Анализ неопределенности и внутренних представлений языковых моделей для поиска сгенерированных документов

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

Кафедра интеллектуальных систем ФПМИ МФТИ
Специализация: Интеллектуальный анализ данных

21 декабря 2024

Цель работы

Исследуется проблема поиска сгенерированных документов. Стандартные методы поиска не являются устойчивыми к смене генератора или смене домена, поэтому нужно предложить более устойчивые методы.
Требуется предложить метод поиска сгенерированных фрагментов с помощью анализа неопределенности языковой модели, которой передан документ, требующий проверку. Помимо этого, предлагается исследовать методы анализа внутренних представлений языковых моделей, а также внутренних размерностей текстов при обработке текстового документа.

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 $\mathbf{D}$:

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

Тогда задача классификации автора документа записывается как:

$$\phi : \mathbb{D} \to \mathbf{C} \tag{1}$$

Анализ неопределенности

$\boldsymbol{\theta}$ - параметры нашей модели f.

$$y_k = f(D, y_1, \ldots, y_{k-1} \mid \boldsymbol{\theta})$$

$$\boldsymbol{y} = [y_1, \ldots, y_n]^T$$

Maxiumum Sequence Probability

$$\text{MSP}(\boldsymbol{y}|D, \boldsymbol{\theta}) = 1 - P(\boldsymbol{y}|D, \boldsymbol{\theta}) \tag{2}$$

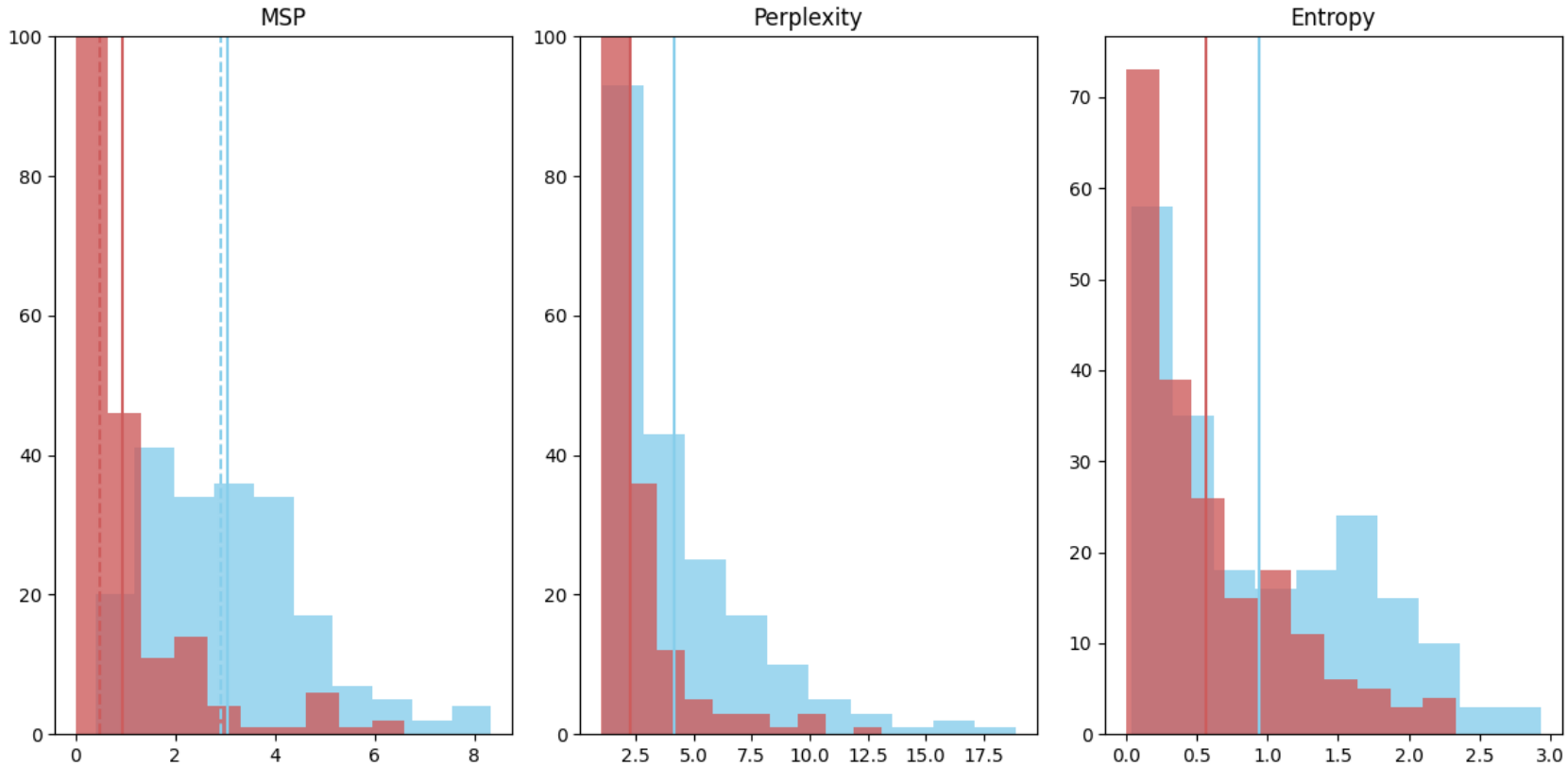
Perplexity

$$P(\mathbf{y}|D, \boldsymbol{\theta}) = \exp\{-\frac{1}{|D|} \log P(\mathbf{y}|D, \boldsymbol{\theta})\} \tag{3}$$

Mean Token Entropy

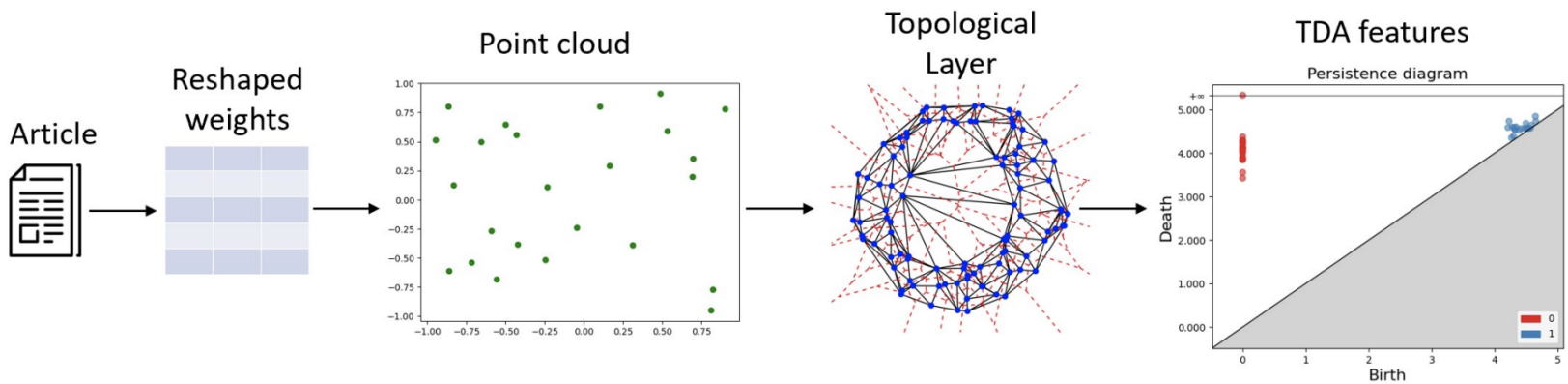
$$\mathcal{H}_T(\mathbf{y}, D; \boldsymbol{\theta}) = \frac{1}{|D|} \sum_{l=1}^{|D|} \mathcal{H}\left(y_l \mid \mathbf{y}_{<l}, D, \boldsymbol{\theta}\right) \tag{4}$$

Вычислительный эксперимент



Было взято 1000 человеческих текстов и 1000 текстов и на них были посчитаны фнукции, описанные на предыдущем слайде: **Maximum Sequence Probability, Perplexity, Mean Token Entropy**.

Использование топологических признаков текста



Основной подход основан на использовании понятий персистентной гомологии конечного набора точек $\mathcal{M}$ в метрическом пространстве с метрикой d.

$E_t = \{(v, u) : v, u \in \mathcal{M}, \quad d(v, u) \leq t\} \quad \forall t \in (0, \infty)$

Персистентная гомология PH_i определена набором признаков размерности i, так PH_0 задается компонентами связности, PH_1 - циклами, и т.д.

Внутреняя размерность

У каждого признака есть своя *продолжительность жизни*, это пара $(t_{\text{birth}}, t_{\text{death}})$, когда данный признак появляется, и когда исчезает.

Введем α-взвешенную сумму, $l(\gamma)$ - продолжительность жизни признака γ.

$$E^i_\alpha(X) := \sum_{\gamma \in \mathrm{PH}_i(X)} |l(\gamma)|^\alpha \tag{5}$$

$$E^0_\alpha(X) \sim Cn^{\frac{d-\alpha}{d}}, n \to \infty \Leftrightarrow \alpha < d \tag{6}$$

$$\dim_{\mathrm{PH}}(\mathcal{M}) = \inf\left\{d \mid \exists C \quad E^0_d(X) \le C \quad \forall X \subset \mathcal{M}, |X| < \infty\right\}. \tag{7}$$

Гипотеза

Внутренняя размерность $\dim_{\mathrm{PH}}(\mathcal{M})$ показывает количество степеней свободы у точки в $\mathcal{M}$.

Персисентная гомологическая размерность (PHD) - метрика, основанная на внутренней размерности текстов, показала себя статистически значимой метрикой для разделения текстов разной природы для первых языковых моделей[1]. Однако для более новых моделей разделимость уже не такая хорошая.

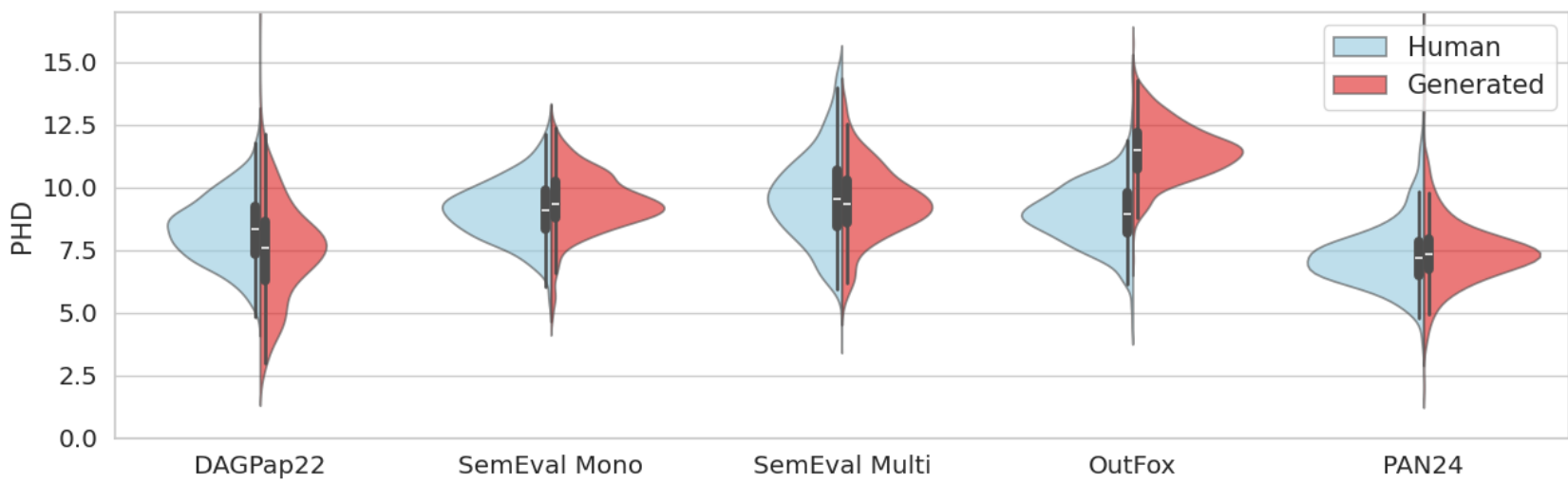
Гипотеза

Можно адаптировать PHD для новых моделей, так что она все еще будет статистической метрикой разделимости текстов.

[1] Tulchinskii et al. Intrinsic Dimension Estimation for Robust Detection of AI-Generated Texts, NeurIPS 2023

Вычислительный эксперимент

Dataset	PHD_{human}	$PHD_{machine}$
OutFox	8.96 ± 1.21	11.48 ± 1.13
SemEval24 Mono	9.11 ± 1.19	9.41 ± 1.2
SemEval24 Multi	9.65 ± 1.81	9.42 ± 1.44
DAGPap22	8.35 ± 1.33	7.48 ± 2.01
PAN24	9.4 ± 1.05	8.52 ± 1.59
MGT-1 Mono	9.19 ± 1.75	8.96 ± 2.24
MGT-1 Multi	8.76 ± 1.85	8.6 ± 2.29



Новый способ подсчитывать внимание

Пусть N - длина текстовй последовательности. Выделим в тексте "якоря" $d_1, .., d_n$, в которых содержится основной смысл текста. Рассмотрим голову внимания h со слоя l модели M. Определим *QK-score* $S_{QK}^{(l,h)}(d_i)$ и *Attention-score* $S_{Att}^{(l,h)}(d_i)$:

$$S_{QK}^{(l,h)}(d_i) = q_N^{(l,h)\top} k_{t_i}^{(l,h)}, \quad S_{Att}^{(l,h)}(d_i) = a_{N,t_i}^{(l,h)}, \quad i \in \{1, 2, ..., n\} \tag{8}$$

Гипотеза

QK-score может лучше решать задачу чем стандартный подсчет внимания и быть более интерпретируемым.

Вычислительный эксперимент

Данные два подхода подсчета внимания сравнивались на задаче ответов на тестовые вопросы MMLU.

	LLaMA...			
Method	**2-13B**	**2-70B**	**3-8B**	**3-70B**
Baseline, Acc	47.4	57.7	60.5	**78.2**
Baseline, PA	34.6	45.9	47.7	**70.1**
QK-score, Acc	**49.7**	**58.9**	**63.0**	77.9
QK-score, PA	**38.3**	**47.1**	**49.3**	67.9

Таблица: Сравнение различных моделей на датасете MMLU. Приводятся следующие метрики: Accuracy (Acc) и Permutation Accuracy (PA).

Дополнительно сравниваются другие датасеты с тестовыми вопросами, а также другие модели. Больше результатов приведено в статье.

Пример сравнения QK-Score и внимания

```
What singer appeared in the 1992 baseball film 'A League
of Their Own'\nOptions:\n
A. Brandy.\n B. Madonna.\n
C. Garth Brooks.\n D. Whitney Houston.\n
E. I don't know.\n F. None of the above.\n
```

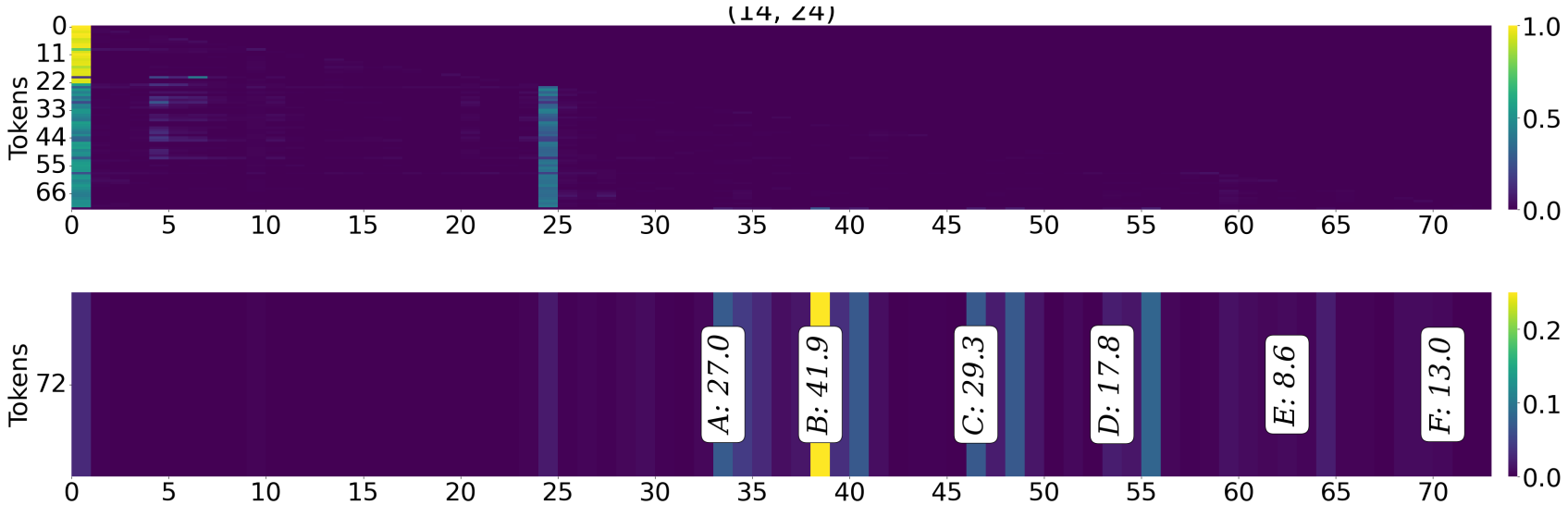


Рис.: Пример когда QK-Score и внимание выдают одинаковый ответ

Пример сравнения QK-Score и внимания

```
What singer appeared in the 1992 baseball film 'A League
of Their Own'\nOptions:\n
A. Brandy.\n B. Madonna.\n
C. Garth Brooks.\n D. Whitney Houston.\n
E. I don't know.\n F. None of the above.\n
```

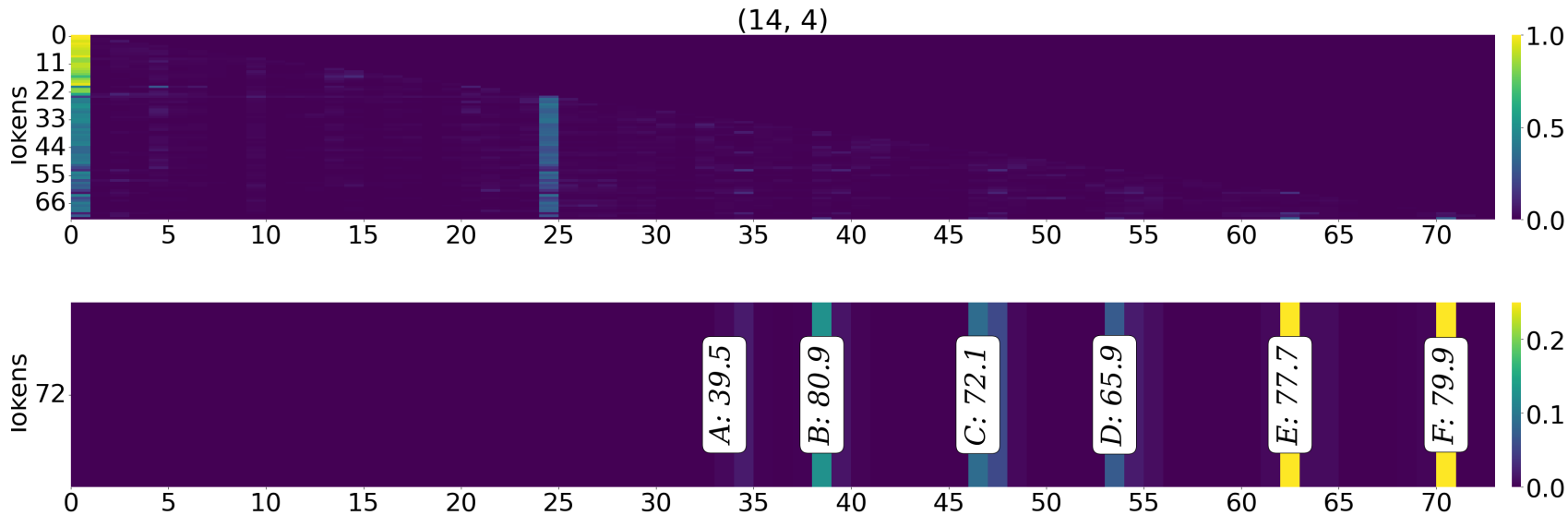


Рис.: Пример когда только QK-Score выдает корректный ответ

Итоги НИР за семестр и планы на следующий семестр

Результаты

1. Опубликована 1 работа на A* конференции, еще 1 принята к публикации, еще 2 в состоянии препринта и ожидают оценки ревьюеров;
2. Получены первые результаты с анализом PHD для сгенерированных текстов и с оценкой неопределенности на них;

Планы

1. Предложить новые эксперименты для проверки гипотез;
2. Необходимо модифицировать PHD чтобы добиться разделимости по текстам от более новых моделей;
3. Модифицировать QK-Score для задачи поиска сгенерированных фрагментов, определить какие якоря могут быть в этой задаче;

Список работ автора по теме НИР

Публикации

1. Listening to the Wise Few: Select-and-Copy Attention Heads for Multiple-Choice QA. *arXiv preprint:2410.02343*
2. Advacheck at GenAI Detection Task 1: AI Detection Powered by Domain-Aware Multi-Tasking, *Proceedings of Workshop on Detecting AI Generated Content at COLING 2025*
3. Are AI Detectors Good Enough? A Survey on Quality of Datasets With Machine-Generated Texts, *arXiv preprint:2410.14677*

Выступления с докладом

1. DeepPavlov 1.0: Your Gateway to Advanced NLP Models Backed by Transformers and Transfer Learning.// *EMNLP, Miami, Florida*