

Maël Houbre

First year PhD student in Natural Language Processing

Education

- 2022–2025 **PhD Student**, *LS2N, University of Nantes*, .
PhD student in NLP and information retrieval.
- 2016–2019 **M.S in Computer science** , *Telecom Nancy*, .
French computer science engineering school. Embedded software speciality.
- 2014–2016 **Classes préparatoires aux grandes écoles**, *Brest Navy High School*, .
Two-years undergraduate intensive course in mathematics and physics to prepare for French engineer schools' competition exams.

PhD studies

- Title *Unsupervised absent keyphrase generation for scientific indexing*
- Supervisors Béatrice Daille, Florian Boudin
- Description Online publishing platforms made scientific publications much easier to share, resulting in a rapid growth of the number of articles available to the community. However this increasing number of publications makes it difficult for researchers to comprehensively find relevant articles for their research. When using keyphrases as a document expansion for indexing, research have showed that keyphrases, especially keyphrases that are not in the source text, that is, absent keyphrases, have a beneficial impact on document retrieval. However, state of the art generative approaches need large datasets to be trained which makes it difficult to have efficient models for several domains and languages. The work of this thesis is therefore focused on how to generate absent keyphrases in an unsupervised manner.

Publications

- o **Maël Houbre**, Florian Boudin, Béatrice Daille: A Large-Scale Dataset for Biomedical Keyphrase Generation. *Proceedings of the 13th International Workshop on Health Text Mining and Information Analysis (LOUHI 2022)*

Work experience

- Sept 2022 – **Teaching assistant**, *Nantes Université*.
Dec 2022 Introduction to computer sciences (32h)
- Janv 2023 – **Teaching assistant**, *Nantes Université*.
Apr 2023 Databases 101 (32h)

Nov 2019– **NLP Engineer**, *French Ministry of Defence*.

Mar 2022 As an NLP engineer, I was essentially in charge of the evaluation of the various algorithms and technical reports that we received during projects.

Missions:

- Develop baseline algorithms
- Prepare training and testing datasets, evaluation scripts for each task
- Evaluate models' performances and write reports
- Responsible for the team's technological watch

Oct 2018– **Graduation Project with Sopra Steria**, *Automatic paraphrase generation for data augmentation*.

Mar 2019

In my senior year in Telecom Nancy, I took part in a project with the company Sopra Steria. We were a team of three students who were in charge of developing a deep learning based paraphrase generation model for Sopra Steria's data science department.

Missions:

- Learn about word/document embedding librairies (FastText, gensim) and compare them using different similarity measures
- Develop a LSTM based generation model

Computer science skills

- Programming languages: Python, C/C++, R
- Machine Learning librairies: PyTorch, Hugging face transformers, Scikit-learn
- Dataset librairies: Pandas, Hugging face datasets
- Scientific skills: Signal processing, Linear algebra, Probability and statistics, Functional analysis