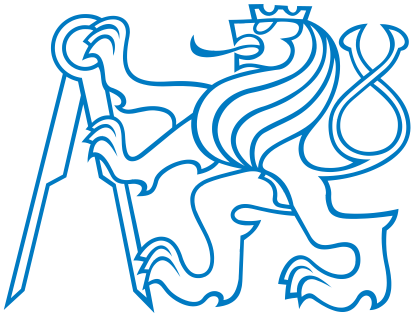
Czech Technical University in Prague

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.

BACHELOR‘S THESIS ASSIGNMENT

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 modelům 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.	prof. Dr. Ing. Jan Kybic	prof. Mgr. Petr Páta, Ph.D.
Supervisor’s signature	Head of department’s signature	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 Procession (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 Procession (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

GPS Global Positioning System

LiDAR Light Detection and Ranging

RAG Retrieval-Augmented Generation

NLP Natural Language Procession

STS Semantic Textual Similarity

QA Question Answering

BERT Bidirectional Encoder Representations from Transformers

Contents

1 Introduction

First, introduce the reader to the research topic. Start with the most general view and slowly converge to the particular field, sub-field, and the challenges you face. You can cite others' work here [1].

1.1 Related works

This section should contain related state-of-the-art works and their relation to the author's work. We usually cite the original works like this [3]. You can also cite multiple papers at once like this [1], [2].

1.2 Contributions

This section should describe the author's contributions to the field of research.

1.3 Mathematical notation

It is a good practice to define basic mathematical notation in the introduction. See Table 1.1 for an example.

$\mathbf{x}$, $\boldsymbol{\alpha}$	vector, pseudo-vector, or tuple
$\hat{\mathbf{x}}$, $\hat{\boldsymbol{\omega}}$	unit vector or unit pseudo-vector
$\hat{\mathbf{e}}_1, \hat{\mathbf{e}}_2, \hat{\mathbf{e}}_3$	elements of the *standard basis*
$\mathbf{X}, \boldsymbol{\Omega}$	matrix
$\mathbf{I}$	identity matrix
$x = \mathbf{a}^\intercal \mathbf{b}$	inner product of $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^3$
$\mathbf{x} = \mathbf{a} \times \mathbf{b}$	cross product of $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^3$
$\mathbf{x} = \mathbf{a} \circ \mathbf{b}$	element-wise product of $\mathbf{a}$, $\mathbf{b} \in \mathbb{R}^3$
$\mathbf{x}_{(n)} = \mathbf{x}^\intercal \hat{\mathbf{e}}_n$	n^{th} vector element (row), $\mathbf{x}, \mathbf{e} \in \mathbb{R}^3$
$\mathbf{X}_{(a,b)}$	matrix element, (row, column)
x_d	x_d is *desired*, a reference
$\dot{x}, \ddot{x}, \dddot{x}, \ddddot{x}$	1^{st}, 2^{nd}, 3^{rd}, and 4^{th} time derivative of x
$x_{[n]}$	x at the sample n
$\mathbf{A}, \mathbf{B}, \mathbf{x}$	LTI system matrix, input matrix and input vector
SO(3)	3D special orthogonal group of rotations
SE(3)	*SO(3)* $\times \mathbb{R}^3$, special Euclidean group

Table 1.1: Mathematical notation, nomenclature and notable symbols.

2 Literature Review

- 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., FastText, 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 How to write thesis in LaTeX

5.1 Versioning with git

Write the LaTeX in such a way that it could be versioned by git, which will help when collaborating with other people. This means writing **one sentence per line**. Even when you use third-party platforms, such as the OverLeaf, you can still share the repository through Git.

5.2 Forming paragraphs

A paragraph is formed in LaTeX by an uninterrupted block of non-empty lines. It is recommended to keep a single sentence per line (helps with versioning using git). A new paragraph is started after an empty line.

This is a new paragraph. It is strongly recommended to **avoid** the use of the *newline* (\\) feature of LaTeX for forming paragraphs as it doesn't format the new paragraph properly (no space at beginning of the new paragraph).

5.3 Linguistic anti patterns

Narrative

We recommend to write your thesis in plural form of the first-person narrative in combination with passive tense, e.g.:

- We discourage the use of any other form, and/or
- any other form is discouraged, but **not**
- I discourage you from using the first-person narrative.

Moreover, avoid "instructional" or "teacher"-like style of writing, such as "Now, we multiply the matrix $\mathbf{A}$ by the scalar c to get the scaled matrix $\mathbf{B}$." A better way of writing the same information would be e.g. "Now, the scaled matrix $\mathbf{B}$ is obtained by multiplying the matrix $\mathbf{A}$ by the scalar c."

Pronouns

The use of pronouns (it, this, they) is strongly **discouraged**. Although, pronouns make it easier for you as a writer to form the flow of the text, pronouns also make it much more difficult for the reader to follow the text. The reader is forced to retain more of the context to substitute and understand what the author meant. Moreover, pronouns can easily become vague (there is more than one way how to interpret them) and can become invalid while making editorial changes to the text, i.e., when moving sentences around. A technical text should be written in a way that makes it as easy to read and comprehend as possible and as hard to misunderstand or misinterpret as possible at the same time.

5.4 Mathematical notation with LaTeX

Take care to use the correct mathematical symbols and common ways of denoting mathematical concepts. Use bold fonts to visually distinguish vectors and matrices ($\mathbf{x}$, $\mathbf{A}$) and

scalars (k, N).

Common errors

A frequent error, carried over from programming languages, is using the asterisk symbol ($*$) to denote multiplication. The asterisk correctly denotes convolution. Similarly, the cross sign ($\times$) typically denotes the cross product (it can also used for stating dimensions, such as $10\,\mathrm{m} \times 10\,\mathrm{m}$) and thus should not be used for scalar multiplication. In English mathematical notation, **scalar multiplication is typically not denoted at all**.

This custom may sometimes make it unclear whether a sequence of letters denotes multiplication of several scalars or a multi-letter variable, such as

$$T = T0 + coeffmeas, \tag{5.1}$$

where the variables in this hypothetical equation are T, $T0$, $coeff$ and $meas$. For this reason, **avoid using multi-letter variable naming** and strive to denote mathematical variables with single letters optionally a with lower or upper index, or other modifiers (`\hat`, `\bar`, etc.). The equation above could be modified to be

$$T = T_0 + cT_{\mathrm{meas}}. \tag{5.2}$$

If the multiplication is still unclear (e.g. when multiplying many single-letter scalars), the `\cdot` symbol may be used such as

$$P \cdot V = n \cdot R \cdot T. \tag{5.3}$$

Equations

Mathematical equations should be numbered and should be a part of a sentence. For example, a discrete LTI system update is described as

$$\mathbf{x}_{[k+1]} = \mathbf{A}\mathbf{x}_{[k]} + \mathbf{B}\mathbf{u}_{[k]}, \tag{5.4}$$

where $\mathbf{x}_{[k]} \in \mathbb{R}^m$ is the state vector at the sample k, $\mathbf{u}_{[k]} \in \mathbb{R}^n$ is the input vector, $\mathbf{A} \in \mathbb{R}^{m \times m}$ is the main system matrix, and $\mathbf{B} \in \mathbb{R}^{m \times n}$ is the system input matrix. Proper punctuation should be used after the equation, as if it were an ordinary object in the sentence.

Do not put any empty lines before the equation. If the sentence that the equation is a part of continues after the equation (as is the case here), do not put empty lines after the equation either. That would create a new paragraph mid-sentence. For an example of how not to do it, the equation

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5.5}$$

describes the logistic function often used in machine learning. Observe how a new paragraph is created for the equation and then for this block of text (compare with the proper typesetting above). Not only does this not look correct, it may also cause incorrect page breaking.

5.5 Using footnotes

Do not be afraid to use footnotes for additional information, such as http links[1]. We use footnote links whenever we want to *point* to a website, rather then to cite it as a source. Like with everything, do not overdo it.

5.6 Referencing document elements

LaTeX allows you to dynamically reference to parts of the documents, such as

- figures: Fig. 5.4, Figure 5.4,
- equations: eq. (5.4), (5.4),
- code: Lst. 5.1,
- and any other object that can contain a `\label`.

Check the section in the `document_setup.tex` that contains useful macros for unifying the references:

```
\newcommand{\reffig}[1]{Fig.~\ref{#1}}
\newcommand{\reflst}[1]{Lst.~\ref{#1}}
\newcommand{\refalg}[1]{Alg.~\ref{#1}}
\newcommand{\refsec}[1]{Sec.~\ref{#1}}
\newcommand{\reftab}[1]{Table~\ref{#1}}
\newcommand{\refeq}[1]{\eqref{#1}}
```

Listing 5.1: LaTeX macros for referencing to document elements.

5.7 Abbreviations with Acronym

Abbreviations are handled by the *acronym* package. Example sentence with abbreviations: "**UAV!** (**UAV!**) is a flying vehicle that commonly uses Light Detection and Ranging (LiDAR) and Global Positioning System (GPS) receiver". Note that the acronyms are only explained once in the document by default. It is good practice to re-explain acronyms used both in the abstract and the rest of the document as the abstract is often presented separately. This can be achieved by resetting the internal status of the acronyms ("forgetting" that they were explained) using the `\acresetall` command after the abstract. Please, read the documentation[2].

5.8 Units of measurements with Siunitx

Typesetting of units has never been more accessible with the Siunitx package. Acceleration is measured in $\mathrm{m\,s^{-2}}$. Gravity accelerates objects at a rate $\approx 9.81\,\mathrm{m\,s^{-2}}$ near the sea level. You can define your units if you want.

5.9 Hyphens and dashes

Hyphens and dashes are the various form of the symbol "-" used in many situations. There are also various ways how to typeset the symbol in LaTeX.

- The *hyphen* is used to compound words, e.g., "the eye-opener". The hyphen is typeset as a single *minus/hyphen* character: `-`.

[1] This repository: https://github.com/ctu-mrs/thesis_template.
[2] Acronym package: http://mirrors.ctan.org/macros/latex/contrib/acronym/acronym.pdf

- The *en-dash* is used to specify ranges of values, e.g., "between 2–10". The en-dash is typeset as two consecutive hyphens characters: `--`.
- The *em-dash* is used to separate complex sentences in place of commas, parenthesis and colons — each with its particular rules. The em-dash is typeset as three consecutive hyphens characters: `---`.

Check the https://www.thepunctuationguide.com/ for all the details.

5.10 Double quotation marks

"Double quotes" in English are composed of a pair of opening (") and closing (") symbols. The opening symbol is typeset as two backtick characters: ``` `` ``` (typically below the `Esc` key on the English keyboard), and the closing quotes as two apostrophes: `''`. The LaTeX engine will convert them automatically to the opening and closing symbols. A more robust solution is to use the `csquotes` package and the `\enquote` command which also takes care of nested quoting and other peculiarities.

5.11 2D Diagrams with Tikz

Tikz is a powerful tool for drawing 2D (and 3D) shapes and diagrams. Check the documentation and examples: https://www.overleaf.com/learn/latex/TikZ_package. The benefit of using *Tikz*, instead of some other third-party drawing program, are:

- fonts are the same as in LaTeX,
- you can typeset math in LaTeX,
- you can use references to other parts of your document,
- you can version the image in git,
- the images are easily adjustable while editing your document.

Check Fig. 5.1 for example.

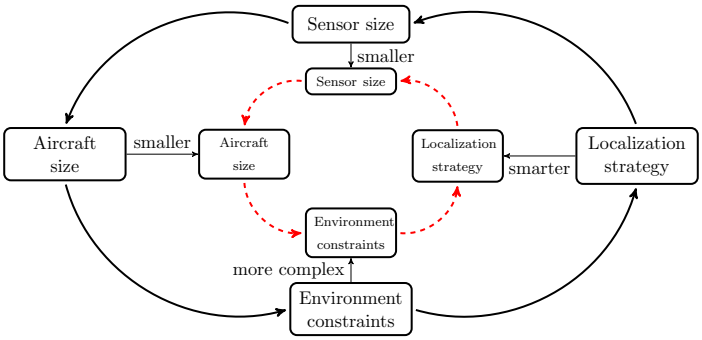


Figure 5.1: Example of a 2D diagram using tikz *PGFPlots*.

5.12 Data plots with PGFPlots

PGFPlots produces nice 2D and 3D data plots from data stored in CSV. The plot parameters can be versioned and easily adjusted by editing the plot definition file.

- Documentation and manual: https://ctan.org/pkg/pgfplots
- Compile the plots individually and then include the pdfs because it can take longer.
- Example located in `fig/plots/example_plot`, see Fig. 5.2.

- You could include the latex file directly. However, it will take longer to compile, and platforms such as Overleaf can have a problem with that.

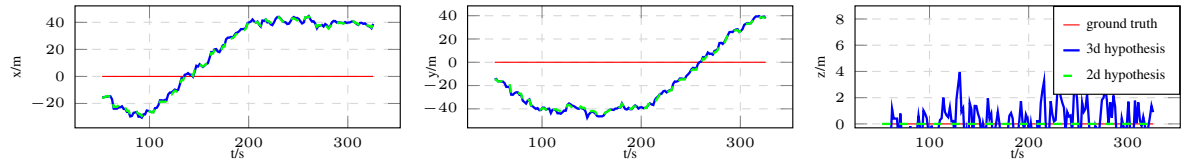


Figure 5.2: Example of a 2D plot using *PGFPlots*.

5.13 3D Plots with Sketch

Sketch is a tool for defining a 3D scene using simple descriptive language. The 3D scene is then converted to *Tikz*, which is later compiled to pdf. The benefits of using *Sketch* are similar to using *Tikz*: LaTeX fonts, versioning using git, and cleanness of the result. See the example image in Fig. 5.3.

- Documentation and manual: `http://www.frontiernet.net/~eugene.ressler/`
- Cross-compilation from *Sketch* to *pdf* using the `fig/sketch/compile_sketch.sh` script.

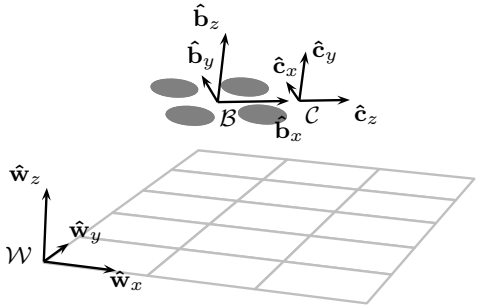


Figure 5.3: Depiction of the used coordinate systems. The image was drawn using *Sketch*.

5.14 Image collages with Subfig

We recommend using the `subfig` package, which provides the `\subfloat` command. It is more versatile than the simpler `subcaption` package. Check Fig. 5.4 for an example.

5.15 Citations with Biblatex

Biblatex is probably the most powerful citation package for LaTeX. It consumes the standard `.bib` file. However, it can sort and filter the citations using the `keywords` tag. Citing references is done using the `cite` command, e.g., [1]. You can also define some nice citation boxes, such as this one:

[1] T. Baca, M. Petrlik, M. Vrba, V. Spurny, R. Penicka, D. Hert, *et al.*, "The MRS UAV System: Pushing the Frontiers of Reproducible Research, Real-world Deployment, and Education with Autonomous Unmanned Aerial Vehicles," *Journal of Intelligent & Robotic Systems*, vol. 102, no. 26, pp. 1–28, 1 May 2021

(a) A UAV, the T650 model.

(b) Another UAV, again, the T650 model.

Figure 5.4: The caption should mention both subfigures, the Fig. 5.4a and the Fig. 5.4b. You can just refer to them as (a) and (b) in the main Figure's caption, but beware, you need to keep it correct as you edit.

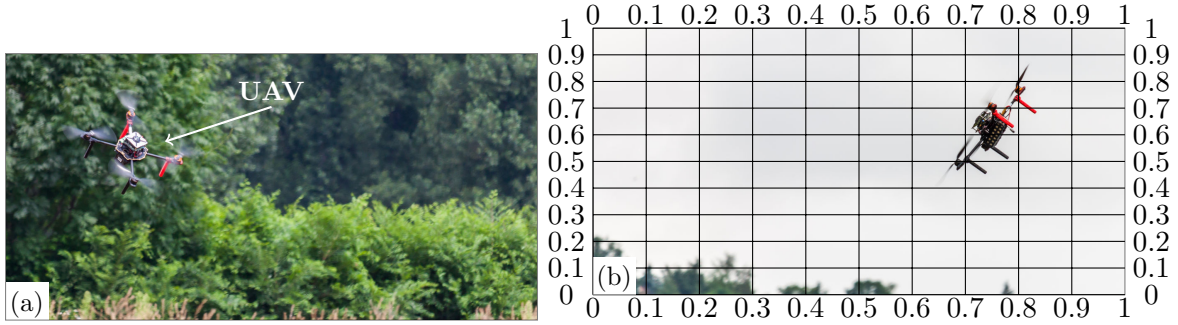


Figure 5.5: Example of using Tikz for image overlays. (a) shows a final product, (b) shows a grid useful for nailing down the coordinates.

5.16 Image overlays with Tikz

Tikz is very useful to create custom image overlays. The overlay can be set such that the image is spanned by Cartesian coordinates $(x, y) \in [0, 1]^2$ Example can be seen in Fig. 5.5.

5.17 General tips

In general, strive to make the paper easy to read and understand, and hard to misunderstand or misinterpret. Here are some more specific tips on how to achieve that (and other general suggestions).

- **Be consistent.** This applies in all contexts. For example, if you decide to use the name "LiDAR", do not mix it with "LIDAR" or "Lidar", do not mix different mathematical notations, ensure your Figures have the same style and use the same graphics for the same concepts, etc.
- After you finish writing or modifying any of:
  - a sentence,
  - a paragraph,
  - a section/chapter,
  - the whole paper/thesis,

  re-read it to make sure that it makes sense, it is coherent and correct, and doesn't

contain typos.

- If you're using a LLM-based tool (ChatGPT etc.) for grammar-proofing or even formulation of sentences, **do not just copy-paste its response** to your query. The previous rule applies doubly here. LLMs tend to often produce confident-sounding nonsense, sentences with reformulated duplicated content, or with a slightly changed meaning. They are a good tool to get inspiration to start writing about a subject, for grammar-checking, or for finding alternative, nice-sounding formulations, but they can lie or warp facts — take care when using them!

6 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.

7 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.

8 References

[1] T. Baca, M. Petrlik, M. Vrba, *et al.*, "The MRS UAV System: Pushing the Frontiers of Reproducible Research, Real-world Deployment, and Education with Autonomous Unmanned Aerial Vehicles," *Journal of Intelligent & Robotic Systems*, vol. 102, no. 26, pp. 1–28, 1 May 2021.

[2] T. Baca, G. Loianno, and M. Saska, "Embedded Model Predictive Control of Unmanned Micro Aerial Vehicles," in *IEEE International Conference on Methods and Models in Automation and Robotics (MMAR)*, IEEE, 2016, pp. 992–997.

[3] A. Benallegue, A. Mokhtari, and L. Fridman, "High-order sliding-mode observer for a quadrotor UAV," *International Journal of Robust and Nonlinear Control: IFAC-Affiliated Journal*, vol. 18, no. 4-5, pp. 427–440, 2008.

A Appendix A