MINISTRY OF EDUCATION AND RESEARCH OF THE REPUBLIC OF MOLDOVA

Technical University of Moldova

Faculty of Computers, Informatics, and Microelectronics

Department of Software Engineering and Automation

Approved for defense
Department head:
Ion FIODOROV, phd, associate professor

„___” ________ 2025

IMPROVING VECTOR SEARCH IN RETRIEVAL-AUGMENTED GENERATION

Master's project

Student:	_____________	**Morari Gheorghe, IS-231M**
Coordinator:	_____________	**Gavrilița Mihail, university lecturer**
Consultant:	_____________	**Cojocaru Svetlana, university assistant**

Chișinău, 2025

ABSTRACT

This thesis explores how semantic operators in a semantic algebra can improve vector search in Retrieval Augmented Generation (RAG) scenarios. Semantic operators manipulate the meaning of words and sentences in a formal, mathematical way, enabling operations that allow for precise comparison and manipulation of meaning. These operators are applied practically through the use of Large Language Models (LLMs), which decompose, manipulate, and recompose the meaning of words and sentences. The decomposed and tagged sentences are then transformed into embeddings that facilitate the search for relevant sources within a vector space, which will allow for more control over the search process and the retrieval of more relevant sources.

REZUMAT

Această teză explorează modul în care operatorii semantici într-o algebră semantică pot îmbunătăți căutarea vectorială în scenarii de Generare Augmentată prin Regăsire (RAG). Operatorii semantici manipulează sensul cuvintelor și propozițiilor într-un mod formal și matematic, permițând operații care oferă posibilitatea de a compara și manipula precis sensul. Acești operatori sunt aplicați practic prin utilizarea modelelor de limbaj mari (LLM), care descompun, manipulează și recompun sensul cuvintelor și propozițiilor. Propozițiile descompuse și etichetate sunt apoi transformate în embedding-uri care facilitează căutarea surselor relevante într-un spațiu vectorial, ceea ce va permite un control mai mare asupra procesului de căutare și regăsirea unor surse mai relevante.

CONTENTS

INTRODUCTION

The field of semantic algebra has long been a theoretical cornerstone in cognitive linguistics and computational semantics. It provides a formal framework for manipulating the meaning of words and sentences using algebraic operators. Historically, the concepts of semantic algebra were ahead of their time, offering a rigorous mathematical model for semantics that could not be fully realized due to the limitations in computational power and data availability.

With the advent of machine learning, particularly the development of Large Language Models (LLMs) and advanced embedding techniques that were developed alongside LLMs, the theoretical constructs of semantic algebra can now be put into practice. These advancements have enabled the creation of high-quality text embeddings that capture the nuanced meanings of words and sentences, making it possible to apply semantic algebra in practical applications.

One of the most promising applications of this synergy between semantic algebra and machine learning is in the domain of vector search. Traditional search engines rely on keyword matching, which often fails to capture the true intent behind a query. In contrast, vector search leverages embeddings to represent the semantic content of both queries and documents in a continuous vector space. By applying semantic algebra to these embeddings, we can perform more sophisticated manipulations and comparisons, leading to more relevant search results.

This thesis explores how the integration of semantic algebra with LLM-generated embeddings can enhance vector search, particularly in Retrieval Augmented Generation (RAG) scenarios. By using semantic operators to refine and manipulate embeddings, the aim is to improve the accuracy and relevance of search results, thereby demonstrating the practical utility of semantic algebra in modern computational linguistics.

BIBLIOGRAPHY

[1] Gao, Tianyu, et al. SimCSE: Simple Contrastive Learning of Sentence Embeddings. arXiv:2104.08821, arXiv, 18 May 2022. arXiv.org, https://doi.org/10.48550/arXiv.2104.08821.

[2] Li, Xianming, and Jing Li. AnglE-Optimized Text Embeddings. arXiv:2309.12871, arXiv, 17 July 2024. arXiv.org, https://doi.org/10.48550/arXiv.2309.12871.

[3] Huang, Zezhou. Disambiguate Entity Matching Using Large Language Models through Relation Discovery. arXiv:2403.17344, arXiv, 29 May 2024. arXiv.org, https://doi.org/10.48550/arXiv.2403.17344.

[4] Wang, Yingxu. "A Semantic Algebra for Cognitive Linguistics and Cognitive Computing." 2013 IEEE 12th International Conference on Cognitive Informatics and Cognitive Computing, 2013, pp. 17–25. IEEE Xplore, https://doi.org/10.1109/ICCI-CC.2013.6622221.

[5] Vaswani, Ashish, et al. Attention Is All You Need. arXiv:1706.03762, arXiv, 2 Aug. 2023. arXiv.org, https://doi.org/10.48550/arXiv.1706.03762.

[6] Templeton, Adly, et al. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." Transformer Circuits Thread, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

[7] Brown, Tom B., et al. Language Models Are Few-Shot Learners. arXiv:2005.14165, arXiv, 22 July 2020. arXiv.org, https://doi.org/10.48550/arXiv.2005.14165.

[8] Schulhoff, Sander, et al. The Prompt Report: A Systematic Survey of Prompting Techniques. arXiv:2406.06608, arXiv, 23 Dec. 2024. arXiv.org, https://doi.org/10.48550/arXiv.2406.06608.

[9] Mikolov, Tomas, et al. Efficient Estimation of Word Representations in Vector Space. arXiv:1301.3781, arXiv, 7 Sept. 2013. arXiv.org, https://doi.org/10.48550/arXiv.1301.3781.

[10] Pennington, Jeffrey, et al. "Glove: Global Vectors for Word Representation." Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, 2014, pp. 1532–43. DOI.org (Crossref), https://doi.org/10.3115/v1/D14-1162.

[11] Lewis, Patrick, et al. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401, arXiv, 12 Apr. 2021. arXiv.org, https://doi.org/10.48550/arXiv.2005.11401.

[12] S. Stratulat, D Prijilevschi, G. Morari, T. Bumbu, A Disambiguation Model for Natural Language Processing. in: Proceedings of the Conference on Mathematical Foundations of Informatics MFOI-2020, January 12-16, 2021, Kyiv, Ukraine, pp.361-381.

[13] G. Kamradt. Needle in a haystack - pressure testing llms. https://github.com/gkamradt/LLMTest_NeedleInAHaystack, 2023.

[14] G. Gerganov. LLM inference in C/C++. https://github.com/ggerganov/llama.cpp, 2024.

[15] Lilian Weng. LLM Powered Autonomous Agents https://lilianweng.github.io/posts/2023-06-23-agent/, 2023.