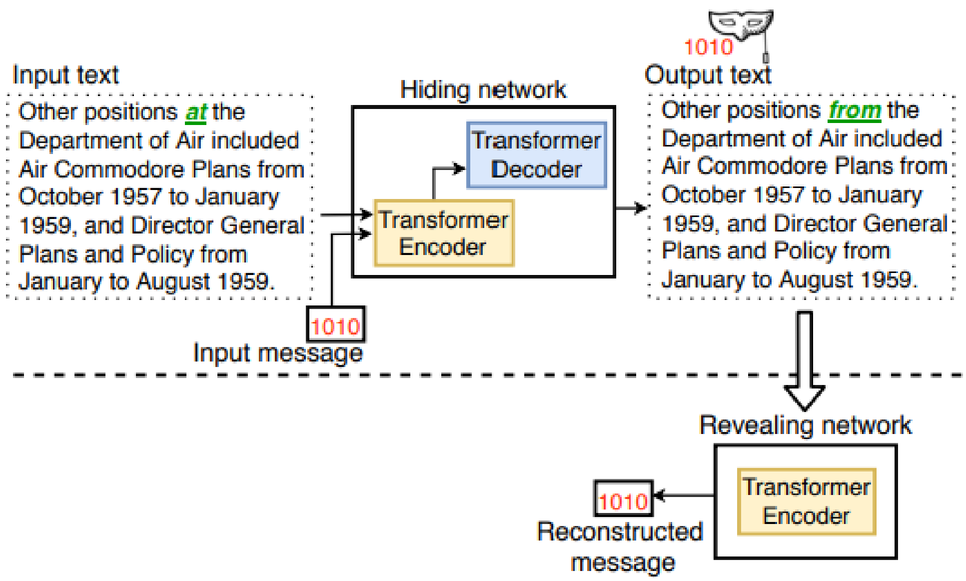
Adversarial Watermarking Transformer: Towards Tracing Text Provenance with Data Hiding

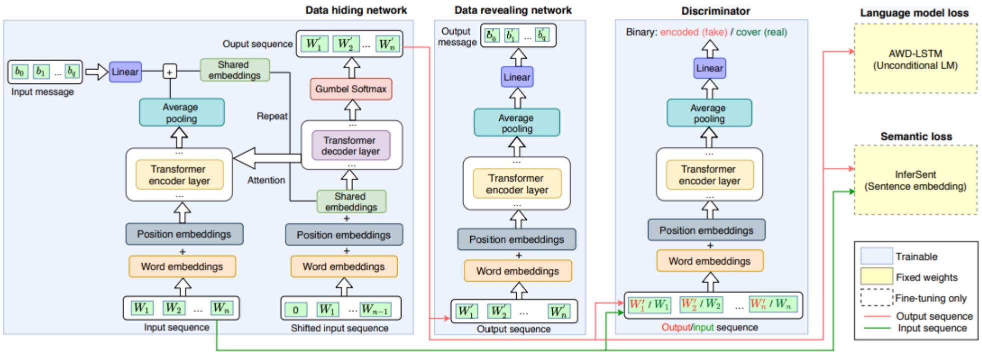
> [IEEE Xplore Full-Text PDF:]

这篇文章就是Text Watermarking下Post-hoc方法中的Neural Network based Approach，讲的还是非常全面，尤其对于Effectiveness, Secrecy和Robustness的评估方法可以借鉴

- 核心思想:
  - Encoder-Decoder + discriminator network，用了Adversarial training
  - Data Hiding：CV中比较成熟，当时NLP里使用还比较少。Encoder把Secret Message融入文本形成Modified Text，Decoder来恢复Secret Message



- NLP中Data Hiding的方式
  - Linguistic Steganography
    - Translation
    - Generate text according to a shared key (require mask)
  - Model Watermarking
    - embedding the watermark into model's weight: requires white-box access
    - assigning specific labels for a trigger set: only requires black-box access
- **Our approach**
  - 使用场景：减少LLM滥用情况，由于hide的secret message可以是multi-bit，所以不仅可以判别是否是machine-generated，还可以包含更多使用者信息，如API key可以被当作secret message
  - 具体操作方法:



1. Data Hiding Network：

   用Transformer Encoder编码Sentence，再把Sentencing Embedding和Message Embedding加起来，送入Transformer Decoder，变换成Modified Text：(W1′, W2′, W3′, ..., Wn′)

2. Data Revealing network：

   把Modified Text：(W1′, W2′, W3′, ..., Wn′)再用Encoder编码，过Linear解出Output Message

3. Discriminator

   比较Modified Text和Real Sentence的差异，希望差异越小越好

- Experiment Result

  1. Effectiveness

     - Text utility: Measured by **SBERT （SBERT比meteor score更注重语义）**

       SBERT-distance可以用来衡量语义

     - bit accuracy: Measured by **bitwise message accuracy**

  2. Secrecy Evaluation：尝试训练一个能识别encoded text和non-encoded text的分类器

  3. Robustness Evaluation

     > 对每种攻击手段，去衡量文本改变的程度和准确度下降程度 文本改变程度可以用meteor score / SBERT distance衡量

     - Attack（Removal Attack）

       - Random changes and denoising：初级攻击者，不知道AWT的操作细节

         - Random changes: 随机删除部分词语 / 近义词替换
         - Denoising：text reconstructing，重写；这篇文章贡献了一个DAE，可供text reconstructing

       - Re-watermarking and de-watermarking：高级攻击者，知道AWT的所有操作细节，但是不知道AWT的具体模型结构和weight

         类似逆向工程，看看攻击者能不能重新train一个迷惑AWT