DefRAG: Web Platform for the Extraction of Definitions from Scientific Articles with Generative Artificial Intelligence

Sara Dueñas-Romero, L. Alfonso Ureña-López and Eugenio Martínez-Cámara

SINAI Research Group. Center for Advanced Studies in ICT (CEATIC). Universidad de Jaén

Abstract

Scientific publications are growing at an unprecedented pace, making it increasingly difficult for researchers to stay updated. Definitions play a key role in understanding core scientific concepts, but identifying them across literature poses a challenge due to their varied and unspecific patterns. To address this, we present DefRAG, a web platform for automatic definition extraction (DE) from scientific papers, built on a Retrieval-Augmented Generation (RAG) architecture. The system retrieves relevant passages and generates coherent definitions for target concepts. We evaluate several large language models (LLMs) as the generative component of DefRAG, including Llama-3-8B-Instruct, Llama-3.2-3B-Instruct, Mistral-7B-Instruct-v0.3, and Salamandra-7B-Instruct. Among these, Llama-3 consistently shows the most reliable performance. Our results highlight not only the effectiveness of this model, but also the suitability of the RAG architecture for the task of scientific DE. DefRAG offers a practical tool to support researchers in navigating complex and growing scientific literature.

Keywords

natural language processing, retrieval augmented generation, definition extraction, large language models.

1. Introduction

In the realm of linguistic and computational research, textual definitions serve as a critical resource for understanding the meaning of specialized terms. Traditionally, definitions have been meticulously compiled in dictionaries and domain-specific glossaries, providing a structured approach to semantic comprehension. However, the manual construction and maintenance of such resources present significant challenges, particularly in rapidly evolving fields with emerging terminology and novel domains [1].

The inherent limitations of manual definition compilation have catalysed the development of computational approaches to definition extraction (DE). DE is a sophisticated natural language processing (NLP) task that aims to automatically identify and extract definitional sentences from text [2]. This ability to discern general text from definitions, which could be considered identifying "vital data" from, in our case, scientific documents directly relates to the idea of keeping scientific knowledge up to date. Furthermore, the potential to discern "future research trajectories" suggests that understanding the definitions of current concepts can provide valuable context for anticipating future developments in a field.

Over time, several methods have been developed, from simple rules [3] to sentence classification methods [4], but all face the challenge of handling variability and context of definitions. Compared to these traditional machine learning approaches, deep learning models have demonstrated superior performance in various information extraction tasks, including relation extraction which DE systems often relied on by identifying '<concept> is a <definiendum>' patterns via semantic role labeling [5].

While these models often achieve state-of-the-art results, they typically require large annotated training corpora to reach their full potential. To overcome this, the use of large language models (LLMs)

SEPLN 2025: 41st International Conference of the Spanish Society for Natural Language Processing, Zaragoza, Spain, 23-26 September 2025.

sduenas@ujaen.es (S. Dueñas-Romero); laurena@ujaen.es (L. A. Ureña-López); emcamara@ujaen.es (E. Martínez-Cámara)
0009-0006-3434-4815 (S. Dueñas-Romero); 0000-0001-7540-4059 (L. A. Ureña-López); 0000-0002-5279-8355 (E. Martínez-Cámara)

has further revolutionized the field of definition extraction. These models, pre-trained on massive amounts of text data, possess remarkable capabilities in text understanding and generation [6].

In this paper, we present DefRAG[1], a web platform for the NLP task of definition extraction of technological concepts from scientific papers. DefRAG relies in LLMs for the generation of a consolidated definition of a given concept from several articles. LLMs tend to hallucinate [7], which is unacceptable in the definition extraction task, since it may drive to crucial errors to users. Accordingly, DefRAG is based on a Retrieval-Augmented Generation (RAG) architecture to minimise the likelihood of hallucination of the definition generation module.

The rest of this paper is structured as follows. Section 2 summarises the salient works related to the definition extraction task. We present the architecture of DefRAG in Section 3. Section 4 presents the evaluation conducted, and section 5 highlights the main conclusions of this work.

2. Related work

Research in definition extraction (DE) has evolved from rule-based approaches and sentence classification algorithms to the application of advanced language models. According to Navigli and Velardi [1], early methods focused on specific patterns and rules to identify definitions in texts. However, these methods were limited by the rigidity of the rules and the inability to handle language variability. Over the years, challenges like DeftEval [8] have encourage researches to develop intricate solutions to this problem.

With the advancement of language models, their ability to understand and generate text more consistently and accurately has been explored. For example, BERT [9] and GPT-3 [6] have shown remarkable ability to understand and generate natural language, being used in diverse tasks within natural language processing (NLP).

2.1. Retrieval-Augmented Generation

As proposed in the mentioned corresponding article [10] and in its current context, RAG refers to a method in which a model, in order to generate text or answer questions, first retrieves relevant information from a corpus of documents, and then uses this retrieved information to improve the quality of its responses.

This approach allows developers to improve the accuracy of large models without the need for exhaustive retraining for specific tasks. Especially suitable for knowledge-intensive tasks, RAG combines a retriever with a *sequence2sequence* generator model, subjected to *end-to-end fine-tuning* for better capture of the same. By dynamically retrieving external information during the inference process, RAG substantially improves response accuracy, addressing problems such as hallucinations [11].

3. DefRAG System

Drawing from our comprehensive analysis of existing definition extraction techniques, we propose an innovative approach leveraging LLMs within a RAG architecture. This methodological advancement addresses critical limitations inherent in traditional definition extraction methods, which often struggle with contextual nuance, domain-specific terminology, and the dynamic nature of scientific discourse. By integrating sophisticated retrieval mechanisms with the advanced linguistic comprehension capabilities of state-of-the-art language models, our system aims to transcend conventional text classification methods.

The primary objective of our proposed system is to dynamically generate precise, contextually rich definitions for specialized technical concepts by intelligently extracting and synthesizing information from a curated knowledge base of scientific literature. This approach not only enhances the accuracy of definition extraction but also provides a flexible, scalable solution for capturing the evolving semantic landscapes across various academic and research domains.

[1] DefRAG web page: https://cetedex.ujaen.es/defrag/

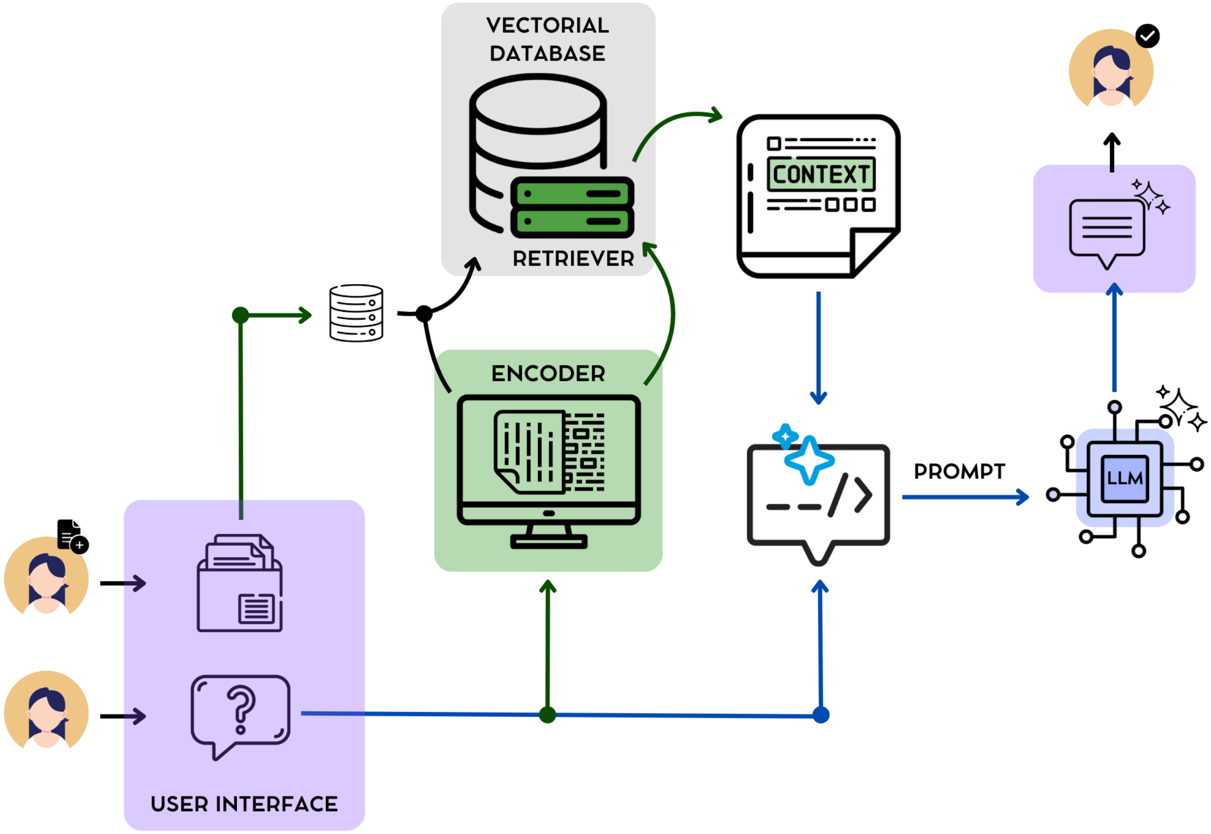


Figure 1: Complete DefRAG system: user interface, architecture and workflow.

In this section, we first describe the internal architecture of DefRAG (see Section 3.1). We then detail the design and implementation of the user interface (see Section 3.2), which serves as a gateway for the public to directly interact with and explore the capabilities of this innovative system.

3.1. DefRAG architecture

As we have already established, this computer system is specialized in the extraction of definitions in natural language from a knowledge base composed of scientific documents.

DefRAG, shown in Figure 1, is designed to improve access to and understanding of scientific information by automating the process of extracting and presenting definitions efficiently. The system features a client-server architecture and consists of the following modules:

- `Information Retrieval (IR) Module`: in the figure, the IR module shown in green, works through relevance queries to the documents that must be indexed in a way that their access is efficient. For this problem, it is necessary to know the source document of the extracts where the retrieval finds the relevant information. All these restrictions are resolved with the implementation of the index through word embedding or vectorization of the texts.
- `Language Generation Module`: It takes the retrieved passages within the prompt as input and produces an output text based on them. It can be any sequence-to-sequence model capable of generating coherent and fluent text, in this case an LLM that uses a decoder-only architecture. This module can be trained on different tasks, such as question answering, summarization, or text generation.
- `User Interaction Module`: means through which users interact with the system. Using a REST API as a base, a set of access points is established that allow communication between the client, where the interface is executed, and the server, where the process of obtaining the

definition is executed. This architecture allows communication between heterogeneous systems independently, adapting to different requirements, improving scalability, maintenance and changes in the internal system.

3.1.1. Extending the knowledge base

The system features robust capabilities for expanding its knowledge base through the integration of new documents, both user-provided and retrieved from external repositories. When new documents are received via HTTP requests, the system processes them using the same methodology applied to the initial knowledge base. Once indexed, the PDF documents are automatically moved to the global knowledge base directory specified in the configuration file, ensuring all knowledge resources are centralized and properly organized.

Additionally, the system can autonomously enhance its knowledge base by retrieving relevant scientific papers from ArXiv.org. The designated function enables targeted document acquisition by accepting theme and result quantity parameters. This functionality creates an ArXiv API client, performs searches based on relevance, and systematically downloads the resulting PDFs to a designated directory.

Each document is properly named according to its theme and source URL. This automated document retrieval system significantly expands the knowledge base's breadth and depth without requiring manual intervention, keeping the system's information corpus continuously updated with the latest scientific research.

3.1.2. Information Retrieval (IR) Module

The first step of the definition extraction process is retrieval, where the documents or text fragments that are likely to contain the information needed to generate the requested definition are obtained. These fragments are stored in a vector database to make the retrieval process more efficient.

We use as scientific papers database the one provided by one of the projects of the Cátedra Isdefe/CETEDEX-UJA. We use as scientific papers database which consists of 1011 papers in PDF format. We propose to transform all the documents into vectorial representations based on word embeddings, thus optimizing information storage and retrieval. This approach, essential in the RAG model for accurate and fast queries, consists of mapping natural language words or phrases to vectors in a continuous space of reduced dimensions, facilitating the comparison and manipulation of large volumes of textual data. In particular, we use a sentence transformer model [12] trained on a novel pre-training method that inherits the advantages of BERT and XLNet on the Microsoft model "mpnet-base", fine-tuned in a 1B sentence pairs dataset[2] [13].

This implementation leverages Chroma, as the vector database, which is particularly well-suited for handling scientific document collections. Our approach uses a two-phase retrieval strategy:

1. `Single index creation`: when initializing the system, we either create a new index or load an existing one. During index creation, the PDF documents are loaded, then split into manageable chunks. This splitting strategy maintains semantic coherence while optimizing for retrieval performance, with configurable chunk size, set to 2000 character per chunk, and overlap, set to 100, parameters to balance context preservation against computational efficiency. The chunked documents are then embedded and stored in the Chroma vector database, which persists the index to disk for future use. If an index already exists in the specified directory, it's simply loaded without reprocessing the documents.
2. `Information Retrieval`: For the actual retrieval operation, the module takes a concept query and passes it to the retriever component. The retriever searches the vector database for the most relevant document chunks based on embedding similarity. The retrieved fragments are then formatted into *summaries* that serve as context for the subsequent generation phase.

[2] https://huggingface.co/sentence-transformers/all-mpnet-base-v2

This implementation optimizes both storage and retrieval efficiency while keeping semantic relevance, which is crucial when dealing with large volumes of scientific literature where precise information extraction is essential for generating accurate definitions and explanations.

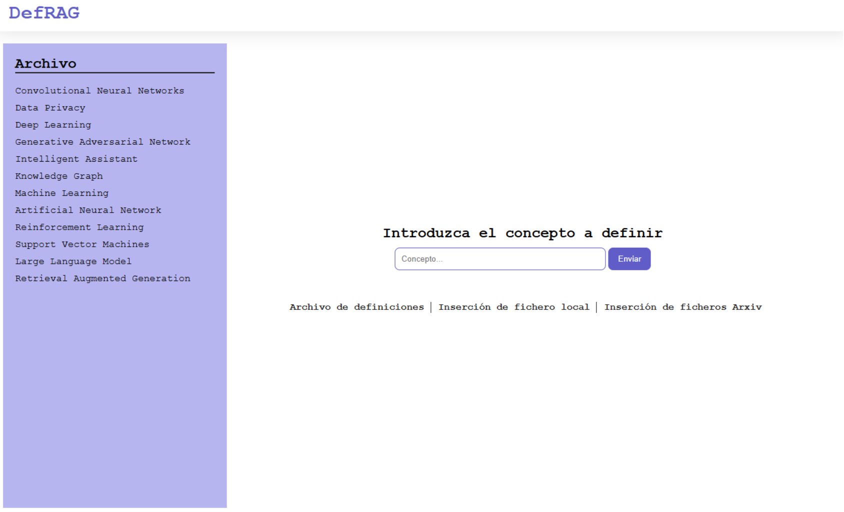

DefRAG
Archivo
Convolutional Neural Networks
Data Privacy
Deep Learning
Generative Adversarial Network
Intelligent Assistant
Knowledge Graph
Machine Learning
Artificial Neural Network
Reinforcement Learning
Support Vector Machines
Large Language Model
Retrieval Augmented Generation
Introduzca el concepto a definir
Concepto...
Enviar
Archivo de definiciones | Inserción de fichero local | Inserción de ficheros Arxiv

Figure 2: DefRAG system: user interface main menu.

3.1.3. Language Generation Module

Next, once the relevant context fragments are retrieved, we proceed to generating the desired answer via **prompting**, which involves the construction of the prompt that is presented to the untrained LLM. This prompt is composed of the information retrieved in the previous step, together with a personalized message indicating to the model the specific task to be performed. An example of such a message would be: "*Based on this information, you must generate a definition for the concept <concept>*".

This prompt is then sent to the inference server, where the designated LLM is waiting to be prompted. The server waits for a response and this output is then reinserted in the main workflow.

The output of the system is the generated definition, which is presented to the user in an accessible way and is characterized by being multi-document, i.e., elaborated from the consolidated information of several retrieved documents [14]. In addition, responsible behaviour is ensured by providing the sources of the documents used to generate the definition, allowing the user to verify the information.

Finally, the server sends to the client the elaborated definition together with the documents used as sources, presenting everything in a clear and appropriate manner. Once this task is completed, the definition, together with its sources, is stored for future queries, thus facilitating access to previously generated and verified information.

3.2. User Interface

The system is presented through an intuitive web interface, Figure 2, designed to offer a simple and accessible user experience. From the main page, users can access all the functionalities of the system. The main goal of the system is to ensure that the user can request the definition of a concept and the platform displayes the corresponding definition in english along with the references used to generate it. Moreover, DefRAG provides two additional functions related to the ability to expand the knowledge base of papers.

3.2.1. Adding individual papers.

The user can insert a local document by accessing corresponding menu option, which displays a form that facilitates the upload of a PDF file from the user's device allowing its renaming before saving.

3.2.2. Bulk adding of papers.

The user can add documents from ArXiv on a specific topic. In this scenario, the user selects the desired option, fills the search criteria, and the system downloads the requested documents. Subsequently, the system processes the documents, incorporates them into the index, and notifies the user with a success message.

4. Evaluation and results

An extensive knowledge base of scientific articles (1011 papers provided by the sponsoring entity, see Section 3.1.2) extracted from ArXiv, a pre-publication service for scientific articles widely used by the scientific community, especially in the areas of engineering and science, has been used.

4.1. Large Language Models used

The RAG architecture is based on the use of an LLM, so we have evaluated the performance of 4 LLMs, namely:

- `Llama-3-8B-Instruct`: an auto-regressive language model that uses an optimized transformer architecture, in particular, this is an instruction tuned model optimized for dialogue use cases via supervised fine-tuning (SFT) and reinforcement learning with human feedback (RLHF) to align with human preferences for helpfulness and safety [15].
- `Llama-3.2-3B-Instruct`: similar to the previous one, this model is also an instruction tuned model optimized for dialogue use case like agentic retrieval and summarization tasks.
- `Salamandra-7B-Instruct`: 7B instructed version of the transformer-based decoder-only language model that has been pre-trained from scratch on highly multilingual data that comprises text in 35 European languages and code by Language Technologies Unit in Barcelona Supercomputing Center (BSC) [16].
- `Mistral-7B-Instruct`: developed by the French company Mistral AI, this model is a fine-tuned version of the open source base model Mistral-7B [17].

4.2. Evaluation metrics

In order to evaluate the performance and accuracy of the models automatically, evaluation metrics have been used to compare the generated response with one or more reference definitions. Additionally, the accuracy and performance of the retrieval module has been evaluated.

These evaluation metrics have been collected using RAGAS [18], a library that provides tools to supercharge the evaluation of LLMs. The metrics values are obtained by a method called LLM-as-judge, defined as the use of LLMs to evaluate objects, actions or decisions based on predefined rules, criteria or preferences [19]. Formally defined as

$$E \leftarrow P_{LLM}(x \oplus C) \tag{1}$$

where a language model (PLLM) assigns a probability to the combination ($\oplus$) of an input (x) - which can be text, image or video - with a context (C), which is usually a prompt or dialogue history information. The final result of this process is the evaluation E.

RAGAS utilizes LLM-as-judge with GPT-3.5-turbo-16k [3] to compute numeric metrics for NLP performance. For this evaluation, we have employed the following metrics:

- `Faithfulness`: Factual consistency of the generated answer against the given context. An answer is ‘faithful’ if all its claims can be inferred by the given context.

[3] We clarify that this is the default judge LLM used by the RAGAS framework.

Table 1
Unified results for all Models.

Model	Concept	Faith	CP	Factual	Semantic
Llama-3-8B-Instruct	CNN	**1.0000**	**1.0000**	**0.5500**	**0.9600**
	Intelligent Assistant	**1.0000**	**0.9700**	**0.7500**	**0.9300**
	University of Jaén	**1.0000**	**1.0000**	**1.0000**	**0.9300**
	Machine Learning	0.6700	**1.0000**	**0.9200**	0.9300
Llama-3.2-3B-Instruct	CNN	**1.0000**	0.9500	0.6400	0.9500
	Intelligent Assistant	**1.0000**	**0.9700**	0.3600	0.9400
	University of Jaén	**1.0000**	**1.0000**	0.0000	0.6900
	Machine Learning	0.7500	**1.0000**	0.7200	0.8800
Salamandra-7B-Instruct	CNN	**1.0000**	0.9500	0.3800	0.9500
	Intelligent Assistant	0.3300	**0.9700**	0.6000	0.8500
	University of Jaén	0.0000	**1.0000**	0.0000	0.7000
	Machine Learning	0.9400	**1.0000**	0.6700	**0.9500**
Mistral-7B-Instruct	CNN	**1.0000**	0.9500	0.4500	0.9500
	Intelligent Assistant	**1.0000**	**0.9700**	0.3300	0.8900
	University of Jaén	0.0000	**1.0000**	0.0000	0.7000
	Machine Learning	**0.9500**	**1.0000**	**0.7500**	**0.9500**

Table 2
Scientific concepts extracted for the evaluation. The out of domain concepts are those that are not in the scientific papers of the knowledge base.

In/Out of Scope	**Concept**
In	Absolute Error, Convolutional Neural Networks (CNN), Data Mining, Data Privacy, Deep learning, Generative Adversarial Network, Intelligent Assistant, Internet Of Things, Knowledge Graph, Large Language Models, Machine Learning, Natural Language Processing, Artificial Neural Network, Reinforcement Learning, Sentiment Analysis, Support Vector Machines, Word Embeddings
Out	Wooden Door, Flower Vase, University of Jaén

- `Context Precision`: Evaluates if all the ground-truth (correct answer) items present in the contexts are favoured. All of the relevant items should appear at the top of the ranking.
- `Factual Correctness`: Overlap between answer and ground-truth, this is done by extracting the statements of each text and comparing their alignment.
- `Semantic similarity`: Resemblance between the answer and the ground-truth, this can offer valuable insights into the quality of the generated response.

While RAGAS's LLM-as-judge approach facilitates large-scale, reproducible evaluation, it may inherit bias from the judge model (GPT-3.5-turbo) and can be sensitive to prompt formulation. Future work should include complementary human evaluation or diverse judge models to mitigate these biases.

4.3. Results

To compute these results, expected definitions for the test concepts have been manually extracted from the knowledge base as well as outside of the scope of the technological domain in order to compare them to the generated answer.

The test concepts are a selection of 20 technical terms that may or may not appear in the said knowledge base, as well as some other concepts outside of the computing domain. This selection has been optimised to ensure the evaluation of correct definitions as well as outside-of-the-scope performance. Table 2 shows the concepts used in the evaluation.

Based on the results shown in Table 1, it is clear that Llama-3-8b-Instruct substantially outperforms the other models. The decision between using Llama-3 and Llama-3.2 was grounded in their behaviour

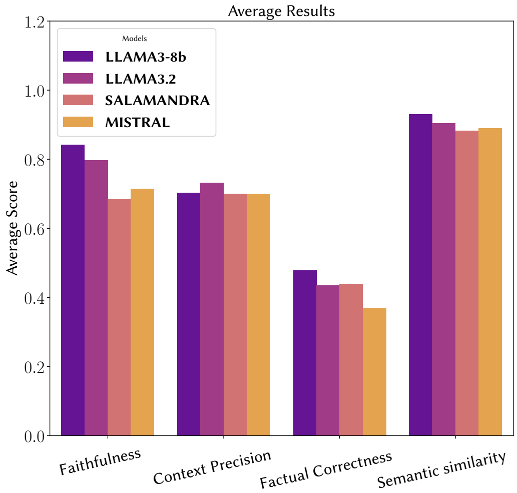


Figure 3: Comparison of average performance metrics across the models.

when faced with the out-of-the-scope concepts, as in the case of ‘University of Jaén’. In these cases, Llama-3 always returns the predefined response informing that the concept is not in the knowledge base. In contrast, models like Mistral or Salamandra generate an incorrect definition, or in other words, it hallucinates.

Llama-3.2 often generates an answer that indicates the same sentiment of the default out-of-the-scope answer, but in this system we also take into account the strict following of the rules indicated on the prompt: ’*If the concept does not appear in the context provided or you cannot extract a definition in it, return the default phrase <Retrieving a definition for the concept is not possible with this context> and nothing else*’.

Therefore, after analysing these results, we conclude that a RAG methodology, specially based on inference on Llama-3 instruct-tuned models, can be used for the extraction of definitions of scientific scope, thus validating the hypothesis of this work.

5. Conclusion

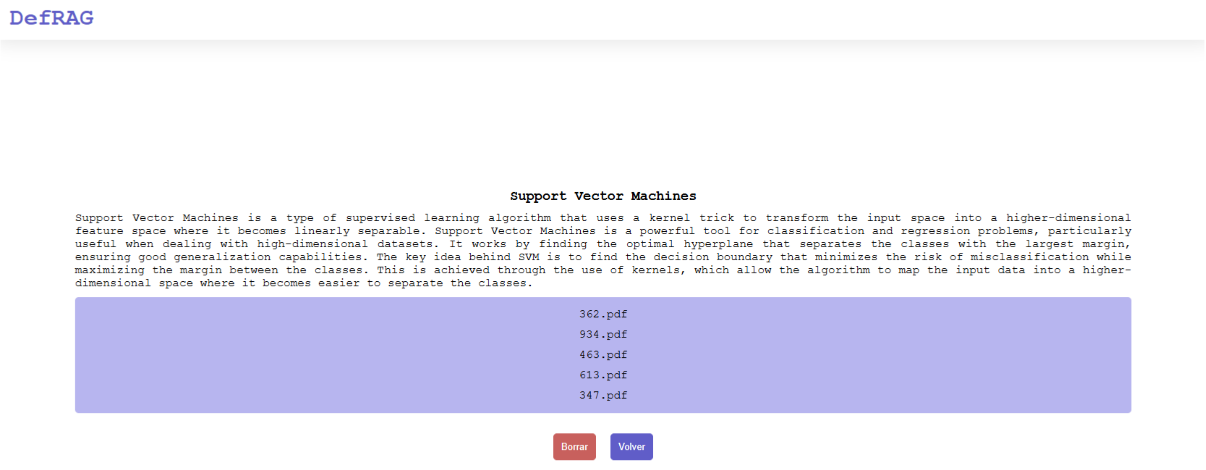
This work has demonstrated the effectiveness of using LLMs, specifically the ones from the Llama 3 family, for definition extraction using a retrieval-augmented generation (RAG) architecture. Our approach addresses and overcomes the limitations of traditional methods and sentence classification algorithms by properly considering context.

Experimentation shows that Llama-3-8b-Instruct offers higher performance and avoids the hallucinations common in other models, standing out as the most suitable choice for this system. The completed results shown in Figure 3, in which the average result for each metric across 20 reference examples is averaged per model, illustrate both the efficiency of LLMs in this task and how useful LLM-as-judge is in the evaluation of RAG architectures.

This approach not only allows the knowledge base to be extended seamlessly, but also provides a system capable of generating coherent, natural-language, multi-document definitions for the user, such us the one shown in Figure 4, marking a significant advance in NLP and opening new avenues for future research.

Several promising research directions emerge from this study. First, we propose extending the current RAG architecture to incorporate multi-lingual definition extraction, which would significantly broaden the system’s applicability across different linguistic contexts.

Another critical avenue for investigation involves developing more sophisticated evaluation metrics that can more comprehensively assess the nuanced quality of generated definitions beyond current



Support Vector Machines

Support Vector Machines is a type of supervised learning algorithm that uses a kernel trick to transform the input space into a higher-dimensional feature space where it becomes linearly separable. Support Vector Machines is a powerful tool for classification and regression problems, particularly useful when dealing with high-dimensional datasets. It works by finding the optimal hyperplane that separates the classes with the largest margin, ensuring good generalization capabilities. The key idea behind SVM is to find the decision boundary that minimizes the risk of misclassification while maximizing the margin between the classes. This is achieved through the use of kernels, which allow the algorithm to map the input data into a higher-dimensional space where it becomes easier to separate the classes.

362.pdf
934.pdf
463.pdf
613.pdf
347.pdf

Figure 4: DefRAG system: user interface definition page.

quantitative measures. Furthermore, integrating active learning techniques could potentially improve the system's ability to identify and learn from ambiguous or edge-case definitional contexts. Lastly, investigating the potential of combining this RAG approach with emerging multimodal language models could open up innovative ways of extracting and representing definitions.

Acknowledgments

This research has been partially funded by the Isdefe/CETEDEX-UJA chair and the projects FedDAP (PID2020-116118GA-%I00), MODERATES (TED2021-130145B-I00), SocialTOX (PDC2022-133146-C21) y CONSENSO (PID2021-122263OB-C21) financed by MCIN/AEI/10.13039/501100011033, *FEDER Una manera de hacer Europa y Unión Europea NextGenerationEU/PRTR*. This work was also funded by the *Ministerio para la Transformación Digital y de la Función Pública and Plan de Recuperación, Transformación y Resiliencia* - Funded by EU – NextGenerationEU within the framework of the project *Desarrollo Modelos ALIA*.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT and Perplexity to assist with LaTeX syntax as well as grammar and spelling checks in English. After using these tools, the authors carefully reviewed and edited the content as needed and take full responsibility for the publication's content.

References

[1] R. Navigli, P. Velardi, Learning word-class lattices for definition and hypernym extraction, in: Annual Meeting of the ACL, 2010. URL: https://aclanthology.org/P10-1134/.

[2] J. L. Klavans, S. Muresan, Evaluation of the DEFINDER system for fully automatic glossary construction, in: Proceedings of the AMIA Symposium, 2001, p. 324.

[3] R. Del Gaudio, A. Branco, Automatic extraction of definitions in portuguese: A rule-based approach, in: J. Neves, M. F. Santos, J. M. Machado (Eds.), Progress in Artificial Intelligence, Springer Berlin Heidelberg, Berlin, Heidelberg, 2007, pp. 659–670.

[4] A. Khoo, Y. Marom, D. Albrecht, Experiments with sentence classification, in: Proceedings of the Australasian Language Technology Workshop 2006, 2006, pp. 18–25.

[5] H. Wang, K. Qin, R. Y. Zakari, G. Lu, J. Yin, Deep neural network based relation extraction: An overview, 2021. arXiv:2101.01907.

[6] T. Brown, B. Mann, N. Ryder, M. Subbiah, Kaplan..., D. Amodei, Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901.

[7] Z. Ji, N. Lee, R. Frieske, T. Yu, D. Su, Y. Xu, E. Ishii, Y. J. Bang, A. Madotto, P. Fung, Survey of hallucination in natural language generation, ACM computing surveys 55 (2023) 1–38.

[8] S. Spala, N. A. Miller, F. Dernoncourt, C. Dockhorn, Semeval-2020 task 6: Definition extraction from free text with the DEFT corpus, 2020. arXiv:2008.13694.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, 2019. arXiv:1810.04805.

[10] P. Lewis, E. Perez, A. Piktus, F. Petroni, V. Karpukhin, N. Goyal, H. Küttler, M. Lewis, W. tau Yih, T. Rocktäschel, S. Riedel, D. Kiela, Retrieval-augmented generation for knowledge-intensive NLP tasks, 2021. arXiv:2005.11401.

[11] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. arXiv:2312.10997.

[12] N. Reimers, I. Gurevych, Sentence-BERT: Sentence embeddings using siamese BERT-Networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, 2019.

[13] K. Song, X. Tan, T. Qin, J. Lu, T.-Y. Liu, MPNet: Masked and permuted pre-training for language understanding, 2020. arXiv:2004.09297.

[14] A. Artikis, T. Eiter, A. Margara, S. Vansummeren, Dagstuhl seminar on the foundations of composite event recognition, ACM SIGMOD Record 49 (2021) 24–27. doi:10.1145/3456859.3456865.

[15] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, A. Yang, A. Fan, A. Goyal, A. Hartshorn, A. Y. et al., The llama 3 herd of models, 2024. arXiv:2407.21783.

[16] A. Gonzalez-Agirre, M. Pàmies, J. Llop, I. Baucells, S. D. Dalt, D. Tamayo, J. J. Saiz, F. Espuña, J. Prats, J. Aula-Blasco, M. Mina, A. Rubio, A. Shvets, A. Sallés, I. Lacunza, I. Pikabea, J. Palomar, J. Falcão, L. Tormo, L. Vasquez-Reina, M. Marimon, V. Ruíz-Fernández, M. Villegas, Salamandra technical report, 2025. arXiv:2502.08489.

[17] A. Q. Jiang, A. Sablayrolles, A. Mensch, C. Bamford, D. S. Chaplot, D. de las Casas, F. Bressand, G. Lengyel, G. Lample, L. Saulnier, L. R. Lavaud, M.-A. Lachaux, P. Stock, T. L. Scao, T. Lavril, T. Wang, T. Lacroix, W. E. Sayed, Mistral 7b, 2023. arXiv:2310.06825.

[18] S. Es, J. James, L. Espinosa-Anke, S. Schockaert, Ragas: Automated evaluation of retrieval augmented generation, 2023. arXiv:2309.15217.

[19] J. Gu, X. Jiang, Z. Shi, H. Tan, X. Zhai, C. Xu, W. Li, Y. Shen, S. Ma, H. Liu, S. Wang, K. Zhang, Y. Wang, W. Gao, L. Ni, J. Guo, A survey on LLM-as-a-judge, 2025. arXiv:2411.15594.