УДК 004

Наумов Максим Денисович, магистрант, Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара

Борисова Светлана Павловна, доцент кафедры математики и бизнес-информатики, Самарский национальный исследовательский университет имени академика С.П. Королева, г. Самара

МЕТОД ГИБРИДНОГО АНАЛИЗА СЕМАНТИЧЕСКОГО СХОДСТВА ТЕКСТОВ С ПРИМИНЕНИЕМ TF-IDF И ЭМБЕДДИНГОВ НЕЙРОННЫХ СЕТЕЙ

Аннотация

В работе описан комбинированный алгоритм автоматизированного текстового анализа, предназначенный для определения семантической близости текстовых данных. Предлагаемый подход сочетает традиционную статистическую модель TF-IDF и современные методы нейросетевого представления текста в виде эмбеддингов. Такая интеграция позволяет не только сопоставлять текстовые фрагменты между собой, но и формализованно измерять степень их смыслового сходства.

Повышение точности достигается за счёт объединения анализа частотных характеристик лексики с контекстно-зависимым моделированием семантики, реализуемым нейронными сетями. Алгоритм включает несколько последовательных этапов: предварительную обработку текстового материала, преобразование данных в векторное пространство и вычисление евклидовой метрики для оценки расстояния между полученными семантическими представлениями.

Annotation

The paper describes a combined algorithm for automated text analysis designed to determine the semantic similarity of textual data. The proposed approach integrates the traditional statistical TF-IDF model with modern neural network–based text representations in the form of embeddings. This integration makes it possible not only to compare textual fragments but also to formally quantify the degree of their semantic similarity.

Improved accuracy is achieved through the combination of frequency-based lexical analysis and context-dependent semantic modeling implemented by neural networks. The algorithm consists of several sequential stages, including text preprocessing, transformation of data into a vector space, and the computation of the Euclidean metric to estimate the distance between the resulting semantic representations.

Ключевые слова: обработка текста, оценка смысловой близости текстовых данных, евклидово расстояние, обработка естественного языка.

Keywords: text processing, semantic similarity assessment, Euclidean distance, natural language processing.

Современные информационные платформы функционируют в условиях постоянного роста объёма текстовой информации, что обуславливает актуальность задач автоматизированного определения семантического сходства текстов [1, 2]. Подобные задачи возникают при поиске релевантных документов, тематической кластеризации, обработке пользовательских запросов, а также при сопоставлении описаний, выраженных разными языковыми средствами, но передающих сходное содержание.

Традиционные методы обработки текста, базирующиеся на анализе частотных характеристик лексических единиц, отличаются относительной простотой реализации и прозрачностью интерпретации результатов. Однако такие подходы ограниченно отражают контекстуальные и смысловые связи. Нейросетевые модели представления текста, напротив, способны учитывать семантические зависимости, но требуют значительных вычислительных

ресурсов и могут демонстрировать чувствительность к шумовым данным. В связи с этим целесообразным представляется объединение статистических и нейросетевых инструментов в рамках интегрированной методики.

Разработанный способ оценки семантической близости текстовой информации реализуется в несколько этапов. Сначала проводится предварительная обработка данных: удаление служебных символов, приведение текста к единому регистру, токенизация, а при необходимости — лемматизация. Для обеспечения сопоставимости материалов допускается автоматический перевод текстов на общий язык.

Далее выполняется преобразование текстов в векторное пространство с использованием модели TF-IDF. Каждый документ представляется в виде числового вектора, отражающего значимость терминов с учётом их частотности в конкретном тексте и распространённости в корпусе. Это позволяет оценить лексическую близость сравниваемых материалов.

Одновременно формируются нейросетевые векторные представления. С этой целью применяется предобученная модель SentenceTransformer, преобразующая предложения и документы в плотные векторы фиксированной размерности. Полученные эмбеддинги инкапсулируют контекст и семантические связи между словами.

Сравнение текстов в обоих векторных пространствах осуществляется посредством вычисления евклидовой метрики. Сокращение расстояния между соответствующими векторами свидетельствует о большей степени смысловой близости. Итоговый показатель сходства формируется путём агрегирования значений, полученных как на основе TF-IDF, так и на основе нейросетевых эмбеддингов.

Метод TFIDF позволяет представить текст в виде вектора весов терминов, вычисляемых по формуле [1]:

$$TFIDF(t, d, D) = TF(t, d) \times IDF(t, D) \quad (1)$$

где: TF(t,d) – частота термина t в документе d, IDF(t,d) – обратная частота документа в корпусе D.

Данный метод позволяет выделять лексически значимые элементы текста, однако не обеспечивает полноценного учета смысловой эквивалентности различных формулировок.

Векторные представления, полученные с использованием нейросетевых моделей, базируются на трансформерных архитектурах и ориентированы на моделирование семантических и контекстных зависимостей между словами и устойчивыми выражениями [3–5]. В результате тексты, передающие близкое содержание при различающемся наборе лексических единиц, отображаются в таком пространстве на минимальном расстоянии друг от друга.

Количественная мера сходства векторов определяется с применением евклидовой метрики, которая вычисляется следующим образом:

$$D(A, B) = \qquad (2)$$

где: A и B – сравниваемые векторы.

В отличие от метрик, основанных на анализе угла между векторами, евклидова мера учитывает абсолютные различия координат и позволяет трактовать сходство как степень пространственной близости объектов в многомерном признаковом пространстве.

Интегральный показатель семантической близости формируется посредством нормирования и последующего объединения расстояний, рассчитанных как для TF-IDF представлений, так и для нейросетевых эмбеддингов. Такой механизм обеспечивает сбалансированное учёт лексической составляющей и глубинных смысловых связей.

Программная реализация алгоритма выполнена на языке Python. Для построения TF-IDF векторов и вычисления расстояний применяются инструменты библиотеки scikit-learn, тогда как формирование нейросетевых представлений осуществляется с использованием пакета sentence-transformers. Численные операции и работа с многомерными массивами реализованы средствами NumPy, что обеспечивает эффективную обработку высокоразмерных векторов.

На вход системы поступает корпус текстовых документов, который проходит этапы предварительной очистки и преобразования в векторное пространство. Затем для каждой пары текстов вычисляется евклидова дистанция в соответствующих представлениях, после чего формируется сводная матрица семантического сходства.

Апробация предложенного метода проводилась на текстовых выборках различной тематики и структуры. Результаты экспериментов продемонстрировали, что комбинированный подход обеспечивает более стабильные и точные оценки по сравнению с применением исключительно TF-IDF либо только нейросетевых моделей.

Наиболее выраженный эффект наблюдается при анализе текстов, содержащих перефразированные конструкции и синонимичные выражения. В таких ситуациях нейросетевые представления нивелируют ограничения статистической модели, тогда как TF-IDF сохраняет чувствительность к значимым терминам.

Таким образом, предложена интегрированная методика оценки семантической близости текстовой информации, основанная на совместном использовании TF-IDF и нейросетевых эмбеддингов с применением евклидовой метрики. Подход позволяет учитывать как лексические признаки, так и контекстуально-смысловые характеристики, обеспечивая более комплексное и надежное сопоставление текстов.

Дальнейшее развитие исследования предполагает адаптацию методики к узкоспециализированным предметным областям, а также анализ влияния различных стратегий агрегирования расстояний на итоговую оценку семантического сходства.

Литература

- Ермаков А. С., Кузнецов Д. Ю. Анализ эффективности статистических и нейросетевых подходов к измерению семантического сходства текстов // Информационные технологии и вычислительные системы. 2022. Вып. 3.

- Астафьев С. И., Тихомиров И. А. Комбинированные методы автоматизированной обработки текстовой информации: интеграция частотных признаков и контекстных эмбеддингов // Вестник компьютерных и информационных технологий. 2023. № 5 (215).
- Королев М. С. Использование трансформерных моделей типа BERT для выявления смысловой близости русскоязычных документов // Программная инженерия. 2021. Том 12, № 4.
- Васильев В. Г., Петров А. Н. Применение математического аппарата оценки расстояний в многомерных векторных пространствах при анализе текстов // Системы управления, связи и безопасности. 2024. № 1.
- Бондаренко И. С. Развитие методов векторного представления лексики: от статистических моделей к крупным языковым системам (LLM) // Вопросы искусственного интеллекта. 2023. № 2.

Literature

- Ermakov A. S., Kuznetsov D. Yu. Analysis of the effectiveness of statistical and neural network approaches to measuring semantic similarity of texts. *Information Technologies and Computing Systems*. 2022. Issue 3.
- Astafiev S. I., Tikhomirov I. A. Hybrid methods for automated text processing: integration of frequency-based features and contextual embeddings. *Bulletin of Computer and Information Technologies*. 2023. No. 5 (215).
- Korolev M. S. Application of transformer-based BERT models for detecting semantic similarity of Russian-language documents. *Software Engineering*. 2021. Vol. 12, No. 4.
- Vasiliev V. G., Petrov A. N. Application of mathematical methods for distance evaluation in multidimensional vector spaces in text analysis. *Systems of Control, Communication and Security*. 2024. No. 1.

- Bondarenko I. S. Evolution of vector representations of words: from statistical models to large language models (LLMs). *Issues of Artificial Intelligence*. 2023. No. 2.