УДК 004.82

АКТУАЛЬНЫЕ ПРОБЛЕМЫ АССИМЕТРИЧНОГО СЕМАНТИЧЕСКОГО ПОИСКА НА ОСНОВЕ МНОГОМЕРНЫХ ВЕКТОРНЫХ ПРЕДСТАВЛЕНИЙ ТЕКСТОВЫХ ДАННЫХ

Рябинкин Г.М., аспирант

Белорусский государственный университет информатики и радиоэлектроники
г. Минск, Республика Беларусь

Гецевич Ю.С. – канд. технич. наук

Аннотация. В статье рассматриваются проблемы и подходы к семантическому поиску, основанному на методах глубокого обучения и векторном представлении текстовых данных. Особое внимание уделяется асимметричному семантическому поиску, когда запрос и документы различаются по объему и лексике. Анализируются методы формирования векторных пространств, выбор единиц кодирования (слова, предложения, документы) и метрики семантической близости. Рассматриваются проблемы, возникающие при выборе модели и способа обучения, а также влияние обучающего корпуса на качество результатов поиска.

Ключевые слова. Семантический поиск, векторный поиск, обработка естественного языка, глубокое обучение, векторное сходство, семантическое сходство, векторное представление текста.

Повсеместная цифровизация, охватывающая все сферы общественной жизни вынуждает на постоянной основе манипулировать значительными объемами текстовых данных. Одной из важнейших проблем при работе с информацией становится ее эффективный поиск, позволяющий быстро найти релевантные сведения среди огромных массивов данных. С ростом количества доступной для обработки информации классические поисковые системы полнотекстового поиска перестают справляться с возложенными на них задачами, в первую очередь, при организации взаимодействия в системе «человек-машина», где пользователь ожидает от системы не примитивный поиск по количеству вхождений лексем из поискового запроса, а релевантную выборку на основе семантического сходства с пользовательским запросом [1]. В данной связи, все большую актуальность приобретают подходы, основанные на семантическом поиске, базой технической реализации которых выступают методы глубокого обучения в области обработки естественного языка [2]. Ключевая особенность рассматриваемых методов – модели представления семантической информации на основе многомерных векторов [3].

В настоящее время, несмотря на широкое распространение семантического поиска на основе многомерных векторных представлений текстовых данных, в названном подходе существует ряд проблем, требующих тщательного рассмотрения и решения. Целью работы является осуществление общего обзора задач и подходов, применяемых в рамках семантического поиска текстовой информации и выявление актуальных проблем в данной области. Актуальность исследования обусловлена стремительным ростом объемов доступной текстовой информации, что делает классические поисковые системы менее эффективными. Современные задачи информационного поиска требуют более интеллектуального подхода, основанного на семантическом анализе текстов.

Научная новизна работы заключается в комплексном анализе задач и методов семантического поиска, с особым акцентом на асимметричный семантический поиск. Исследование направлено на выявление существующих ограничений и проблем в данной области, а также рассмотрение перспективных техник построения семантического векторного пространства и методов оценки семантической близости фрагментов текста. В отличие от традиционных подходов, в данной работе акцент сделан на рассмотрение асимметричных поисковых моделей, рассматриваются особенности работы с специализированными корпусами текстов, что позволяет глубже понять влияние названных факторов на релевантность извлекаемых данных и заложить основу для дальнейших исследований, направленных на улучшение качества взаимодействия между пользователем и интеллектуальной системой поиска.

Семантический поиск предполагает определение релевантности информации поисковому запросу на основе семантической близости, т.е. сходства контекстуального значения и общего смысла запроса с сопоставляемой единицей информации, а не на основе количества вхождений ключевых слов и лексических соответствий запроса и документа [4]. Семантические поисковые системы позволяют решать значительно более широкий спектр задач, предоставлять эргономичный интерфейс взаимодействия с пользователем, исключать нерелевантные результаты из поисковой выборки, эффективно работать с синонимами, сокращениями, орфографическими ошибками, обобщениями и т.д. [1].

Для реализации систем семантического поиска может использоваться комбинация целого ряда походов, включающая в себя приемы, основанные на лексическом сходстве, оценке частотности

упоминания ключевых терминов, тезаурусах (идеографических словарях), онтологиях (графах знаний). Значительный прогресс в данной области был достигнут в свете применения достижений методов глубокого обучения для обработки естественного языка. Так, использование моделей глубокого обучения для формирования векторного представления слов и предложений, задействование рекуррентных нейронных сетей и трансформеров позволило значительно продвинуться при решении классических задач обработки естественного языка, таких как классификация, установление кореферентности и восполнение отсутствующей информации. Все это позволяет говорить и качественных успехов применения методов глубокого обучения для выявления семантики обрабатываемых текстовых данных, что, несомненно, нашло широкое применение при решении задач семантического поиска [5].

Именно векторное представление текстовых данных в пространстве, сформированном в результате обучения моделей глубокого обучения, является представлением семантической информации содержащейся в тексте [6]. Так, слова, появляющиеся в схожем контексте в корпусе текстов, на котором происходило обучение модели, будут представлены векторами, находящимися на меньшем удалении друг от друга в рассматриваемом векторном пространстве.

Вместе с этим, существуют различные способы формирования векторных пространств семантических представлений, которые показывают неодинаковую эффективность при решении задач семантического поиска. В первую очередь, на способность выделения семантического сходства слов или частей текста влияет задача, в рамках которой происходило обучение модели глубокого обучения [7]. Так, модель, предназначенная для задач классификации текста, будет успешно определять семантическую близость между словами, которые позволяют отнести текст к конкретному классу. Например, модель, обученная соотносить диагноз (класс) со списком симптомов расположит вектора, представляющие симптомы чаще встречающиеся для одних и тех же классов, на меньшем удалении друг от друга [8]. В тоже время, если задачей модели будет восполнение недостающей информации – предсказание случайным образом пропущенных слов из описания симптомов, то она может в меньшей степени учитывать сходство симптомов между собой, и будет стремиться к восстановлению пропущенных данных, опираясь на вероятностные связи между словами.

Другим важным аспектом формирования векторных представлений является архитектура модели. Например, трансформернные модели, такие как BERT или SBERT (Bidirectional Encoder Representations from Transformers, Sentence-BERT), используют механизм самообращенного внимания, позволяющий учитывать контекст всей последовательности слов [9]. Это делает их эффективными в задачах анализа значений слов в зависимости от окружающих элементов текста и позволяет улучшить способность модели к обобщению в ходе обучения [10]. Однако модели на основе рекуррентных нейронных сетей (Recurrent Neural Network, RNN) или долгосрочной краткосрочной памяти (Long-Short Term Memory, LSTM) могут иметь преимущества при обработке последовательных данных, где важен порядок слов.

Также значительное влияние на качество семантических представлений оказывает объем и качество обучающего корпуса. Чем разнообразнее и насыщеннее данные, тем более точными и релевантными становятся векторные пространства, созданные моделью. В последнее время активно развиваются методы дообучения (fine-tuning) [11], которые позволяют адаптировать предварительно обученные модели, уже имеющие высокую обобщающую способность, к конкретным задачам, улучшая их способность выявлять семантические связи исходя из контекста, заданного конкретным корпусом текстов, что особенно актуально для задач семантического поиска на ограниченном объеме текстовых данных, относящихся к узкой области профессиональной деятельности (например, базы знаний, основанные на медицинских данных, корпусе юридических текстов, финансовых документов и т.п.).

Таким образом, выбор модели и способа обучения играет ключевую роль в эффективности семантического поиска, определяя, насколько точно алгоритм сможет различать значения и контексты слов. Кроме того, необходимо учитывать масштаб и специфику обучающего корпуса. Более разнородные данные из разных предметных областей, позволяют формировать универсальные семантические представления. В то же время модели, обученные на узкоспециализированных данных, могут демонстрировать высокую эффективность в контексте конкретных доменов, например, медицинских или юридических текстов. В свою очередь, тонкая настройка и дообучение на дополнительных данных или применение методов усиленного обучения (Reinforcement Learning) помогает корректировать векторные пространства семантических представлений, делая их более точными и релевантными.

При выборе способа кодирования текстовых данных в векторные представления перед исследователями возникают вопросы выбора размерности вектора и определения базовой единицы кодирования – слово, предложение, некоторая произвольно выбранная смысловая единица или даже документ целиком. Выбор размерности вектора и базовой единицы кодирования играет решающую роль в эффективности представления текстовых данных. От этого зависит, насколько точно модель сможет улавливать смысловые связи и проводить поиск схожих текстов.

Если в качестве базовой единицы кодирования выбрать слово, это позволит учитывать лексическое значение каждого элемента текста, но может ограничивать возможность анализа более

сложных смысловых конструкций. Также, такой подход не позволяет справиться с проблемами наличия в языке многозначных и синонимичных слов, которые могут встречаться как в корпусе текстов, так и в поисковом запросе. Данная проблема особенно актуальна при работе с текстами узкой направленности, где, как правило, используется весьма ограниченный словарь, а сами слова встречаются в составе множества различных по значению составных терминов.

Кроме того, кодирование каждой лексической единицы текста отдельным вектором требует значительных затрат на носители информации, так как многомерные вектора содержат в себе на порядки больше информации, чем исходная лексическая единица. Данную проблему можно обойти, учитывая наличие взаимно однозначного соответствия между единицей сформированного словаря и ее векторным представлением, однако, необходимость многократного кодирования текста в векторные представления при каждой операции поиска делает такой подход требовательным к вычислительным ресурсам и в данной связи малоприменимым для решения задач семантического поиска.

Использование предложения в качестве единицы кодирования дает преимущество в обработке смысловых конструкций, позволяя учитывать контекст на более высоком уровне. При кодировании на уровне смысловых единиц, таких как фразы или тематические блоки, можно достичь еще большей гибкости в обработке текста, но этот метод требует дополнительной разметки данных и более сложных алгоритмов кластеризации.

Помимо этого, при выборе в качестве единицы кодирования достаточно больших блоков исходного документа возможна проблема «усреднения» семантического представления фрагмента текста. Так, наложение друг на друга векторов большого количества семантических единиц посредством вычисления среднего значения или пулинга будет способствовать потере семантических особенностей единицы кодирования, что может снизить релевантность поисковой выборки [12].

Наконец, кодирование целого документа в единое векторное представление актуально для задач, связанных с категоризацией и классификацией текстов, но еще больше способно нивелировать нюансы семантического представления информации, теряя на этапе объединения множества векторов их специфические особенности. Такие модели применяются, например, в рекомендательных системах и анализе больших объемов текстовой информации но ввиду названных особенностей могут хуже справляться с задачами семантического поиска по поисковым запросам пользователей [13].

Особенно остро эта проблема проявляется при решении задач ассиметричного семантического поиска – когда поисковый запрос и единица поиска (документ) несопоставимы между собой по объему информации и/или диапазону используемой лексики. Примером такого сценария может служить система семантического поиска юридических документов для пользователей, не имеющих соответствующей экспертизы. В данном случае очевидны несоответствие объема пользовательского запроса и искомого документа, а также высокая вероятность использования в запросе некорректных терминов или разговорных выражений, которые необходимо поставить в соответствие строгим формулировкам юридических текстов из базы знаний.

Оптимальный выбор единицы кодирования зависит от конкретных задач, стоящих перед поисковой системой, и может определяться структурой текстов базы знаний, особенностями их разметки, общим объемом корпуса текстов.

Еще одним фактором, оказывающим влияние на эффективность поиска, может являться способ определения семантической близости между векторными представлениями семантической информации в сформированном векторном пространстве. Так, среди наиболее распространенных подходов можно выделить метрики, основанные на сходстве и различии векторов, а также дистанции между ними в многомерном пространстве [14].

Так, коэффициент косинусового подобия векторов, позволяет установить сходство между ними, выраженное скаляром в диапазоне от -1 до 1, основываясь на угле между векторами в многомерном пространстве. При этом, на значение полученного коэффициента не оказывает влияния абсолютное значение векторов. В случае же выражения сходства через скалярное произведение, для случая ненормализованных векторов, на результат будут оказывать влияния не только угол, но и абсолютное значение векторных представлений рассматриваемых семантических единиц. Еще более чувствительным к абсолютным значениям будут результаты вычислений, основанных на евклидовом расстоянии между векторами.

Выбор метрики выражения сходства векторных представлений во многом зависит от итогового сформированного векторного пространства и того, какую в нем роль играет угол и абсолютное значение векторов для их семантической близости, задач, стоящих перед поисковой системой (мультимодальный поиск, ассиметричный текстовый поиск, рекомендательная система поиска схожих документов и т.д.).

Таким образом, семантический поиск, основанный на многомерном векторном представлении текстовых данных и методах глубокого обучения, открывает широкие возможности для эффективной обработки и анализа текстовой информации. Однако, для достижения высокого качества поиска необходимо учитывать ряд факторов, включая выбор модели, способа обучения, единицы кодирования и метрики семантической близости. Особое внимание следует уделять асимметричному семантическому

поиску, где различия между запросом и документами требуют применения специализированных подходов.

В рамках дальнейших исследований в данной области планируется осуществить адаптацию более эффективных методов обучения моделей к задачам семантического поиска, изучить проблемы применения методов поиска к различным предметным областям и узкоспециализированным корпусам текстов, исследовать и формализовать критерии выбора основных подходов к формированию семантического векторного пространства, выбору размерности векторов и используемых метрик. Решение этих задач позволит создать интеллектуальные системы поиска, способные эффективно выделять семантику текста и предоставлять наиболее релевантные результаты поиска в различных сценариях, в том числе в рамках ассиметричного поиска и работы с узкоспециализированными корпусами текстов.

Список использованных источников:

1. Pre-training Tasks for Embedding-based Large-scale Retrieval / W.-C. Chang [et al.] // 8th International Conference on Learning Representations, ICLR 2020 / ed. by A. Rush [et al.]. – Addis Ababa, Ethiopia : OpenReview.net, 2020. – P. 12.

2. Bantsevich, K. Integration of Large Language Models with Knowledge Bases of Intelligent Systems / K. Bantsevich [et al.] // Open semantic technologies for intelligent systems (OSTIS) : сборник научных трудов / BSUIR. – Minsk, 2023. – Vol. 7 – P. 213–218.

3. Embeddings for Efficient Literature Screening: A Primer for Life Science Investigators / C. Galli, C. Cusano, S. Guizzardi [et al.] // Metrics. – 2024. – Vol. 1, № 1. – P. 1.

4. Kasenchak, R. T. What is Semantic Search? And why is it important? / R. T. Kasenchak // Information Services and Use, 2019. – Vol. 39, № 3. – P. 205–213.

5. Yurchuk, I. Extending Monolingual Asymmetric Semantic Search Models For Multilingual Query Processing Using Knowledge Distillation / I. Yurchuk, D. Boiko // Proceedings of the Information Technology and Implementation (IT&I) Workshop: Intelligent Systems and Security (IT&I-WS 2024: ISS) / ed. by V. Snytyuk [et al.]. – Kyiv, Ukraine : CEUR Workshop Proceedings, 2024. – P. 1–10.

6. Wang, L. Improving Text Embeddings with Large Language Models / L. Wang [et al.] // ACL 2024 : Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) / ed. by L.-W. Ku, A. Martins, V. Srikumar. – Bangkok, Thailand : Association for Computational Linguistics, 2024. – P. 11897–11916.

7. Mahmoud, A. F. A. Enhancing Semantic Search Precision through the CBOW Algorithm in the Semantic Web / A. F. A. Mahmoud [et al.] // Engineering, Technology & Applied Science Research, 2025. – Vol. 15, № 1. – P. 19522–19527.

8. Zubiaga, A. Exploiting Class Labels to Boost Performance on Embedding-based Text Classification / A. Zubiaga // CIKM '20: Proceedings of the 29th ACM International Conference on Information & Knowledge Management / ed. by M. d'Aquin [et al.] event-place: Virtual Event, Ireland. – New York, NY, USA : Association for Computing Machinery, 2020. – P. 3357–3360.

9. Walsh, H. S. Semantic Search With Sentence-BERT for Design Information Retrieval / H. S. Walsh, S. R. Andrade // ASME 2022 International Design Engineering Technical Conferences and Computers and Information in Engineering Conference : International Design Engineering Technical Conferences and Computers and Information in Engineering Conference (Volume 2: 42nd Computers and Information in Engineering Conference (CIE)). – American Society of Mechanical Engineers Digital Collection, 2022. – P. 9.

10. Fujishiro, N. Accuracy of the Sentence-BERT Semantic Search System for a Japanese Database of Closed Medical Malpractice Claims / N. Fujishiro, Y. Otaki, S. Kawachi // Applied Sciences, 2023. – Vol. 13, № 6. – P. 4051.

11. Kroshchanka, A. Kroshchanka, A. Reduction of neural network models in intelligent computer systems of a new generation / A. Kroshchanka // Otkrytye semanticheskie tehnologii proektirovanija intellektual'nyh sistem [Open semantic technologies for intelligent systems], 2023. – Vol. 7 – P. 127–132.

12. Ai, Q. Analysis of the Paragraph Vector Model for Information Retrieval / Q. Ai, L. Yang, J. Guo, W. B. Croft // Proceedings of the 2016 ACM International Conference on the Theory of Information Retrieval : ICTIR '16 / ed. by B. Carterette [et al.] event-place: Newark, Delaware, USA. – New York, NY, USA : Association for Computing Machinery, 2016. – P. 133–142.

13. Riyadh, M. LLM-assisted Vector Similarity Search / M. Riyadh [et al.] arXiv:2412.18819 [cs]. – arXiv, 2024.

14. Pan, J. J. Survey of vector database management systems / J. J. Pan, J. Wang, G. Li // The VLDB Journal, 2024. – Vol. 33, № 5. – P. 1591–1615.

UDC 004.82

ACTUAL PROBLEMS OF ASYMMETRIC SEMANTIC SEARCH BASED ON MULTIDIMENSIONAL VECTOR REPRESENTATIONS OF TEXT DATA

Rabinkin H.M.

Belarusian State University of Informatics and Radioelectronics, Minsk, Republic of Belarus

Hetsevich Y.S. – PhD in Technical Sciences

Annotation. The article considers problems and approaches to semantic search based on deep learning methods and vector representation of text data. Particular attention is paid to asymmetric semantic search, when the query and documents differ in volume and vocabulary. The methods of forming vector spaces, the choice of encoding units (words, sentences, documents) and semantic similarity metrics are analyzed. The problems arising in the choice of a model and training method, as well as the influence of the training corpus on the quality of search results are considered.

Keywords. Semantic search, vector search, natural language processing, deep learning, vector similarity, semantic similarity, text embeddings.