Detection of machine-generated fragments in text

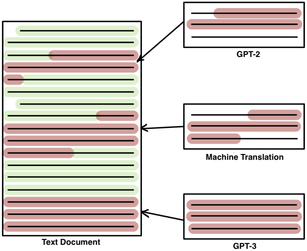
Anastasiya Voznyuk

Moscow Institute of Physics and Technology

Expert: Andrey Grabovoy

2023

Goal of research



Suggest a model, that will detect machine-generated fragments in text and classify them according to their origin model. Number of model is fixed and know.

Problem statement

Let

$$\mathbb{D} = \left\{ \left[t_j \right]_{j=1}^{n} \quad | \quad t_j \in \mathbf{W}, n \in \mathbb{N} \right\}$$

be the space of documents, **W** is the alphabet.
Given set of N documents

$$\mathbf{D} = \bigcup_{i=1}^{N} D^i, D^i \in \mathbb{D}.$$

C is a set of $K + 1$ labels for classification, where 0 is the label of human-written fragment, $\{1...K\}$ are the labels representing corresponding K language models, participated in generating **D**.

$$\mathbb{T} = \left\{ \left[t_{s_j}, t_{f_j}, C_j \right]_{j=1}^{J} | t_{s_j} = t_{f_{j-1}}, s_j \in \mathbb{N}_0, f_j \in \mathbb{N}, C_j \in \mathbf{C} \right\},$$

where J is a number of fragments, t_{s_j} and t_{f_j} are start and end of the j-th fragment.

Problem statement

Our model is

$$\phi : \mathbb{D} \to \mathbb{T} \qquad \phi : \mathbf{g} \circ \mathbf{f},$$

f is mapping, responsible for text segmentation.
g is mapping, responsible for classifying obtained fragments.

The **quality criteria** is macro-averaged precision and recall, where S is ground truth fragmentation and R is predicted fragmentation. We compare segments on sentence level.

$$prec(S, R) = \frac{1}{|R|} \sum_{r \in R} \frac{|\bigcup_{s \in S} (s \cap r)|}{|r|},$$

$$rec(S, R) = \frac{1}{|S|} \sum_{s \in S} \frac{|\bigcup_{r \in R} (s \cap r)|}{|s|},$$

Dataset

There wasn't any open datasets with consistent texts from several authros, where some authors are LLMs and the positions of author change is known. That's why we did our own one:

We took the Medium dataset with articles from Medium.com.

- Cropped it to the length of 4000 tokens
- Randomly picked from 2 to 4 paragraphes
- Generated paragraphes with LLaMA with prompt consisting of previous paragraph of picked paragraph
- Replaced the picked paragraphes with the artifical ones

Result - dataset with 10000 documents.

Baseline experiment

We didn't change how documents in our dataset are divided into paragraphs, and that's why every paragraph could be labeled as **0** (machine-generated) or **1** (human-generated). Then, we fine-tuned RoBERTa-XLM on our texts for classification task on the paragraph-level. After that we took the **[CLS]** token and tried to clustered these vectors with cosine distance. We got two distinctive clusters with cosine values close to 1 and -1, which shows high confidence of the model.

The same experiment but on sentence-levels showed much worse results, a lot of cosine distances were closer to 0 rather than to 1 and -1.

Future Work

- Weaken the hypothesis "One paragraph - one author" and develop a method for style-change detection within paragraph with a method of sliding window
- Add other LLM's as authors (i.e Alpaca-7b or Mistral-7b) and try to distinct several models.

Literature

- **German Gritsay et al.**, 2022, Automatic Detection of Machine Generated Texts: Need More Tokens
- **Sebastian Gehrmann et al.**, 2019, GLTR: Statistical Detection and Visualization of Generated Text
- **Mikhail P. Kuznetsov et al.**, 2016, Methods for intrinsic plagiarism detection and author diarization
- RUATD Competition 2022
- **Adaku Uchendu et al.** Authorship Attribution for Neural Text Generation