Can Watermarks Survive Translation? On the Cross-lingual Consistency of Text Watermark for Large Language Models

> https://arxiv.org/abs/2402.14007

- Observation：当前的水印算法对翻译攻击不鲁棒，也就是说从加了水印的文本在被翻译成其它语言时检测的成功率大大降低
- 本文贡献：
  - 一个新的攻击方式：Cross-lingual Watermark Removal Attack (CWRA)
    - 辨析CWRA和re-transaltion：
      - re-translation：把response翻译成另一个语言再翻译回来
      - CWRA：把prompt翻译成另一个语言，送入LLM得到另一个语言的watermarked输出，再翻译回去
  - 两个在这种攻击下maintain consistency的factor

    从KGW入手，KGW如果想在翻译过后还能保持consistency，有两个factor：
    - 不同语言但语义相近的token需要落在同一个partition里，either red/green
    - 不同语言但语义相近的prefix对应的partition需要基本相同
  - 一个能有效抵御这种攻击的新水印算法
    - 之前的工作SIR其实已经对第二个factor做了优化，方式是不用hash function做partition，而是先对context用一个multilingual embedding model转换成embedding，然后训了一个transform model，让partition的结果和semantic context embedding强相关
    - 本文主要对第一个factor做优化：

      想规避的情况是什么呢？例如SIR在partition的时候，对于一个multilingual tokenizer，我 和 I 这两个token语义非常相近，但是可能没落在同一个red/green list里。

      solution：对vocabulary的partition不再以token为单位，而是以semantic cluster为单位，把语义相近的词绑定在同一个组里，以group为单位加bias。从原来的给green token加bias，改成给green clusters加bias。