УДК 004.048

ИССЛЕДОВАНИЕ КУЛЬТУРНЫХ ОСОБЕННОСТЕЙ СТРАН С ПОМОЩЬЮ КОНТЕКСТУАЛЬНЫХ ПРЕДСТАВЛЕНИЙ СЛОВ БОЛЬШИХ ЯЗЫКОВЫХ МОДЕЛЕЙ НА ПРИМЕРЕ МОДЕЛЕЙ, ОБУЧЕННЫХ НА КОРПУСАХ РОССИЙСКИХ И АМЕРИКАНСКИХ ГАЗЕТ

В.А. Сергеев
Институт проблем управления им. В.А. Трапезникова РАН
Россия, 117997, Москва, Профсоюзная ул., 65
E-mail: sergeev.bureau@gmail.com

Ключевые слова: большая языковая модель, контекстуальное векторное представление, Берт, культурные особенности.

Аннотация: в рамках работы по данной теме, были обучены две большие языковые модели, одна на корпусе публикаций из американских газет, другая на корпусе публикаций из российских газет. В качестве базовых векторов для сравнения были использованы названия шкал карты культурных ценностей Инглхарта. Для каждой из моделей произведено сравнение векторных представлений отобранных терминов на основе косинусной меры.

1. Введение

В последнее время большие языковые модели все чаще используются как инструмент в решении повседневных задач [16]. Однако вместе с этим, большие языковые модели находят применение в качестве инструмента для проведения исследований [3, 17]. Например, существуют подходы, в рамках которых большие языковые модели используются в качестве персон или агентов, обладающих характерными особенностями [14]. Появляются работы использующие большие языковые модели для задач исследования общества [2].

Традиционные методы оценки культурных особенностей опираются на опросы и интервью. Например, Р. Инглхарт на основе опросов оценивал страны и размещал их на карте культурных ценностей см., например [11]. В предложенном докладе рассматривается вопрос применения такого инструмента как большая языковая модель к задаче выявления культурных особенностей рассматриваемого общества. Для данной задачи были с нуля обучены две большие языковые модели, одна на основе публикаций из ведущих российских газет, другая на основе публикаций из ведущих американских газет. Газеты выбраны в качестве источника текстов для построения моделей, с одной стороны, за широкий охват освещаемых тем, с другой за наличие рецензирования, что предполагает некоторый уровень значимости публикуемой новости, в отличии, например от постов в соц. сетях. В данной работе использовались газетные публикации

с 2015 по 2019 год, поэтому, построенные модели отражают только события имевшие место быть в эти годы.

2. Данные

В работе использованы два корпуса данных доступных для некоммерческого использования.

Корпус публикаций из американских газет на английском языке. В качестве базы, использован корпус А. Томпсона «All the news 2.0» [6], содержащий публикации из ведущих американских газет с 2016 по 2020 гг. Из него были удалены публикации 2020 года, добавлены публикации 2015 года из источника «All the news» [8], удалены газеты с низким уровня доверия см., например [9], такие как «Buzzfeed», а также удалены публикации из источников посвященных узким тематикам, например, такие как «Refinery 29». В корпусе для обучения модели были оставлены только публикации за 2015 – 2019 годы.

Корпус публикаций из российских газет на русском языке был сформирован на основе корпуса И. Гусева [7]. В него также были добавлены некоторые публикации с сайтов ria.ru и lenta.ru. В корпусе для обучения модели были оставлены только публикации за 2015 – 2019 годы.

Для обучения каждой из моделей было отобрано по 1 млн. публикаций, а для валидации по 100 тысяч публикаций.

3. Метод

В отличии от более ранних моделей типа word2vec или Fasttext векторные представления слов в архитектуре Берт – контекстуальны, т.е. зависят от окружающих слов в предложении см., например [4]. Архитектура Берт отлично зарекомендовала себя в широком спектре задач, модели на ее основе позволили получить улучшение на многих бенчмарках, относительно доминировавших прежде моделей на основе статического векторного пространства [1, 13].

Настройка параметров модели производилась в соответствии с рекомендациями из [5, 15]. В частности, размер словаря равен 32768, а максимальная длина входной последовательности 128 токенов. В качестве источника векторных представлений использованы значения с последних четырех слоев модели. Для текстов каждого из имеющихся корпусов были с нуля обучены отдельные модели. В соответствии с процедурой обучения, слова из исследуемого корпуса текстов, токенизируются и векторизуются. Векторы для представления слов называются эмбеддингами или векторными представлениями. В соответствии с дистрибутивной гипотезой см., например [12], две лингвистические единицы, встречающиеся схожем контексте, имеют тенденцию находиться рядом в векторном пространстве модели, а расстояния между соответствующими им векторными представлениями отражают меру их семантической похожести. В качестве меры сходства двух векторов используется мера косинусного сходства. Для двух векторов A, B размерности n, косинусное сходство вычисляется как:

$$S_c(A,B) = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \cdot \sqrt{\sum_{i=1}^{n} B_i^2}}.$$

Для иллюстрации возможности выявления культурных особенностей исследуемых сообществ или стран, было рассмотрено косинусные сходство между векторными представлениями для ценностей, положенных Р. Инглхартом в основание

классификации стран: традиции, рациональность, выживание, самореализация (базовые векторы), и векторными представлениями для слов «Россия» и «США», заданными с соответствующим контекстом. Определения для шкал взяты из [10] и переведены на русский язык. Для слов «Россия» и «США» использованы широко распространённые описания.

- Традиционные ценности подчеркивают важность религии, связей между родителями и детьми, уважения к власти и традиционных семейных ценностей. Люди, которые принимают эти ценности, также отвергают развод, аборты, эвтаназию и самоубийство. Эти общества имеют высокий уровень национальной гордости и националистических взглядов.
- Секулярно-рациональные ценности имеют противоположные предпочтения традиционным ценностям. Эти общества уделяют меньше внимания религии, традиционным семейным ценностям и авторитету. Развод, аборт, эвтаназия и самоубийство считаются относительно приемлемыми. (Суицид не обязательно более распространен.).
- Ценности выживания – особое внимание уделяется экономической и физической безопасности. Это связано с относительно этноцентричным мировоззрением и низким уровнем доверия и толерантности.
- Ценности самовыражения – высокий приоритет защите окружающей среды, растущей толерантности к иностранцам, геям и лесбиянкам и гендерному равенству, а также растущим требованиям участия в принятии решений в экономической и политической жизни.
- США или Соединенные Штаты Америки, широко известные как Соединенные Штаты или Америка, – это страна, расположенная в основном в Северной Америке. США состоит из 50 штатов, федерального округа, пяти крупных некорпоративных территорий, девяти малых отдаленных островов и включает 326 индейских резерваций. Граничит по суше с Канадой на севере и с Мексикой на юге, а также имеет морские границы с рядом других стран. С населением более 334 миллионов человек. Национальная столица Соединенных Штатов — Вашингтон, округ Колумбия, а самый густонаселенный город и главный финансовый центр – Нью-Йорк.
- Россия или Российская Федерация – это страна, охватывающая Восточную Европу и Северную Азию. Это самая большая страна в мире по площади, охватывающая одиннадцать часовых поясов. Столица страны и крупнейший город – Москва. Санкт-Петербург – второй по величине город и культурная столица России. Население страны составляет 146447424 человека. Официальным языком на всей территории страны является русский. Другие крупные города страны включают Новосибирск, Екатеринбург, Нижний Новгород, Челябинск, Красноярск, Казань, Краснодар и Ростов-на-Дону.

4. Результаты

Для каждой из обученных моделей, были получены значения косинусной меры между рассматриваемыми векторами, соответствующими словам «Россия», «США» и четырьмя базовыми векторами. На рис. 1 представлены результаты сравнения того, как «Россия» и «США» соотносятся с базовыми векторами на основе российских и американских публикаций.

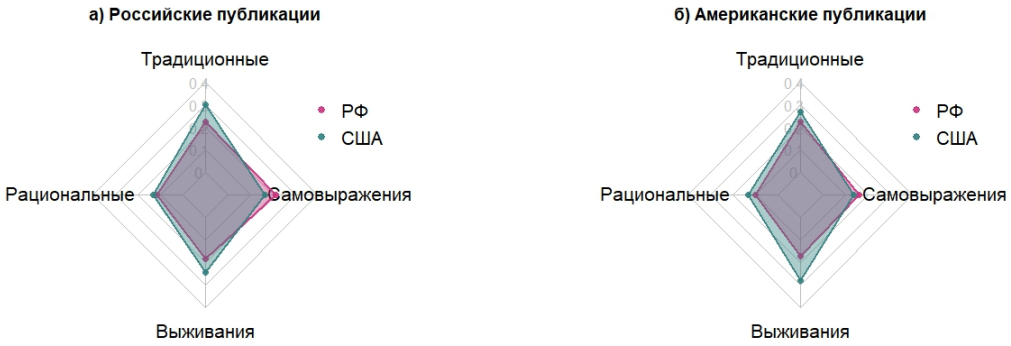

Рис. 1. Значения меры косинусной похожести для слов «Россия» и «США» с базовыми векторами, а) для модели обученной на российских публикациях; б) для модели обученной на американских публикациях.

Из приведенных рисунков видно, что в российских и американских газетах Америка освещается как страна, более ориентированная на традиционные ценности и ценности выживания, в формулировках из [10]. Россия несколько выделяется в сторону ценностей самовыражения. Следует также отметить высокую степень схожести полученных результатов для двух моделей, обученных на различных данных. Из результатов видны полученные противоречия с картой культурных ценностей Рональда Инглхарта седьмой волны (2017-2022 гг.), согласно которой в США ценности самовыражения преобладают над ценностями выживания, а в России рациональные ценности преобладают над традиционными.

5. Заключение

Проведенные сравнения векторных представлений демонстрируют возможность использования больших языковых моделей в качестве источника информации о культурных особенностях. В качестве развития данного исследования, представляется актуальным проведение сравнения на моделях обученных на корпусах содержащих большее количество публикаций.

Список литературы

1. Ann B., Duyen N., Tu T., Weimer L., Jannidis F. To BERT or not to BERT – Comparing Contextual Embeddings in a Deep Learning Architecture for the Automatic Recognition of four Types of Speech, Thought and Writing Representation // SwissText/KONVENS. 2020.
2. Argyle L.P., Busby E.C., Fulda N., Gubler J.R., Rytting C., Wingate D. Out of One, Many: Using Language Models to Simulate Human Samples // Political Analysis. 2023. Vol. 31, No. 3. P. 337-351. doi:10.1017/pan.2023.2.
3. Bubeck S., Chandrasekaran V., Eldan R., Gehrke J.A., Horvitz E., Kamar E., Lee P., Lee Y.T., Li Y., Lundberg S.M., Nori H., Palangi H., Ribeiro M., Zhang Y. Sparks of Artificial General Intelligence: Early experiments with GPT-4. ArXiv abs/2303.12712. 2023.
4. Devlin J., Chang M., Lee K., Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding // Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2019. Vol. 1. P. 4171-4186.
5. Geiping J., Goldstein T. Cramming: Training a Language Model on a Single GPU in One Day // ICML'23: Proceedings of the 40th International Conference on Machine Learning. 2023. P. 11117-11143.
6. https://components.one/datasets/all-the-news-2-news-articles-dataset (дата обращения 15.01.2024).
7. https://huggingface.co/datasets/IlyaGusev/ru_news (дата обращения 15.01.2024).
8. https://www.kaggle.com/datasets/snapcrack/all-the-news (дата обращения 15.01.2024).

9. https://www.pewresearch.org/journalism/2020/01/24/u-s-media-polarization-and-the-2020-election-a-nation-divided/ (дата обращения 15.01.2024).
10. https://www.worldvaluessurvey.org/WVSContents.jsp (дата обращения 15.01.2024).
11. Inglehart R., Welzel C. Modernization, Cultural Change and Democracy: The Human Development Sequence. New York: Cambridge University Press, 2005. 344 p.
12. Jurafsky D., Martin J. Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition. NJ.: Prentice Hall PTR, 2000. 934 p.
13. Korogodina O., Koulichenko V., Karpik O., Klyshinsky E. Evaluation of Vector Transformations for Russian Static and Contextualized Embeddings // GraphiCon 2020 – Proceedings of the 30th International Conference on Computer Graphics and Machine Vision. 2021. Vol. 2. P. 349-357.
14. Kovač, G., Sawayama, M., Portelas, R., Colas, C., Dominey, P.F., Oudeyer, P. Large Language Models as Superpositions of Cultural Perspectives. ArXiv abs/2307.07870. 2023.
15. Peter I., Berchansky M., Levy O. How to Train BERT with an Academic Budget // Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021. P. 10644-10652.
16. Trichopoulos G., Konstantakis M., Alexandridis G., Caridakis G. Large Language Models as Recommendation Systems in Museums // Electronics. 2023. Vol. 12, No. 18. P. 3829. https://doi.org/10.3390/electronics12183829.
17. Xiang D., Bashlovkina V., Han F., Baumgartner S., Bendersky M., What do LLMs Know about Financial Markets? A Case Study on Reddit Market Sentiment Analysis // Companion Proceedings of the ACM Web Conference 2023.