Detection of machine-generated fragments in text

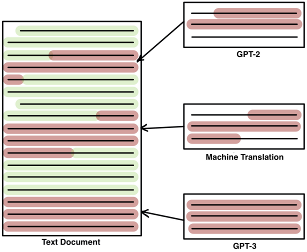
Anastasiya Voznyuk

Moscow Institute of Physics and Technology

Course: My first scientific paper
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 known.

Problem statement

Let

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

be the space of documents, $\mathbf{W}$ is the alphabet.
Given set of N documents

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

$\mathbf{C}$ 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 $\mathbf{D}$.

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

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.

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

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

Proposed method

Transform feature vectors to minimise the variance within groups by target author label.

Let $(\mathbf{b}_i)_{i=1}^{n}$ be sequence of basic feature vectors $\mathbf{b}_i \in \mathbb{R}^{n_b}$.

Transformation of vectors:

$$T : \mathbb{R}^{n_\mathrm{b}} \to \mathbb{R}^{n_\mathrm{t}} \quad T(x) = \mathbf{W}^\mathsf{T} x \quad \mathbf{W} \in \mathbb{R}^{n_\mathrm{b} \times n_\mathrm{t}}$$

Number of fragments with author $c \in \mathbf{C}$

$$N_c = \sum_{i=0}^{K+1} [C_i = c]$$

μ_c is centroid of the transformed feature vectors with label c:

$$\mu_c = \frac{1}{N_c} \sum_{i=1}^{K+1} T(\mathbf{b}_i)[C_i = c].$$

$$L_\mathrm{c} = \sum_{c=0}^{C} \frac{1}{N_c} \sum_{i=1}^{K+1} \| T(\mathbf{b}_i) - \mu_c \|^2 [C_i = c]$$

Computational experiment

Segmentation Pipeline

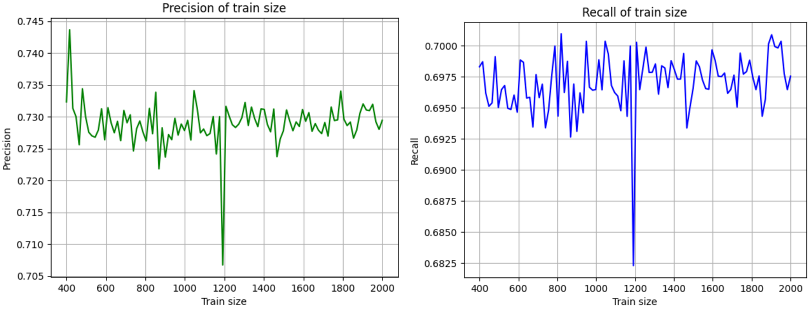
Binary segmentation: 2000 generated documents with 5-6 fragments by human and GPT-2, each consists of of 500-600 tokens.
Multiclass segmentation: 2000 generated documents with 5-6 fragments by for 3 authors, each consists of of 500-600 tokens. Same for 4 and 5 authors.
Multiclass Classifiaction Take documents with several language model as authors with human among authors. Label all non-human fragments 'machine-generated' and fed them to fine-tuned BERT model to classify the fragments.

Multuclass classification of fragments

	precision	recall	f1-score	support
0	0.27	0.25	0.26	267
1	0.16	0.12	0.13	246
2	0.23	0.28	0.25	238
3	0.16	0.19	0.18	252
4	0.19	0.18	0.18	244
5	0.15	0.15	0.15	246
6	0.26	0.24	0.25	264
7	0.17	0.17	0.17	243
accuracy			0.20	2000
macro avg	0.20	0.20	0.20	2000
weighted avg	0.20	0.20	0.20	2000

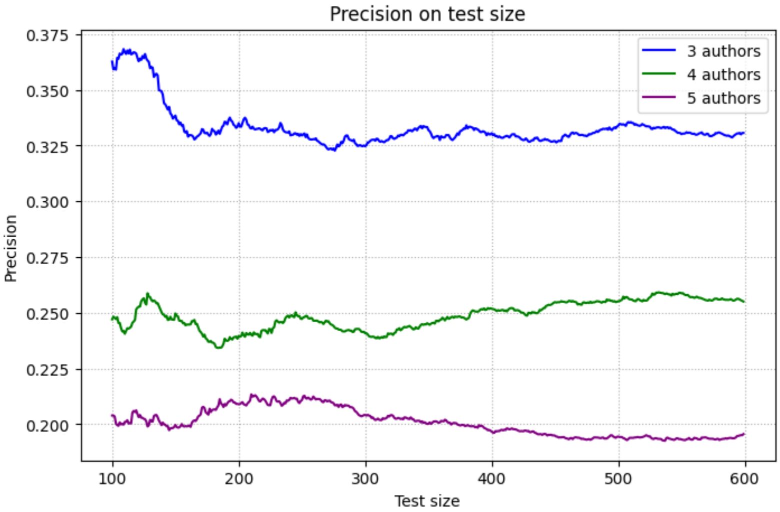
Accuracy is 20%, showing the context is needed when classyfying authors of fragments. Second reason is similarity of language models and their generating style.

Binary segmentation of documents



Only are 2 labels for fragments: human and machine. Both precision and recall don't depend on train size. Precision is around 73%, recall is around 69.75%.

Multiclass segmentation of documents



The best precision is 33% for 3 authors, the worst is 20% for 5 authors.

Multiclass segmentation of documents

The best precision is 92% for 5 authors, the worst is 86% for 3 authors.

Conclusion

New model to detect large machine-generated fragments by different LLM in the documents is suggested;

Results for binary segmentation : 73% precision, 70% recall;

Results for multiclass segmentation : 33% precision, 86% recall for 3 authors;

Results for classification: only 20% accuracy on 8 classes.

Next:

- improving classificator;
- improving the process of segmenting;
- adding CRF to fix problems of very short fragments after segmentation.

.

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