ВЕРОЯТНОСТНО-СМЫСЛОВЫЕ МОДЕЛИ ОЦИФРОВАННЫХ ТЕКСТОВЫХ ДАННЫХ

А. С. Собчишен, А. В. Звягинцева

Донецкий государственный университет, Россия
E-mail: sobchishien@bk.ru, zviagintsevaav@gmail.com

Анализируются модели преобразования текстов в векторные представления, а также использование соответствующих векторных пространств при изучении вероятностной природы смыслов слов в русском языке. Применяемый для этого подход основан на статистическом анализе группы слов, содержащих смысл, по отношению к аналогичной группе бессмысленных слов. Слова соответствующих групп представлялись в виде математических объектов в сформированном трехмерном вероятностном пространстве оцифрованных слов, где последовательности букв алфавита была поставлена в соответствие последовательность целых чисел. Слова со смыслом взяты из филологического словаря, бессмысленные слова образованы с помощью генераторов случайных чисел. В качестве критерия оценки вероятностной природы смыслов использована статистическая вероятность положения слова в пространстве оцифрованных слов. Получены статистические распределения для изучаемых групп слов из трех букв. Результаты работы подтверждают возможность построения вероятностно-смысловых моделей оцифрованных текстовых данных. Показано, что в словах со смыслом наблюдаются статистические закономерности, определяющие местоположение отдельных букв в словах.

PROBABILISTIC SEMANTIC MODELS OF DIGITIZED TEXTUAL DATA

A. S. Sobchishen, A. V. Zviagintseva

The article analyzes models for transforming texts into vector representations, as well as the use of corresponding vector spaces in studying the probabilistic nature of word meanings in the Russian language. The approach used for this purpose is based on the statistical analysis of a group of words containing meaning in relation to a similar group of meaningless words. The words of the corresponding groups were represented as mathematical objects in the formed three-dimensional probabilistic space of digitized words, where a sequence of letters of the alphabet was put in correspondence with a sequence of integers. Words with meaning were taken from a philological dictionary, meaningless words were formed using random number generators. The statistical probability of a word's position in the space of digitized words was used as a criterion for assessing the probabilistic nature of meanings. Statistical distributions were obtained for the studied groups of three-letter words. The results of the work confirm the possibility of constructing probabilistic-semantic models of digitized text data. It is shown that statistical patterns are observed in words with meaning, determining the location of individual letters in words.

Одной из ключевых задач анализа лингвистических данных является изучение вероятностной природы смыслов в языковых единицах. Оптимальный подход получения вероятностных моделей в данном случае связан с построением количественных статистических моделей значимых единиц языка. Начинать такой анализ следует с установления вероятностных характеристик однородных

групп слов естественных языков. Основная гипотеза исследования в этом случае будет связана с тем, что в словах со смыслом могут наблюдаться скрытые статистические закономерности, которые объясняются существованием условных вероятностей, характеризующих местоположение отдельных букв в словах. Для эффективной обработки лингвистических данных следует использовать различные методы оцифровки слов или текстов в различные цифровые или векторные представления. Это позволит применить апробированные статистические методы обработки количественных данных. Рассмотрим некоторые способы представления текстов математическими объектами, а также возможности построения многомерных вероятностных пространств в этом случае.

Преобразование текстов в векторные представления в моделях ИИ.

Имеются три наиболее популярные модели искусственного интеллекта, используемые для анализа текста и преобразования его в векторные представления: *Word2Vec, GloVe* и *BERT*. Эти модели позволяют извлекать семантическую информацию из текстов, оптимально решая разнообразные задачи обработки естественного языка (NLP) [1–4].

Модель *Word2Vec* реализует метод для создания векторных представлений слов, разработанный Google. По архитектуре Word2Vec содержит модель нейронной сети для обучения Continuous Bag of Words (CBOW), которая позволяет распознавать текущее слово на основе окружающих его слов, а также нейросетевой алгоритм обучения Skip-gram, позволяющий предсказывать соседние слова, исходя из текущего слова.

Модель *GloVe* (Global Vectors for Word Representation) – это алгоритм, который создает векторные представления на основе статистики совместного появления слов в больших текстовых корпусах. Техника векторизации текстов основана на матрице совместного появления слов, когда весовые матрицы становятся векторами слов после обучения.

Модель *BERT* (Bidirectional Encoder Representations from Transformers) по архитектуре представляет собой двунаправленную модель на базе трансформеров. В процессе реализации техники векторизации осуществляется токенизация (разбиение текста на токены), проводится двунаправленное обучение (Masked Language Model), когда осуществляется замена некоторых токенов на модель MASK и предсказание замаскированных токенов, используя контекст текста слева и справа.

Все три модели предназначены для создания векторных представлений как слов, так и целых текстов: *Word2Vec* использует нейронные сети для получения векторных представлений; *GloVe* оптимизирует матрицу совместного появления слов, чтобы получить векторные представления; *BERT* применяет двунаправленные трансформеры для создания представлений слов и токенов.

Размерность векторных представлений может различаться в зависимости от модели: *Word2Vec* и *GloVe* часто используют вектора размером 300 измерений. Это связано с тем, что 300-мерные вектора предоставляют хороший баланс между размером и точностью; *BERT* использует вектора размером 768 измерений

для модели BERT-base, но их можно уменьшить до 300 измерений для улучшения вычислительной эффективности.

Другие методы, модели и технологии оцифровки текстов, например, *Bag-of-Words*, *TF-IDF*, *RNN* и *LSTM*, *Text-to-Speech*, *Audio-Signal Processing* и т.п. учитывают частоту появления слов в документах или зависимости между элементами текста, применяют синтактический анализ текста и семантические сети, используют преобразования текстов в речевой или звуковой сигнал и т.д. [3 –7].

Все перечисленные выше методы и модели создают векторные представления, которые жестко привязаны к контексту анализируемых текстовых корпусов, используют различные методы и алгоритмы оцифровки текстов, которые не являются универсальными, формируют векторы очень большой размерности, часто применяют эвристические методы и т.д. Получаемые векторные представления слов сложно интерпретировать как координаты многомерных пространств и, в связи с чем, применять апробированные математические методы для анализа оцифрованных данных.

Представление слов в дискретных пространствах цифровых объектов.

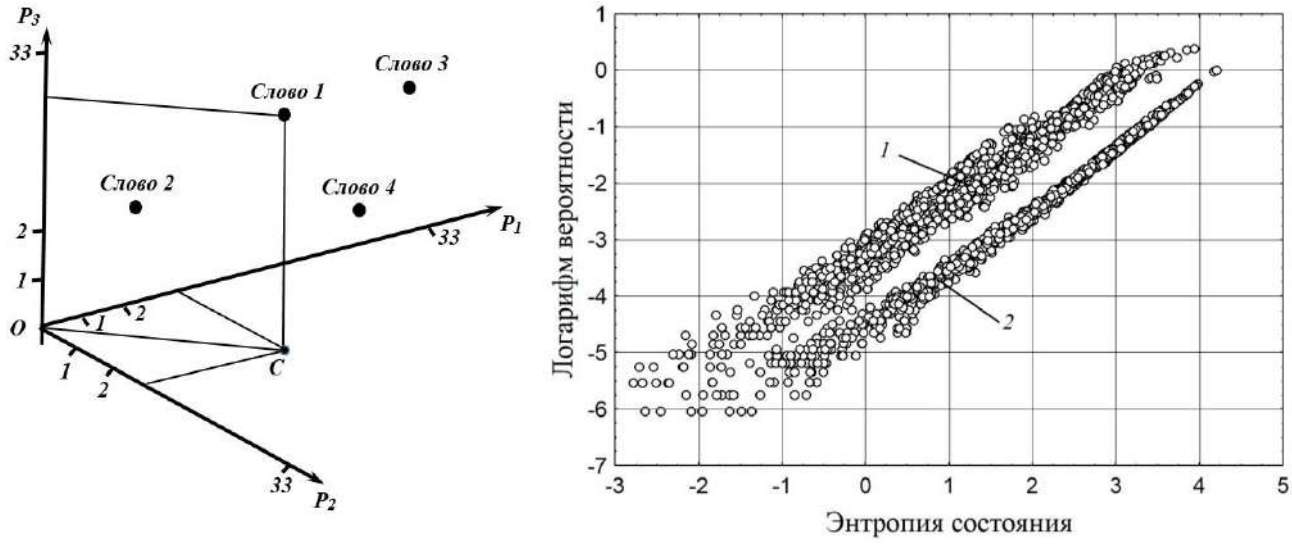
Слова, состоящие из определенного количества букв, могут быть бессмысленными или иметь определенный смысл. Если рассматривать словари из определенно заданного количества букв, то их можно представить как лингвистические системы, где каждое слово (как несущее смысл, так и бессмысленное) является состоянием такой системы. В математическом плане такой системе может быть поставлено в соответствие некоторое пространство состояний.

Если каждому состоянию присвоить некую меру, например, заданную вероятность состояния – вероятность наблюдения слова в определенной области пространства, то можно говорить о существовании вероятностного пространства состояний. Используем метод оцифровки слов, который позволяет получить многомерные векторные модели слов небольшой размерности и найти их вероятностные распределения [8].

По аналогии с результатами работы [8] рассмотрим однородную группу слов из одинакового количества букв, например, из трех букв, которые содержатся в филологическом словаре. Местоположение букв в слове свяжем с тремя координатными осями пространства состояний. В свою очередь, местоположение букв в алфавите определим дискретной последовательностью целых чисел 1, 2, 3…, 33, где букве «а» соответствует число каждой координатной оси, равное 1, букве «б» – число 2 и т.д., букве «я», соответственно, – число 33. Для наглядности координатные оси трехмерного пространства состояний примем в такой последовательности: первая буква слова – это абсцисса, вторая буква – ордината и третья буква – аппликата. Последовательность оцифровки слов представлена на рис., а. Событием наблюдения слова в заданном объеме пространства будем считать совместное событие одновременного выбора целых значений оцифрованных букв, которые формируют слово.

Статистическую вероятность данного события будем оценивать алгоритмически путем применения вычислительных алгоритмов [9, 10]. В трехмерном пространстве состояний представим две группы одинаковых по количеству слов

– группу слов со смыслом и группу бессмысленных слов. Для этого из филологического словаря сформируем массив из 1264 слов со смыслом, состоящих из трех букв. В качестве опорного слова примем наиболее часто встречаемое слово в этой группе – «год». Группу бессмысленных слов образуем с помощью генераторов случайных чисел с равномерным распределением.



а) б)

Сравнение слов по их смыслосодержанию:
а) схема представления вероятностного пространства состояний;
б) статистическая модель групп слов со смыслом и без смысла:
1 – слова со смыслом; 2 – бессмысленные слова

Путем обработки данных для группы из трех букв со смыслом было получено статистическое распределение вида

$$w_s = -3{,}277 + s;\ s = 0{,}899 \ln\left(\frac{x_1}{x_{1_0}}\right) + 0{,}657 \ln\left(\frac{x_2}{x_{2_0}}\right) + 0{,}999 \ln\left(\frac{x_3}{x_{3_0}}\right). \qquad (1)$$

Аналогично для группы из трех букв без смысла

$$w_b = -4{,}457 + s;\ s = 0{,}905 \ln\left(\frac{x_1}{x_{1_0}}\right) + 0{,}812 \ln\left(\frac{x_2}{x_{2_0}}\right) + 0{,}906 \ln\left(\frac{x_3}{x_{3_0}}\right). \qquad (2)$$

Здесь w_s, w_b – статистическая вероятность наблюдения слов со смыслом и бессмысленных слов; s – энтропия состояния (слова); x_k – значения оцифрованных букв для различных слов; x_{k_0} – значения оцифрованных букв опорного слова. Результаты обработки данных представлены на рисунке 1, б.

Из приведенных данных видно, что в словах со смыслом наблюдаются статистические закономерности, определяющие местоположение отдельных букв в словах. Если бессмысленные слова распределены в пространстве состояний абсолютно равновозможно, то слова, содержащие смысл, распределены по неравновозможным статистическим законам. Аналогичные результаты при оценке вероятностных распределений для слов из четырех букв были получены в работе

[8], в которой предложен метод оценки смыслосодержания в оцифрованных словах и способы создания шкал для семантических измерений.

Таким образом, существующие гипотезы о вероятностной природе смыслов и наличии вероятностно-смысловых моделей естественных и искусственных языков, например [11], могут быть обоснованы путем статистического анализа лингвистических данных и построения вероятностных моделей языковых явлений.

СПИСОК ЛИТЕРАТУРЫ

1. *Naseem U., Razzak I., Khan S. K., & Prasad M.* A Comprehensive Survey on Word Representation Models: From Classical to State-of-the-Art Word Representation Language Models // ACM Transactions on Asian and Low-Resource Language Information Processing, 2021. Vol. 20 (5).

2. *Elov B. B., Kamroeva Sh. M., Alayev R. H. [et al.]* Methods of processing the uzbek language corpus texts // International Journal of Open Information Technologies. 2023. Vol. 11. No 12. P. 143-151.

3. *Савенков П. А., Ивутин А. Н.* Методы анализа естественного языка в задачах детектирования поведенческих аномалий // Известия Тульского государственного университета. Технические науки. 2022. № 3. С. 358-366.

4. *Zhiwei F.* Formal Analysis for Natural Language Processing // Springer Nature. 2023. 796 p.

5. Handbook of Research on Natural Language Processing and Smart Service Systems. Pazos-Rangel, Rodolfo Abraham, Florencia-Juarez, Rogelio, Paredes-Valverde, Mario Andrés, Rivera, Gilberto // IGI Global. 2020. 554 p.

6. *Частикова В. А., Козачек К. В., Гуляй В. Г.* Методы обработки естественного языка в решении задач обнаружения атак социальной инженерии // Вестник Адыгейского государственного университета. Серия 4: Естественно-математические и технические науки. 2021. № 4 (291). С. 95-108.

7. *Xiong Z., Shen Q., Xiong Y., Wang Y., Li W.* New generation model of word vector representation based on CBOW or skip-gram // Computers, Materials and Continua. 2019. Vol. 60 (1).

8. *Аверин Г.В.* О вероятностной природе смыслов в дискретных языковых единицах // Системный анализ и информационные технологии в науках о природе и обществе, №1(12)–2(13), 2017. С. 11–18.

9. *Аверин Г. В.* Системодинамика: теория и приложения / Изд. 2-е перераб. и доп. Донецк: ООО «НПП «Фолиант», 2022. 535 с.

10. *Звягинцева А. В.* Вероятностные методы комплексной оценки природно-антропогенных систем / Под науч. ред. д.т.н., проф. Г. В. Аверина. М. : Изд. дом «Спектр», 2016. 258 с.

11. *Налимов В. В.* Вероятностная модель языка. О соотношении естественных и искусственных языков. 2-е изд., перераб. и доп. М. : Наука, 1979. 303 с.