

CZECH TECHNICAL UNIVERSITY IN PRAGUE
FACULTY OF ELECTRICAL ENGINEERING
DEPARTMENT OF CYBERNETICS
CZECH INSTITUTE OF INFORMATICS, ROBOTICS AND CYBERNETICS



From FastText to Transformer Models, and their Application in Retrieval-Augmented Generation

Bachelor's Thesis

Yauheni Zviazdou

Prague, May 2024

Study programme: Open Informatics
Branch of study: Artificial Intelligence and Computer Science

Supervisor: Ing. Jan Šedivý, CSc.

Acknowledgments

Firstly, I would like to express my gratitude to my supervisor.

I. Personal and study details

Student's name: **Zviazdou Yauheni** Personal ID number: **507333**
Faculty / Institute: **Faculty of Electrical Engineering**
Department / Institute: **Department of Cybernetics**
Study program: **Open Informatics**
Specialisation: **Artificial Intelligence and Computer Science**

II. Bachelor's thesis details

Bachelor's thesis title in English:

From FastText to Transformer Models, and their Application in Retrieval-Augmented Generation

Bachelor's thesis title in Czech:

Od FastText k Transformer model m a jejich aplikace v Retrieval-Augmented generování

Guidelines:

Review the representation of words, sentences, and paragraphs, progressing from traditional methods like FastText to advanced transformer-based models such as BERT. The primary focus is to evaluate selected representations using analogy tests and confusion matrix. Use the UPV corpus set for evaluation.

In the second part of the study, emphasis will shift towards selecting optimal representations for Retrieval-Augmented Generation (RAG) algorithms. The investigation will determine the most efficient embeddings and optimal text chunk size for question-answering tasks, particularly in the context of natural language answers generation from technical manuals. Conduct a comprehensive evaluation with a particular focus on suggesting an optimal representation model that balances factuality and CPU requirements.

Bibliography / sources:

- [1] FastText documentation and tutorials, <https://fasttext.cc/>
- [2] LangChain documentation, <https://js.langchain.com/docs>
- [3] Word2Vec tutorials, <https://www.tensorflow.org/text/tutorials/word2vec>

Name and workplace of bachelor's thesis supervisor:

Ing. Jan Šedivý, CSc. Big Data and Cloud Computing CIIRC

Name and workplace of second bachelor's thesis supervisor or consultant:

Date of bachelor's thesis assignment: **02.01.2024** Deadline for bachelor thesis submission: **24.05.2024**

Assignment valid until: **21.09.2025**

Ing. Jan Šedivý, CSc.
Supervisor's signature

prof. Dr. Ing. Jan Kybic
Head of department's signature

prof. Mgr. Petr Páta, Ph.D.
Dean's signature

III. Assignment receipt

The student acknowledges that the bachelor's thesis is an individual work. The student must produce his thesis without the assistance of others, with the exception of provided consultations. Within the bachelor's thesis, the author must state the names of consultants and include a list of references.

Date of assignment receipt

Student's signature

Declaration

I declare that presented work was developed independently, and that I have listed all sources of information used within, in accordance with the Methodical instructions for observing ethical principles in preparation of university theses.

Date
.....

Abstract

This thesis investigates the application of word and sentence embeddings in Retrieval-Augmented Generation (RAG) for factual Question Answering (QA) tasks using technical manuals. The study explores the effectiveness of traditional FastText embeddings and advanced transformer-based models like Bidirectional Encoder Representations from Transformers (BERT) in capturing semantic relationships within text. We evaluate the quality of these representations using analogy tests and confusion matrix analysis on the UPV corpus set.

Subsequently, we will select optimal representations for RAG algorithms and assess their impact on factual accuracy and computational efficiency during QA. By analyzing the performance with different text chunk sizes, we aim to identify the optimal configuration for factual RAG in technical domains. This research contributes to the field of Natural Language Processing (NLP) by providing insights into selecting effective representations that balance factual accuracy and computational efficiency for QA systems.

Keywords NLP, Word Embedding, Transformers, RAG, QA, Semantic Textual Similarity (STS)

Abstrakt

Tato práce zkoumá aplikaci embeddingů slov a vět v modelu Retrieval-Augmented Generation (RAG) pro úlohy Question Answering (QA) zaměřené na fakta s využitím technických příruček. Studie se zabývá účinností tradičních embeddingů FastText a pokročilých modelů založených na transformátoru, jako je Bidirectional Encoder Representations from Transformers (BERT), při zachycování sémantických vztahů v textu. Kvalitu těchto reprezentací hodnotíme pomocí analogových testů a analýzy confusion matrix na korpusu UPV. Následně vybereme optimální reprezentace pro algoritmy RAG a posoudíme jejich vliv na faktickou přesnost a výpočetní efektivitu během QA. Analýzou výkonu s různými velikostmi textových fragmentů se snažíme identifikovat optimální konfiguraci pro faktické RAG v technických oborech. Výzkum přispívá k oblasti zpracování přirozeného jazyka (Natural Language Processing (NLP)) tím, že poskytuje poznatky o výběru efektivních reprezentací, které vyvažují faktickou přesnost a výpočetní efektivitu pro systémy QA.

Klíčová slova NLP, Word Embedding, Transformátory, RAG, QA, Sémantická podobnost textu

Abbreviations

RAG Retrieval-Augmented Generation

NLP Natural Language Proccession

STS Semantic Textual Similarity

QA Question Answering

BERT Bidirectional Encoder Representations from Transformers

CBOW Continuous Bag-of-Words

ML Machine Learning

NN Neural Network

Contents

1	Introduction	1
2	Literature Review	2
2.1	Traditional word embedding methods	2
2.1.1	Word2Vec [3]	2
2.1.2	GloVe [2]	2
2.1.3	FastText [1]	2
3	Methodology	3
4	Experiments and Results	4
5	Discussion	5
6	Conclusion	6
7	References	7
A	Appendix A	8

■ 1 Introduction

- Briefly introduce the importance of text representation in Natural Language Processing (NLP) tasks.
- Highlight the evolution of text representation methods, from traditional techniques like FastText to advanced transformer-based models like BERT.
- State the research objective: Evaluate different word, sentence, and paragraph representations for Retrieval-Augmented Generation (RAG) algorithms, focusing on technical question answering (QA).

■ 2 Literature Review

There are several methods for word representation.

■ 2.1 Traditional word embedding methods

■ Word2Vec [3]

Word2Vec is algorithm that generates word embedding using information about target word (context). Word2Vec uses Neural Network (NN) and Machine Learning (ML) techniques to generate word embedding for every word in vocabulary during training. As NN architecture are used Continuous Bag-of-Words (CBOW) and Skip-gram, Fig. 2.1.

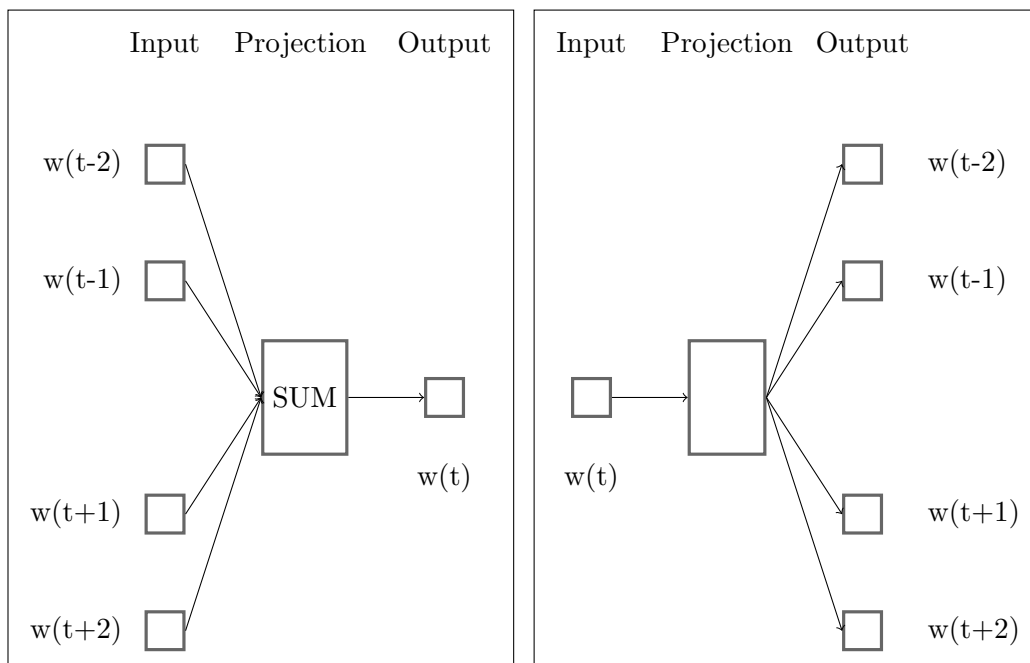


Figure 2.1: CBOW and Skip-gram schemes respectively

■ GloVe [2]

■ FastText [1]

- Discuss traditional word embedding methods like FastText and their limitations.
- Explain the concept of transformer-based models like BERT and their advantages for text representation.
- Review related work on RAG algorithms and their dependence on effective text representations. Discuss existing research on evaluating text representations using analogy tests and confusion matrices.
- Briefly mention the UPV corpus set as the chosen evaluation benchmark.

■ 3 Methodology

- Describe the evaluation process for different text representations.
- - Specify the chosen word, sentence, and paragraph representation models (e.g., Fast-Text, BERT variants).
 - Explain the usage of analogy tests and confusion matrices for evaluation.
 - Detail the selection process for the UPV corpus set and its suitability for technical QA tasks.
- Outline the second part of the study focusing on RAG for technical QA.
- - Explain the RAG algorithm and its reliance on text representations.
 - Describe the evaluation approach for selecting optimal representations for RAG.
 - Mention the factors considered during evaluation, such as embedding efficiency, text chunk size, and factuality of generated answers.

■ 4 Experiments and Results

- Present the results of the evaluation for different text representations using analogy tests and confusion matrices.
- Discuss the findings regarding the effectiveness of each representation model for capturing semantic relationships in technical text.
- Analyze the results from the RAG evaluation, highlighting the impact of different representations and text chunk sizes on answer generation quality and CPU efficiency.
- Identify the representation model that achieves a balance between factuality of answers and computational demands.

■ 5 Discussion

- Interpret the overall findings and their implications for choosing suitable text representations for RAG in technical QA tasks.
- Discuss the strengths and limitations of the chosen evaluation methods.
- Address potential challenges encountered during the study and suggest improvements for future research.

■ 6 Conclusion

Summarize the achieved results. Can be similar as an abstract or an introduction, however, it should be written in past tense.

- Summarize the key takeaways from the research, emphasizing the most effective text representation model for RAG in technical QA based on the evaluation criteria.
- Briefly mention the trade-offs between factuality, CPU usage, and other factors in selecting representations for RAG.
- Suggest potential future research directions, such as exploring other text representation methods or evaluating RAG performance on different datasets.

■ 7 References

- [1] P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov, “Enriching word vectors with subword information,” *arXiv preprint arXiv:1607.04606*, 2016.
- [2] J. Pennington, R. Socher, and C. D. Manning, “Glove: Global vectors for word representation,” in *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. [Online]. Available: <http://www.aclweb.org/anthology/D14-1162>.
- [3] T. Mikolov, K. Chen, G. Corrado, and J. Dean, *Efficient estimation of word representations in vector space*, 2013. arXiv: [1301.3781](https://arxiv.org/abs/1301.3781) [cs.CL].

■ A Appendix A