Методы детекции машинно-сгенерированных фрагментов в документах

Отчет о научно-исследовательской работе
за весенний семестр 2023/2024 учебного года

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

Московский физико-технический институт
(национальный исследовательский университет)
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

18 мая 2024

Цель работы

Предложить модель для определения границы между частью текста, написанной человеком, и продолжением этой части, сгенерированной языковой моделью. Данная граница может быть в любой части текста, но она проходит по словам.
Предлагается использовать трансформерные архитектуры в качестве решения, так как на данный момент именно они показывают наилучшие результаты [1]

[1] Macko et al., 2023, MULTITuDE: Large-Scale Multilingual Machine-Generated Text Detection Benchmark

Общая постановка задачи

Введем пространство документов:

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

где $\mathbf{W}$ - алфавит.
Дан набор из N документов

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

В общем случае, $\forall \mathbf{D} \in \mathbb{D}$

$$\exists \mathbb{T} = \left\{ \left[t_{s_j}, t_{f_j}, C_j \right]_{j=1}^{J} | t_{s_j} = t_{f_{j-1}}, s_j \in \mathbb{N}_0, f_j \in \mathbb{N}, C_j \in \{0, 1\} \right\},$$

где J - количество фрагментов разного авторства, t_{s_j} и t_{f_j} - начало и конец j-ого фрагмента, внутри которого все токены одного авторства. 0 соотвествует человеческому тексту, 1 - машинному тексту.

Частная постановка задачи

В такой постановке модель описывается отображением

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

f отображение для выделения текстовых фрагментов
g отображение для классификации получившихся фргаментов.
Решим задачу в частном случае, когда известно, что

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

В таком случае, задача сводится только к нахождению индекса единственного токена, где происходит смена автора с помощью отображения **f** .

Предлагаемое решение

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

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

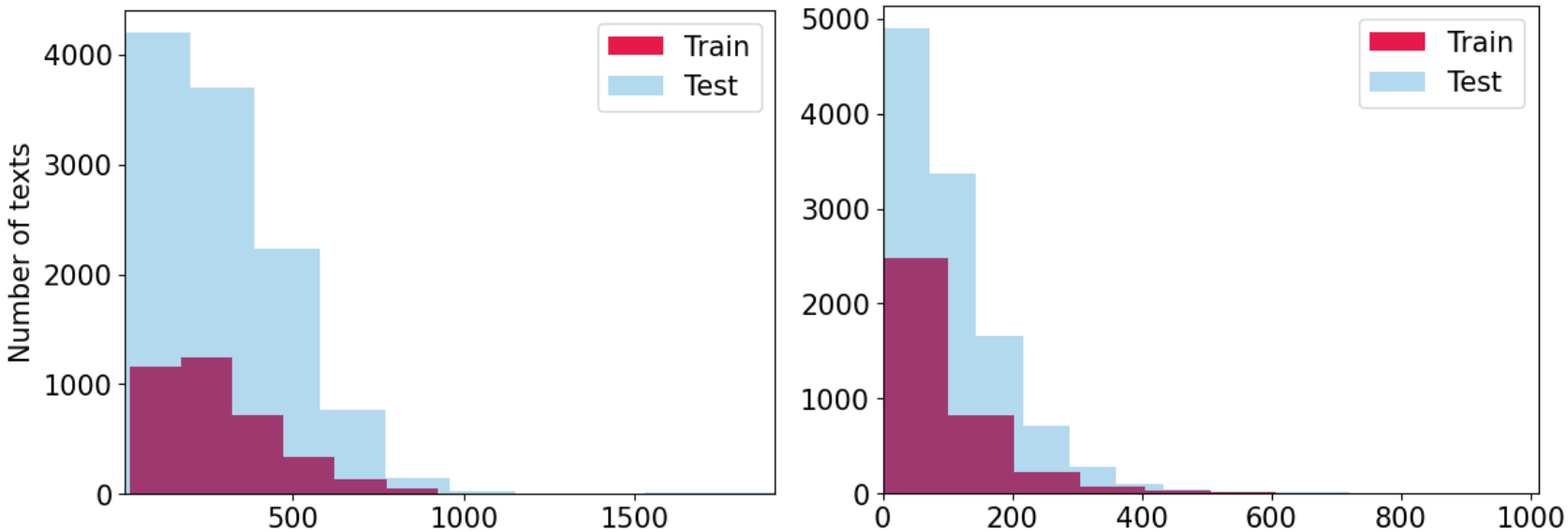
$$I_{\mathbf{D}} = \arg \max_{0 \le i \le |\mathbf{D}|} \psi_\ell(\mathbf{D}, i)$$

Метрикой качества будет MAE (Mean Absolute Error) между предсказанным I_{pred} и истинным I_{true} .
Альтернативное решение: для получения $I_{\mathbf{D}}$ использовать алгоритма Витерби $V(|D|, k)$ для Conditional Random Fields.

Данные

1. Исходные данные были взяты из датасета для соревнования SemEval2024 Task 8 SubtaskC.
2. В тренировочных данных были представлены тексты из датасета PeerReview, продолженные с помощью GPT-4.
3. В тестовых данных дополнительно были тексты из датасета со студентческими эссе Outfox.
4. Тренировочных данных для обучения было недостаточно, поэтому мы дополнительно сделали аугментацию по предложениям для увеличения объем тренировочных данных.

Статистики исходных текстов

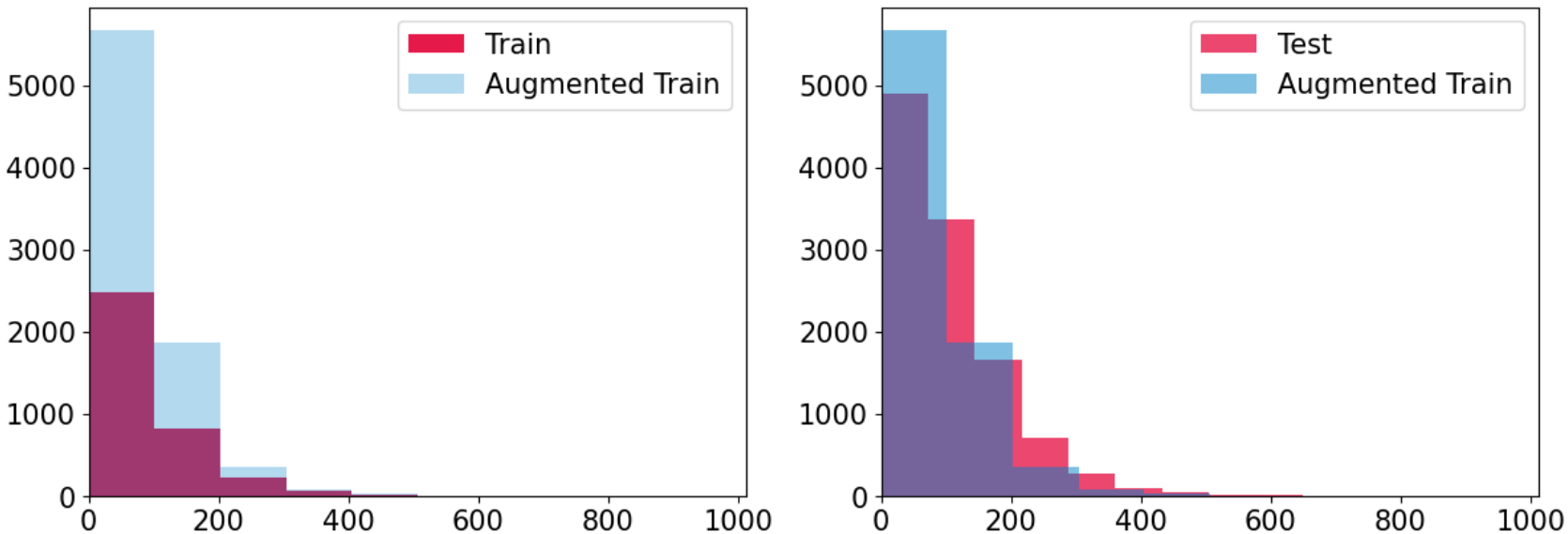


(a) Длины текстов

(b) Позиция смены автора

Рис.: Распределения статистик в текстах после токенизации

Аугментация данных делает распределения похожими



(a) (b)

Рис.: Распределения позиций смены автора в текстах

Эксперимент

Цель эксперимента:
показать, что даже простая аугментация данных позволяет сильно улучшить предсказания, а также отобрать наилучшую модель для решения данной задачи. Поэтому мы сравнивали поведение на двух наборах данных - исходном и аугментированном.
Рассмотренные модели:
в качестве модели ψ было рассмотрено несколько Transformer-based моделей, которые были дообучены для решения задачи классификации токенов.
Были рассмотрены модели Longformer, RoBERTa, DeBERTa.

Результаты

Model	**test**
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**

Таблица: MAE на исходном \ аугментированном датасете.

Итоги НИР за семестр и текущая работа

Результаты

1. Предложено решение для детекции смены авторов в тексте, в случае, когда эта смена авторов происходит единожды, но может быть в произвольном токене
2. Принято участие в соревновании по теме НИР - SemEval 2024 Task 8, и по итогам на post-evaluation стадии был получен новый текущий лучший результат по всему лидерборду

Текущая работа

1. обоснование использования CRF для решения задачи детекции
2. сравнение текущего лучшего метода на текстах, в которых есть несколько моделей-авторов

Выносится на защиту

1. Архитектура решения для детекции смены авторов в текстах, когда смена авторов происходит только на уровне параграфов (результат осеннего семестра)
2. Архитектура решения для детекции смены авторов в тексте, в случае, когда эта смена авторов происходит единожды, но может быть в произвольном токене
3. Использование CRF для задачи детекции границ авторов

Список работ автора по теме НИР

Публикации

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

Выступления с докладом

1. Методы детекции машинно-сгенерированных фрагментов в документах // 66-я Всероссийская научная конференция МФТИ, 2024.