



Факультет компьютерных наук

22 апреля 2025 г.

Шутки в сторону: машинное обучение и интерпретируемый искусственный интеллект в задачах генерации юмористических текстов

Король Михаил БПМИ2310, 2 курс

mkorol@hse.ru

Научный руководитель: д.ф.-м.н. профессор Громов В.А.



Введение

Актуальность, цель и гипотеза

В данный момент ИИ не умеет генерировать юмор. Точнее, из множества сгенерированных шуток, довольно низкий процент окажется действительно смешным. Несмотря на большое количество работ, посвященных теме генерации юмора и юмору в целом, он остается одним из самых сложных явлений для понимания и формализации с точки зрения науки.

Цель: найти качественные различия между обычными и юмористическими текстами для создания методов их автоматической классификации.

Гипотеза: существуют фундаментальные различия в структуре языка, используемого в юмористических и литературных текстах, которые могут быть выявлены и количественно описаны с помощью методов теории хаоса и топологического анализа.



В статье [1] авторы исследуют структуру естественного языка с целью различения текстов, написанных человеком, и текстов, созданных ботами. Для понимания структуры языков авторы собирают словари эмбедингов, после этого по взятым текстам строятся биграмммы, то есть векторизованные два рядом стоящих слова, и производится векторизация Вишарта, авторы используют несколько метрик для оценки кластеризации, после чего статистически доказывают различие в структуре текстов, написанных людьми, и текстов, написанных ботами. Было бы интересно использовать эту методику для проверки гипотезы о том, что структура юмора и литературных текстов имеет статистически значимые семантические различия.



В статье [2] в качестве цели предполагается доказать, что семантические траектории являются хаотическими рядами. Рассматривая слова как вектора, мы хотим изучать тексты как пути динамической системы в фазовом пространстве. Семантическая траектория как раз является таким путем.

В статье вводится метод плоскости Энтропия-Сложность, с помощью которого авторы показывают, что семантическая траектория действительно является хаотичным рядом. Рассматривая юмористические тексты, можно применить этот метод для семантических траекторий в юморе, сделать выводы, к каким последовательностям они относятся, а так же понять, есть ли на этой плоскости явное различие между литературными текстами и юмором.



Мартин, Пластино и Россо (MPR) [3] предлагают подход, позволяющий отличить хаотический ряд от ряда, генерируемого простой детерминированной системой, и от ряда, генерируемого случайным образом. Чтобы использовать такой метод, нужно как-то представить наши текста в виде временного ряда. Собран корпус анекдотов и корпус литературы. Произведена базовая обработка корпусов, которая включает в себя очистку данных и лемматизацию. Далее с помощью словаря эмбедингов получаем ряд векторов.



Рассмотрим наблюдаемую часть временного ряда $y_0, y_1, \dots, y_t, \dots$ и разобьем его на отрезки длины k . В теории их называют z -вектора.

$$z_0 = (y_0, y_1, \dots, y_{k-1})$$

$$z_1 = (y_1, y_2, \dots, y_k)$$

И так далее. Обычно k - небольшая величина. Чтобы ее оценить, можно воспользоваться методом ложных ближайших соседей, добавляя условие на выполнение теоремы Таккенса [4]: размер z -вектора должен удовлетворять условию $m > 2d + 1$, где d — размерность аттрактора. Подробнее см. отчет стр. 6. Суть метода заключается в вычислении двух величин, основываясь на полученных вероятностях, характеризующих исходный временной ряд.



Первая величина – это привычная нам энтропия, но нормированная на ее максимальное значение ($\log m$)

$$0 \leq H \leq 1$$

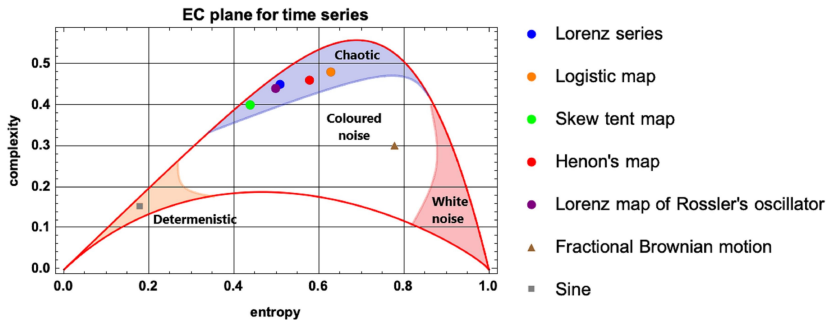
Вторая характеристика носит название сложности, а если быть точным, MPR-сложности (которая названа по первым буквам фамилий ее авторов).

$$C_{\text{MPR}} = Q_0 \cdot H \cdot \|P - P_e\|$$

где P_e – равномерное распределение, то есть: $P_e = \{p_i = 1/N\}$, H – энтропия, Q_0 – нормализующая константа, которая гарантирует, что $0 \leq C_{\text{MPR}} \leq 1$, $\|P - P_e\|$ показывает, насколько уклоняется актуальное распределение от распределения равномерного.



Благодаря этим двум характеристикам получается следующая картина:



Теоретические границы плоскости Энтропия-Сложность



Для анализа семантических путей будем использовать кластеризацию Вишарта¹ [5]. Этот метод был выбран на основе экспериментов, проведенных в исследовании [1], где он показал высокую эффективность на похожей задаче. Для оценки качества кластеризации будем использовать индекс Калински-Харабаша² (CH), который выглядит как:

$$CH_{\text{adj}} = (CH \times \text{ratio_not_noise})^T$$

Где T является гиперпараметром, а ratio_not_noise — количество точек, помеченных как шум, поделенный на общее количество точек.

¹github.com/quynhu-d/stb-semantic-analysis-tools/blob/main/lib/clustering/WishartParallelKD.py

²scikit-learn.org/stable/modules/generated/sklearn.metrics.calinski_harabasz_score.html

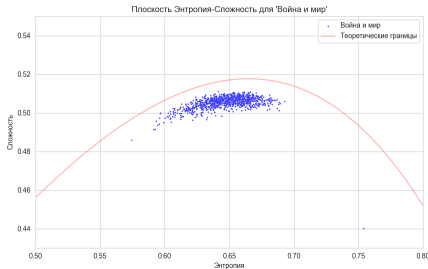


Экспериментальное исследование

- Был собран датасет шуток, содержащий около 90 тыс. анекдотов на разную тематику с различных источников.
- Эмбединги были получены через методы SVD и CBOW, словари эмбедингов были взяты в лаборатории.



С помощью метода ложных соседей были оценены размеры z -векторов. Результат вычислений совпал со значениями, которые использовались в статье [2], авторы в итоге рассматривали значения $n = 2, m = 7 - 8$, а так же $n = 3, m = 4 - 5$. Остальные значения уходили либо в шум, либо в детерминированные процессы. Сначала был воспроизведен результат из самого исследования →

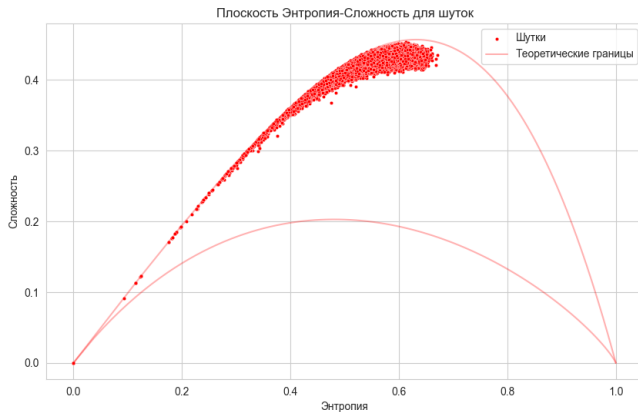


Расположение "Война и мир" на плоскости Энтропия-Сложность



Экспериментальное исследование

Плоскость Энтропия-Сложность

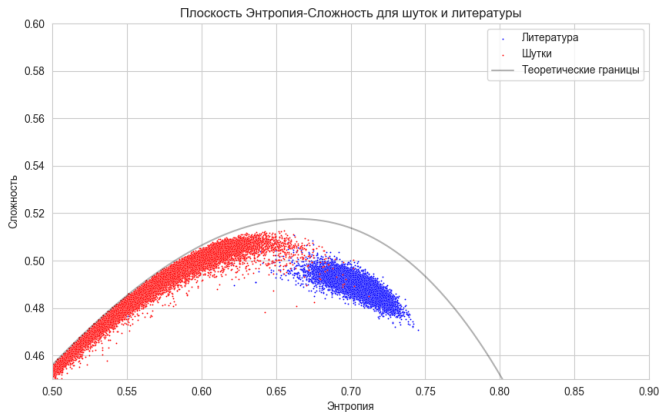


Расположение шуток на плоскости Энтропия-Сложность



Экспериментальное исследование

Плоскость Энтропия-Сложность



Расположение шуток на плоскости Энтропия-Сложность



Данная работа имеет большой потенциал для дальнейших исследований. Семантическое пространство не было рассмотрено должным образом из-за возникших сложностей. Были подготовлены данные для кластеризации, изучена библиотека t-SNE. Можно расширить методологию за счет большего количества уникальных языковых данных. В рамках этого было предпринято участие в создании словаря эмбедингов для двух языков: греческого и иврита.



Процесс создания словарей включал три основных этапа:

- Поиск корпуса
Литературные произведения на греческом и иврите.
- Лемматизация
Были рассмотрены такие популярные библиотеки, как `spacy`, `spark`, `ctlk`, `tree tagger`, `trankit`. В итоге выбор был сделан в пользу Stanza³.
- Построение эмбедингов
SVD, CBOW, Skip-gram

Вычисления производились на суперкомпьютере ВШЭ. Этот результат будет полезен не только дальнейшему исследованию, но и, например, НУГ Поймай бота, которые смогут использовать эти словари в своих исследованиях.

³<https://stanfordnlp.github.io/stanza/>








Благодарность

Исследование выполнено с использованием суперкомпьютерного комплекса
НИУ ВШЭ. [6]



Библиография

-  V. A. Gromov and A. S. Kogan, "Spot the bot: Coarse-Grained Partition of Semantic Paths for Bots and Humans," Feb. 2024.
arXiv:2402.17392 [cs].
-  V. A. Gromov and Q. N. Dang, "Semantic and sentiment trajectories of literary masterpieces," *Chaos, Solitons & Fractals*, vol. 175, p. 113934, Oct. 2023.
-  O. A. Rosso, H. A. Larrondo, M. T. Martin, A. Plastino, and M. A. Fuentes, "Distinguishing Noise from Chaos," *Physical Review Letters*, vol. 99, p. 154102, Oct. 2007.
-  F. Takens, "Detecting strange attractors in turbulence," in *Dynamical Systems and Turbulence, Warwick 1980* (D. Rand and L.-S. Young, eds.), vol. 898, pp. 366–381, Berlin, Heidelberg: Springer Berlin Heidelberg, 1981.
Series Title: Lecture Notes in Mathematics.
-  D. Wishart, "Numerical Classification Method for deriving Natural Classes," *Nature*, vol. 221, pp. 97–98, Jan. 1969.



github.com/DogeSavior3/jokes_course_work