João **Coelho**

MSc · Computer Science and Engineering · IST
PhD Candidate · Electrical Engineering/Language Technologies · IST/CMU
Pittsburgh, PA, USA
412-728-0692 | joaomavc@gmail.com | jmvcoelho.github.io | JMVCoelho | jmvcoelho

Summary

João Coelho holds a MSc degree in Computer Science and Engineering from Instituto Superior Técnico in Lisbon, Portugal. With a background in research projects, João has published his work in international conferences and journals, and specializes in Machine Learning, Natural Language Processing, and Information Retrieval. João is proficient in Python, including machine learning libraries such as PyTorch and TensorFlow, and has extensive experience in training and deploying Transformer models (such as BERT, GPT, and T5) for various tasks. In addition, he has programming experience in C#, Java, and C. Currently a PhD candidate in a Dual-PhD program between Instituto Superior Técnico and Carnegie Mellon University, conducting research work that focuses on improving web retrieval models using document structure, and addressing the context length for long document retrieval and representation learning.

Work Experience

Caixa Mágica Software
Lisbon, Portugal

Machine Learning Engineer / Research — *Sep. 2020 - Sep. 2022*

- Project AppRecommender
  - **Description**: This project aimed to study novel approaches for mobile application search and recommendation.
  - **Tasks**: Created a novel dataset containing metadata on thousands of mobile applications, which involved performing thorough data cleaning and analysis. Proposed a self-supervised training technique to develop a semantic search engine based on neural language models, which also proved highly effective in improving the accuracy of a *More Like This* recommendation system for mobile applications. Additionally created and deployed APIs to facilitate the integration of both the search engine and recommendation system. To ensure the efficacy of these systems, conducted two rounds of user-centered tests, which yielded highly positive results.
  - **Technologies**: Python, PyTorch, Flask, MongoDB, Docker, ElasticSearch.
- Project Digitarq
  - **Description**: This project aims to enhance the current platform of DGLAB's Portuguese Archive, which is renowned as one of the oldest archives globally, housing millions of documents. Key factors involve migrating the existing relational databases to graph databases to optimize their performance and improve search efficiency.
  - **Tasks**: Successfully developed and presented the proof-of-concept to DGLAB, which focused on graph databases, domain-specific ontologies, semantic web, and neural models for entity extraction. This proposal was accepted by the client, leading to one of the company's largest projects. Additionally established the groundwork for the data team by training named entity recognition models, training semantic similarity models for entity disambiguation, implementing and optimizing text clustering techniques, and constructing the initial graph database. Given the unlabeled and under-resourced nature of the Portuguese textual data, techniques such as transfer learning, few-shot learning, and zero-shot learning were employed to achieve optimal results.
  - **Technologies**: Python, PyTorch, Huggingface Transformers, Neo4j, ElasticSearch.
- Other Responsibilities
  - Supervised four interns in two projects, (i) recommendation with collaborative filtering, (ii) slack chatbot.
  - Conducted a workshop at FISTA'21, regarding ElasticSearch.

Caixa Mágica Software
Lisbon, Portugal

Intern Developer — *Jul. 2020 - Sep. 2020*

- Project Helpdesk
  - **Description**: This project aimed to build an helpdesk platform, where questions supported by a FAQ section are automatically answered.
  - **Tasks**: Tested multiple approaches for FAQ retrieval under low resources. Integrated the best one within the previously built platform. Deployed the final product.
  - **Technologies**: Python, scikit-learn, Tensorflow, ReactJS, MySQL, Docker, Gitlab runners for CI/CD.

Xpand IT
Lisbon, Portugal

Intern Developer — *Jul. 2019 - Sep. 2019*

- Project XBot
  - **Description**: This project aimed to build an internal proactive chatbot, which would send notifications or ask questions to the team.
  - **Tasks**: Implemented the backend of the bot, namely the proactive capabilities and possible message scenarios (notification, yes/no question, polls). Integrated a language understanding tool (LUIS). Deployed the final product in a Microsoft Teams channel.
  - **Technologies**: C#, .NET, Entity Framework, MySQL, Azure, LUIS.

Teaching

IST (Instituto Superio Técnico)
Lisbon, Portugal

Teaching Assistant — *Sep. 2021 - Aug. 2023*

- Search and Planning (MEIC) - 2021/2022, 2022/2023
- Artificial Intelligence (LEIC) - 2021/2022, 2022/2023

CMU (Carnegie Mellon University)
Pittsburgh, PA, USA

Teaching Assistant — *Aug. 2024 - Present*

- Large Language Models: Methods and Applications - Fall 2024

Formal Education

CMU (Carnegie Mellon University) — *Pittsburgh, PA, USA*
PhD in Language Technologies — *Sep. 2022 - Present*
- Dual Degree PhD. Conducting research on Neural Information Retrieval (emphasis on long, structured documents).

IST (Instituto Superior Técnico) — *Lisbon, Portugal*
PhD in Electrical and Computer Engineering — *Sep. 2022 - Present*
- Dual Degree PhD. Conducting research on Neural Information Retrieval (emphasis on long, structured documents).

IST (Instituto Superior Técnico) — *Lisbon, Portugal*
MSc in Computer Science and Engineering (Intelligent Systems/Language Technologies) [18.18/20] — *Sep. 2019 - Nov. 2021*
- Merit Distinctions: 2019/2020 and 2020/2021.
- Thesis: *Geographical Question Answering Leveraging Neural Language Models for Passage Retrieval* [19/20]

IST (Instituto Superior Técnico) — *Lisbon, Portugal*
BSc in Information Systems and Computer Engineering [15.46/20] — *Sep. 2016 - Jun. 2019*
- Merit Distinctions: 2017/2018 and 2018/2019.

ESPS (Escola Secundária de Ponte de Sor) — *Ponte de Sor, Portugal*
High School [18/20] — *Sep. 2013 - Jun. 2016*
- Science and Technology Area.

Other Courses

European Summer School in Information Retrieval — *Lisbon, Portugal*
Attendee — *Jul. 2022*
- Week-long summer school focusing on Information Retrieval.

Deep Learning For Language Analysis Summer School — *Koln, Germany (Virtual)*
Attendee — *Aug. 2021*
- Week-long summer school focusing on Machine Learning for NLP.
- NVIDIA certificate: *Building Transformer-Based Natural Language Processing Applications.*

Lisbon Machine Learning Summer School — *Lisbon, Portugal (Virtual)*
Attendee — *Jul. 2021*
- Week-long summer school focusing on Machine Learning for NLP.

Publications

International Journals

1. João Coelho, Beatriz Paula, Diogo Mano, Carlos Coutinho, João Oliveira, Ricardo Ribeiro, and Fernando Batista, **Semantic Similarity for Mobile Application Recommendation under Scarce User Data**. Engineering Applications of Artificial Intelligence, 2023.

International Conferences

1. João Coelho, Bruno Martins, João Magalhães, Jamie Callan, and Chenyan Xiong, **Dwell in the Beginning: How Language Models Embed Long Documents for Dense Retrieval**. Proceedings of the Annual Meeting of the Association for Computational Linguistics, 2024

2. João Coelho, João Magalhães, and Bruno Martins, **Improving Neural Models for the Retrieval of Relevant Passages to Geographical Queries**. Proceedings of the ACM SIGSPATIAL International Conference on Advances in Geographic Information Systems, 2021

3. João Coelho, António Neto, Miguel Tavares, Carlos Coutinho, João Oliveira, Ricardo Ribeiro and Fernando Batista, **Transformer-based Language Models for Semantic Search and Mobile Applications Retrieval**. Proceedings of the International Joint Conference on Knowledge Discovery and Information Retrieval, 2021.

4. João Coelho, António Neto, Miguel Tavares, Carlos Coutinho, Ricardo Ribeiro and Fernando Batista , **Semantic Search of Mobile Applications Using Word Embeddings**. Proceedings of the Symposium on Languages, Applications and Technologies, 2021.

Languages

English
C1 Level
- Certified by Oxford School
- Certified by Duolingo (140)

Portuguese
Native