Методы детекции машинно-сгенерированных фрагментов в документах на основе 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 работы

Исследуется задача детекции машинно-сгенерированных фрагментов в текстовых документах.

Проблема:

Выделять в текстовых документах фрагменты человеческого текста и фрагменты, написанные с помощью языковых моделей.

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:

Рассмотреть различные частные случаи задачи детекции с ограничениями по количеству авторов в документе или с ограничениями в позициях смены авторов.

Общая постановка задачи

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

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

Дан набор из N документов

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

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

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

Рассматривается три подзадачи: классификация автора всего документа, детекция фрагментов в документах по заданным позициям в документе и детекция смены стиля в произвольной позиции в документе.

Постановки подзадач

Классификация автора документа

$$\phi : \mathbb{D} \to \mathbf{C}.$$

Метрика: точность или F1-мера.

Детекция фрагментов

Для каждого документа $d \in \mathbb{D}$ существует представление в виде разбиения на фрагменты различного авторства:

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

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

Постановки подзадач

Тогда модель детектора определяется как композиция отображений:

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

где **f** — отображение для выделения текстовых фрагментов, а **g** - отображение для классификации получившихся фрагментов.
Метрика: мера Жаккара для истинного разбиения и разбиения, полученного с помощью модели.

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

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

Метрика: средняя абсолютная ошибка между истинной позицией индекса смены и предсказанной позицией смены авторства.

Методы решения

Классификация автора документа

Пусть для $d \in \mathbb{D}$ были получены векторные представления $\mathbf{h} = (\mathbf{h}_{\text{CLS}}, \mathbf{h}_1, ..., \mathbf{h}_n)$ с помощью BERT или другого энкодера. Тогда автором документа будет:

$$\hat{y} = \arg\max(\text{softmax}(\mathbf{W} \cdot \mathbf{h}_{\text{CLS}} + \mathbf{b})),$$

где $\mathbf{W}, \mathbf{b}$ - обучаемые параметры модели.

Детекция фрагментов

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.

$$p(y_1, \ldots, y_n | \mathbf{x}_1, \ldots, \mathbf{x}_n) = p(\mathbf{y}|\mathbf{x}),$$

где $y_i \in \mathbf{C}$ - метки авторов

Методы решения

$$p(\mathbf{y}|\mathbf{x}) = \frac{\exp \Phi(\mathbf{x}, \mathbf{y})}{\sum_{\mathbf{y}' \in \mathcal{Y}^n} \exp(\Phi(\mathbf{x}, \mathbf{y}'))},$$

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

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

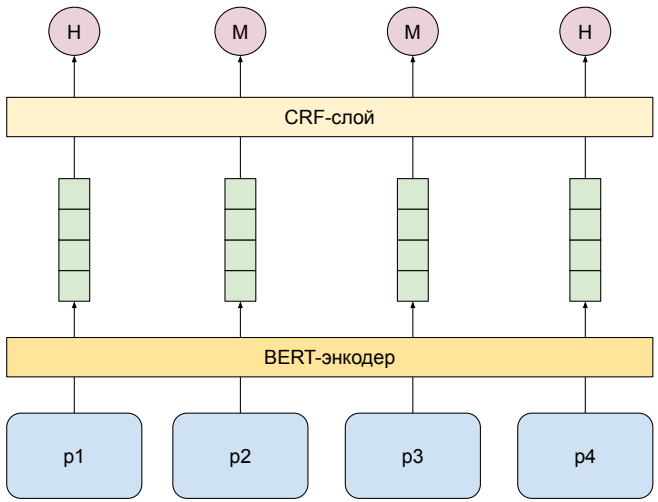


Рис.: Схема модели с марковской линейной цепочкой

Методы решения

Функция потерь для выборки документов $\mathbf{X} \in \mathbb{D}$:

$$\mathcal{L}(\mathbf{Y}, \mathbf{X}) = -\sum_{i=1}^{|\mathbf{X}|} \log(p(\mathbf{y}^i|\mathbf{x}^i)) =$$

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

Введём функцию $\pi[i][j] = \log \sum_{\substack{y' \in \mathcal{Y}^i \\ y'[-1] = \mathcal{Y}[j]}} \exp(\Phi(\mathbf{x}, \mathbf{y}))$, где $j \in \mathbf{C}$,

$1 \leq i \leq n$

Методы решения

Детекция смены стиля

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

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

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

Эксперименты с бинарной классификацией

Метод	Точность
TF-IDF	0.64223
BERT-base-multilingual	0.73430
RoBERTa-base	0.63847
XLM-RoBERTa-base	0.72661
XLM-RoBERTa-large	0.76777
DeBERTa-v3-base	0.72661
mDeBERTa-v3-base	0.76662
ruBERT	0.77288

Таблица: Точность бинарной классификации различных подходов.Цветом выделены модели, предобученные на корпусе русских текстов

Для бинарной классификации документов был взят набор текстами на русском языке с соревнования RuATD.

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

Таблица: Точность детекции

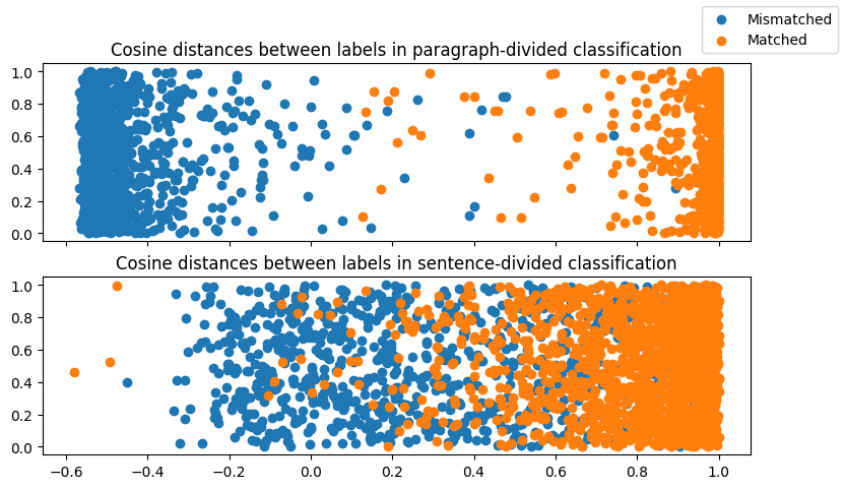
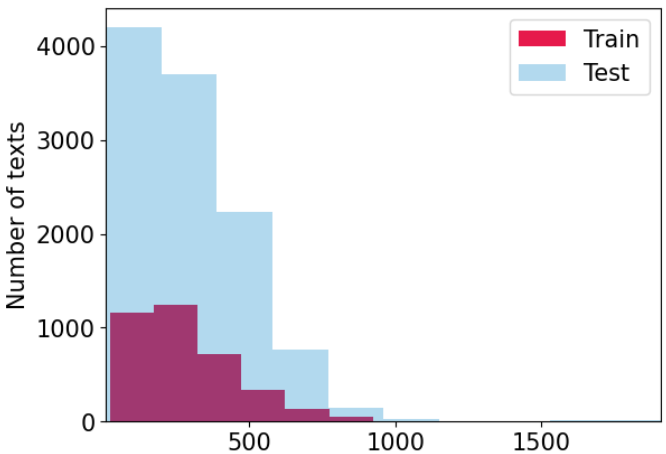
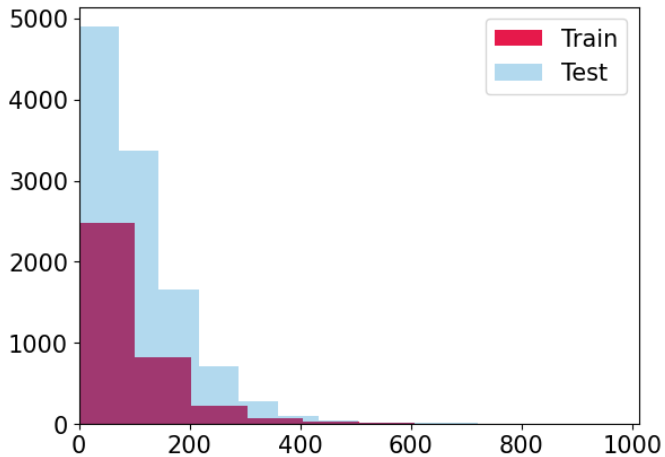


Рис.: Разделение векторных представлений параграфов и предложений с помощью косинусной близости

Эксперименты со сменой стиля



(a) Длины текстов



(b) Позиция смены автора

Рис.: 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 увеличение размера выборки и внесения разнообразия в позиции смены автора.

Эксперименты со сменой стиля

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	-

Таблица: Метрика MAE на исходных и новых (аугментированных) данных. Дополнительно приведено лучшее решение, получившее первое место по результатам соревнования

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. Leveraging Transfer Learning for Detecting Boundaries of Machine-Generated Texts // Proceedings of the 18th International Workshop on Semantic Evaluation (SemEval-2024) (In Printing).