, 2020.

[2] Kathleen C Fraser, Hillary Dawkins, and Svetlana Kiritchenko. Detecting ai-generated text: Factors influencing detectability with current methods. *Journal of Artificial Intelligence Research*, 82:2233–2278, 2025.

[3] Sebastian Gehrmann, Hendrik Strobelt, and Alexander M Rush. Gltr: Statistical detection and visualization of generated text. *arXiv preprint arXiv:1906.04043*, 2019.

[4] Eric Mitchell, Yoonho Lee, Alexander Khazatsky, Christopher D Manning, and Chelsea Finn. Detectgpt: Zero-shot machine-generated text detection using probability curvature. In *International conference on machine learning*, pages 24950–24962. PMLR, 2023.