DOI 10.37539/2949-1991.2025.35.12.019

Акулов Динис Наилевич, магистрант,
Автономная некоммерческая организация высшего образования «Российский новый университет»
Akulov Dinis Nailevich,
Autonomous Non-commercial Organisation of Higher Education “Russian New University”

МОДЕЛИ ПРЕДСТАВЛЕНИЯ СМЫСЛА ТЕКСТА В ЗАДАЧАХ АВТОМАТИЧЕСКОГО СРАВНЕНИЯ ДОКУМЕНТОВ
MODELS OF TEXT MEANING REPRESENTATION IN AUTOMATIC DOCUMENT COMPARISON

Аннотация. В современную эпоху автоматическое сравнение текстов имеет критически важное значение. Ключевой задачей при сравнение текстов является представление текста в числовом формате, понятном для машины. В данной статье анализируется эволюция методов представления текстов – от статистических подходов до современных контекстуальных моделей.

Abstract. In the modern era, automated text comparison is critically important. A key challenge in text comparison is representing text in a machine-readable numerical format. This article analyzes the evolution of text representation methods, from statistical approaches to modern contextual models.

Ключевые слова: Представление смысла текста, семантическая схожесть, обработка естественного языка, сравнение документов, NLP.

Keywords: Text meaning representation, semantic similarity, natural language processing, document comparison, NLP.

В последние годы объем текстовых данных в мире растет с беспрецедентной скоростью. В связи с этим многим организациям, а также частным лицам необходимо эффективно обрабатывать, анализировать и сопоставлять различные данные. Поэтому на сегодняшний день особую актуальность приобретает автоматическое сравнение документов, которое лежит в основе множества критически важных приложений таких как: информационный поиск, выявление плагиата, группировка документов (например, новостей) по темам, системы рекомендаций и т.д. Центральная задача, объединяющая подобные приложения, – это количественная оценка семантической (смысловой) схожести между двумя или более единицами текста. Стоит отметить, что машины не могут напрямую работать с необработанным текстом; им необходимо числовое представление, которое отражает смысл текста. Качество этого представления напрямую определяет успешность любой системы сравнения документов.

История развития моделей представления смысла текста – это путь от простых подсчетов к глубоким, контекстно-зависимым моделям, которые способны улавливать смысловые нюансы [1]. Так, ранние подходы к представлению смысла текста были основаны на частотном распределении слов и статистическом анализе. Модель «Мешок слов» (Bag of Words или BoW) является одним из самых простых и фундаментальных подходов. Суть BoW заключается в представлении текстового документа в виде вектора, где каждое измерение соответствует уникальному слову из совокупности всех документов, а значением является частота встречаемости слова в данном документе. При данном подходе сравнение 2 документов сводится к вычислению метрики схожести между их векторами (например,

косинусного расстояния). Если векторы указывают в одном направлении, это означает, что документы используют схожий набор слов с похожей частотой.

Однако у такого подхода есть ряд значимых ограничений:

1. Игнорирование порядка слов. Так, BoW полностью теряет информацию о структуре предложения.

2. Проблема синонимии. Используя BoW невозможно понять, что слова «автомобиль», «авто» и «машина» имеют схожий смысл. Для нее это 3 разных, не связанных измерения в векторе.

3. Разреженность и размерность. Векторы получаются очень большими и разреженными, что крайне неэффективно для вычислений.

Модель TF-IDF (Term Frequency – Inverse Document Frequency) является дальнейшим развитием BoW. Она также создает вектор частот, однако взвешивает каждое слово, чтобы оценить его важность. TF – это частота слова в документе, а IDF – это логарифмическая мера того, насколько редко слово встречается во всех документах данного корпуса.

Таким образом, TF-IDF отлично подходит для поиска информации и выделения ключевых слов. При сравнении документов данная модель придает больший вес словам, которые являются уникальными и важными для данного текста, и меньший – общеупотребительным словам.

Несмотря на большую эффективность по сравнению с BoW, TF-IDF также не решает ряд фундаментальных проблем: она игнорирует порядок слов и не имеет представления о различных семантических связях (синонимах, антонимах).

Революция в обработке естественного языка произошла с появлением векторных представлений слов (word embeddings). Вместо разреженных векторов в тысячи измерений, эти модели представляют слова в виде плотных векторов из нескольких сотен измерений. Ключевая идея заключается в том, что слова, встречающиеся в схожих контекстах, должны иметь близкие векторы.

Word2Vec и GloVe – это два наиболее популярных подхода к созданию статических представлений [2]. Word2Vec использует «неглубокую» нейронную сеть для предсказания либо слова по его контексту (модель CBOW), либо контекста по слову (модель Skip-gram). GloVe, в свою очередь, строится на основе глобальной матрицы совместной встречаемости слов во всем корпусе. Данные модели успешно определяют семантические отношения. Например, они позволяют выполнять векторную арифметику:

1. Вектор «король» – вектор «мужчина» + вектор «женщина» ≈ вектор «королева».

2. Вектор «Москва» – вектор «Россия» + вектор «Япония» ≈ вектор «Токио»

Для сравнения ряда документов (а не отдельных слов) необходимо агрегировать векторы всех слов документа. Простой способ агрегирования – это усреднение векторов или взвешенное усреднение, где в качестве весов используются значения TF-IDF.

Тем не менее, и у данного подхода есть свои ограничения:

1. У каждого слова есть только один вектор, независимо от контекста. Вектор слова «лук» будет одинаковым в предложениях «мой сын стрелял из лука» и «тетя резала лук».

2. Усреднение векторов – это метод, который часто «размывает» истинный смысл документа.

3. Модели не могут обрабатывать редкие слова, которых не было в обучающем словаре.

Модель FastText решила проблему работы с редкими словами, определяя векторы не для целых слов, а для n-грамм символов, из которых состоят слова. В связи с этим FastText очень полезен для сравнения документов, содержащих опечатки, редкие слова, жаргонизмы или неологизмы, так как модель может сгенерировать вектор даже для неизвестного слова из его составных частей [3].

Однако наиболее важный прорыв в представление смысла текста связан с появлением контекстуальных моделей, которые были основаны на архитектуре Transformer. В отличие от статических представлений, эти модели генерируют динамический вектор для каждого слова, который зависит от его окружения в предложении.

ELMo (Embeddings from Language Models) стал одним из первых популярных контекстуальных подходов. Он использует двунаправленные нейронные сети (bi-LSTM) для анализа контекста слева и справа, создавая представление, которое меняется в зависимости от предложения. Хотя ELMo и стала прорывной моделью, тем не менее, революцией в обработке естественного языка была BERT. Эта модель использует архитектуру Transformer для одновременного анализа всего текста в обе стороны (глубокая двунаправленность). Так, BERT обучается на двух задачах:

1. Masked Language Model: модель предсказывает случайно «спрятанные» (замаскированные) слова в предложении.
2. Next Sentence Prediction: модель определяет, является ли второе предложение логическим продолжением первого.

Хотя BERT можно использовать для сравнения текстов, данный подход вычислительно очень затратен. Поэтому для задач семантического сравнения был разработан Sentence-BERT (SBERT). SBERT использует сиамскую (siamese) архитектуру, где два документа/предложения независимо прогоняются через один и тот же BERT, а затем их выходные векторы сравниваются (например, с помощью косинусной схожести). SBERT обучен так, что семантически близкие предложения получают в векторном пространстве близкие векторы [4]. На сегодняшний день SBERT стал стандартом для задач семантической схожести. Он сочетает глубокое контекстуальное понимание BERT с эффективностью, необходимой для масштабного сравнения.

Однако и трансформеры сталкиваются с определенными вызовами:

1. Трансформеры требуют значительных вычислительных ресурсов для работы.
2. Модели, обученные на общих текстах (например, на текстах Википедии), могут плохо работать со специализированными документами (медицинскими, юридическими). В связи с этим возникает потребность в доменно-специфичных моделях (к примеру, BioBERT для биомедицинских текстов).
3. Модели, обученные на созданных человеком текстах, наследуют предвзятости (гендерные, расовые и т.д.), что необходимо учитывать при их применении.

Итак, эволюция моделей представления текста прошла путь от простых частотных счетчиков слов до сложных нейронных архитектур, способных улавливать контекст и смысловые нюансы. Для задач автоматического сравнения документов это означает переход от сопоставления по ключевым словам к глубокому семантическому сопоставлению текстов.

Если для простых задач (например, группировка документов по темам) все еще могут быть достаточны TF-IDF, FastText, то для сложных и требующих высокой точности программ (выявление плагиата, семантический поиск) стандартом стали контекстуальные модели на основе трансформеров. На практике же выбор модели представления смысла текста для задачи автоматического сравнения документов зависит от баланса между требуемой точностью и доступными вычислительными ресурсами.

Будущие исследования, вероятно, будут сосредоточены на повышении эффективности трансформеров, их адаптации к мультимодальным данным (сравнение текста с изображениями, аудио или видео), улучшении кросс-языкового сравнения и решении проблем, связанных с предвзятостью данных.

Список литературы:

1. Danish S. M. H. Comparative Analysis of BERT and TF-IDF for Textual Semantic Similarity Assessment / S. M. H. Danish, S. M. E. Hasnain, H. Ashraf, R. Rukaiya // 2024 26th International Multi-Topic Conference (INMIC). – 2024. – P. 1-6.

2. Rakshit P. A supervised deep learning-based sentiment analysis by the implementation of Word2Vec and GloVe Embedding techniques / P. Rakshit, A. Sarkar // Multimedia Tools and Applications. – 2025. – Vol. 84, No. 2. – P. 979-1012.

3. Rana A. Semantic Similarity Analysis using FastText / A. Rana, A. Pant, N. Rawat, P. Rawat, S. Vats, V. Sharma // 2024 IEEE 3rd World Conference on Applied Intelligence and Computing (AIC). – 2024. – P. 454-460.

4. Li R. Siamese bert architecture model with attention mechanism for textual semantic similarity / R. Li, L. Cheng, D. Wang, J. Tan // Multimedia Tools and Applications. – 2023. – Vol. 82, No. 30. – P. 46673-46694.