Методы детекции машинно-сгенерированных фрагментов в документах на базе анализа смены стиля

Анастасия Евгеньевна Вознюк
Научный руководитель: к.ф.-м.н. А. В. Грабовой

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных
Направление: 01.03.02 Прикладные математика и информатика

15 июня 2024

Цель детекции машинно-сгенерированных фрагментов

Проблема

В документах все чаще и чаще встречаются фрагменты, написанные языковыми моделями, и необходимо уметь обнаруживать такие фрагменты.

Цель

Предложить методы детекции фрагментов различного авторства в документах в случае смены авторов по фиксированным позициям и в случае единственной смены авторов — смены стиля.

Решение

Использовать контекст фрагментов и искать смену стилистики.

Детекция фрагментов

Детекция смены стиля

Общая постановка нашей задачи

Определим документ как конечную последовательность символов из заданного алфавита $\mathcal{W}$. Пространство документов:

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

Дан набор из N документов

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

Определим множество авторов, тексты которых встречаются в наборе $\mathcal{D}$:

$$\mathcal{C} = \{0, \ldots, k-1\}.$$

Детекция фрагментов в документе

Для каждого документа $d \in \mathbb{D}$ существует разбиение на фрагменты различного авторства:

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

где J — количество фрагментов разного авторства в документе,
t_{s_j} и t_{f_j} — начало и конец j-ого фрагмента одного автора,
C_j — автор j-ого фрагмента.

Модель детектора

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

где $\mathbf{f}$ — модель выделения фрагментов, а
$\mathbf{g}$ — классификатор авторов.

Определение авторов параграфов

Пусть для документа $d \in \mathbb{D}$ известны его параграфы $\mathcal{P} = (p_1, \ldots, p_n)$ и $\mathbf{x} = (\mathbf{x}_1, \ldots, \mathbf{x}_n)$ — векторные представления параграфов.

Вероятностная модель для определения авторов

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = p(\mathbf{y}|\mathbf{x}) = \frac{\exp \Phi(\mathbf{x}, \mathbf{y})}{\sum\limits_{\mathbf{y}' \in \mathcal{Y}^n} \exp(\Phi(\mathbf{x}, \mathbf{y}'))},$$

где $y_i \in \mathcal{C}$ - метка автора для параграфа p_i,
$\mathcal{Y}^n$ — все возможные последовательности меток длины n,
а функция $\Phi(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d$ определена как :

$$\Phi(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{n} \Big(\log \phi_{\texttt{EMIT}}(y_i \to x_i) + \log \phi_{\texttt{TRANS}}(y_{i-1} \to y_i) \Big),$$

где $\phi_{\texttt{EMIT}}$ определяет вероятность появления метки y_i для $\mathbf{x}_i$, а $\phi_{\texttt{TRANS}}$ определяет вероятность появления метки y_i и y_{i-1} в качестве соседей,

Задача оптимизации для определения авторов параграфов

Функция потерь для выборки документов $\mathcal{D}$ и последовательностей меток $\hat{\mathcal{Y}}$:

$$\mathcal{L}(\hat{\mathcal{Y}}, \mathcal{D}) = -\sum_{i=1}^{|\mathcal{D}|} \log(p(\mathbf{y}^i|\mathbf{x}^i)) = -\sum_{i=1}^{|\mathcal{D}|} \log\Big[\frac{\exp\Phi(\mathbf{x}^i, \mathbf{y}^i)}{\sum\limits_{\mathbf{y}' \in \mathcal{Y}^n} \exp(\Phi(\mathbf{x}^i, \mathbf{y}'))}\Big] =$$

$$= \sum_{i=1}^{|\mathcal{D}|} \Big(\underbrace{\log\Big[\sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp(\Phi(\mathbf{x}^i, \mathbf{y}'))\Big]}_{\pi_n} - \Phi(\mathbf{x}^i, \mathbf{y}^i)\Big),$$

$\mathcal{Y}^n$ — все возможные последовательности меток длины n,

Быстрый пересчёт функции потерь

$$\pi_i = \log \sum_{\mathbf{y} \in \mathcal{Y}^i} \exp(\Phi(\mathbf{x}, \mathbf{y})),$$

$$\pi_i^j = \log \sum_{\substack{\mathbf{y} \in \mathcal{Y}^i \\ y^i = j}} \exp(\Phi(\mathbf{x}, \mathbf{y})),$$

где $j \in \mathcal{C}$, $1 \leq i \leq n$

Утверждение 1

π_i^j выражается через π_{i-1}^j

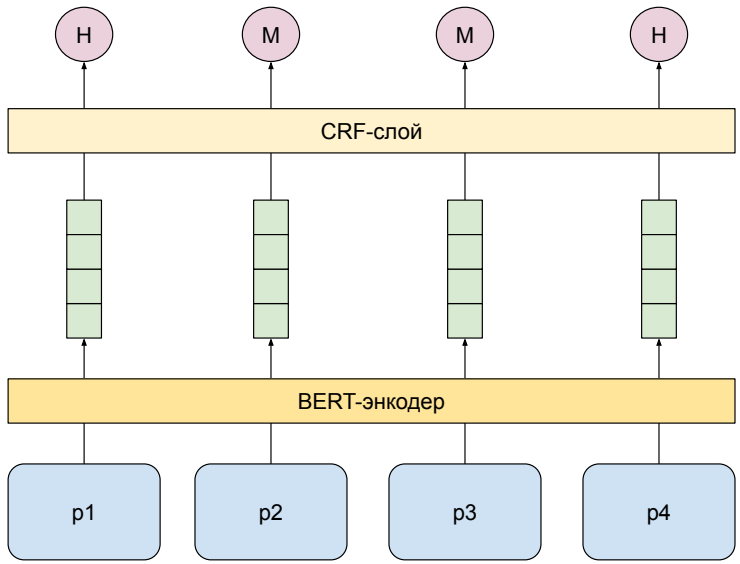


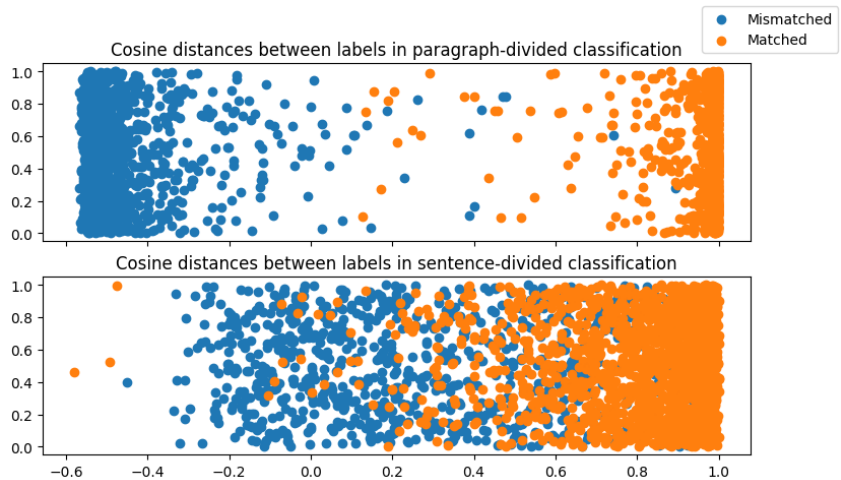
Схема модели для определения авторов параграфов

Эксперименты с детекцией фрагментов по параграфам

Для детекции фрагментов по параграфам был сгенерирован новый датасет на основе статей с Medium.com. В текстах статей из 4-6 параграфов и некоторые параграфы заменяли на машинно-сгенерированные.

Метод	Точность
RoBERTa	0.89
RoBERTa-CRF	0.94

Точность детекции



Разделение векторных представлений параграфов и предложений с помощью косинусной близости

Детекция смены стиля

Пусть для документа $d \in \mathbb{D}$ известно, что

$$\exists I_d \in \mathbb{N}_0 \quad \mathbf{g}([t_0, t_I)) = 0, \quad \mathbf{g}([t_{I+1}, t_{|\mathbf{D}|})) = 1$$

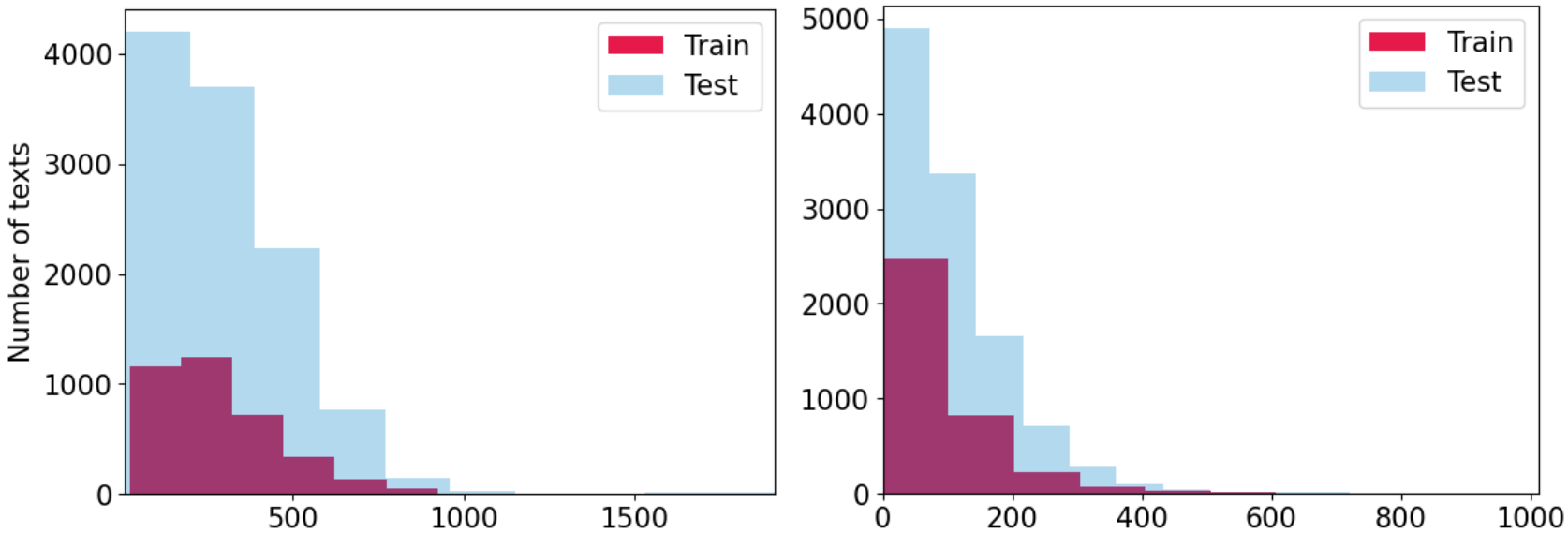
Тогда необходимо с помощью модели $\mathbf{f}$ найти индекс смены авторов.

Введем функцию-детектор с параметром скользящего окна ℓ, которая для токена в документе оценивает его вероятность быть токеном, в котором сменяются авторы:

$$\psi_\ell : \mathbb{D} \times \mathbb{N}_0 \quad \psi_\ell(d, i) = \mathbb{P}(t_i = 1 | t_{i-\ell}, \ldots, t_{i-1})$$

$$I_d = \arg \max_{0 \le i \le |d|} \psi_\ell(d, i)$$

Эксперименты со сменой стиля



(a) Длины текстов

(b) Позиция смены автора

Распределения статистик в текстах после токенизации

Для детекции смены стиля данные были взяты из набора данных с соревнования SemEval2024 Task 8 SubtaskC. Данные были дополнительно аугментированы для увеличения размера выборки и внесения разнообразия в позиции смены автора.

Эксперименты со сменой стиля

Для оценки качества предлагается использовать метрику MAE — Mean Absolute Error, среднее отклонение предсказанного индекса от истинного, считается на уровне слов в документе.

Модель	**Исходный датасет**	**Новый датасет**
RoBERTa-base	31.56	30.71
RoBERTa-large	25.25	20.66
longformer-base	23.16	22.94
longformer-large	22.97	20.33
DeBERTaV3-base	16.12	13.98
DeBERTaV3-large	15.16	**13.38**
Top 1 соревнования	15.68	-

Метрика MAE на исходных и новых (аугментированных) данных. Дополнительно приведено лучшее решение с таблицы результатов соревнования. Longformer предлагался в качестве базового решения

Выносится на защиту

1. Модель детекции смены авторов в текстах, когда смена авторов происходит только на уровне параграфов с помощью марковской линейной цепочки.
2. Модель детекции смены авторов в тексте, в случае, когда эта смена авторов происходит единожды, но может быть в произвольной позиции в документе с помощью моделей на основе трансформеров.

Публикации

1. A. Voznyuk et al.. Leveraging Transfer Learning for Detecting Boundaries of Machine-Generated Texts // Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024) (на опубликовании).