Antoine Simoulin, PhD

antoine.simoulin@gmail.com

For the last five years, I worked as a senior data scientist for Quantmetry, a consulting agency in Paris, France. Parallel to that, I continued to strengthen my technical background as Quantmetry sponsored my PhD at the Laboratoire de Linguistique Formelle from Paris Cité University. By combining these two profiles, I am able to translate complex technical problems into long-term objectives, realistic roadmaps, and rational solutions. I conducted multiple data sciences projects for large companies and deployed effective solutions in production. My role went beyond implementation since I contributed from ideation, framing, deployment, to monitoring.

Education

University of Paris Cité — Paris, France
PhD, Computer Science — Feb. 2019 – July 2022
My PhD entitled, *Sentence embeddings and their relation with sentence structures*, focuses on Natural language Processing methods for building sentence embeddings. Advised by Prof. Benoit Crabbé, member of LLF lab.

École Polytechnique — Paris, France
Dual master program (MSc), Data Science — 2016 – 2017
The leading French research, academics, and innovation institution.

ENSTA Paris — Paris, France
Master of science (MSc), Simulation and Mathematical Engineering — 2013 – 2017
French engineering school accessible through selective *classe préparatoire*. Last year advised by Prof. Pierre Carpentier, director of UMA lab.

Work Experiences

Quantmetry — Paris, France
Senior data scientist, NLP — Apr . 2017 – July 2022
Together with a data-science team of business and technical consultants, I participated in multiple projects to automate or optimize processes. I led end-to-end projects for real-world problems, such as putting neural models in production for classifying, summarizing, and automating email replies in one of the largest French insurance companies.

Crédit Agricole Corporate and Investment Banking — New York, USA
Quantitative analyst intern — Sept. 2015 – Aug. 2016
I implemented and improved Monte-Carlo's algorithms using CUDA on graphic card for the capital calculation of an internal insurance.

Skills

Languages	**French**: Native; **English**: Fluent (TOEIC 965/990); **German**: Working knowledge
Programming Skills	In-depth knowledge of **Python** , working knowledge **C, C++**, and familiar with **R, Matlab**
	In-depth knowledge of **SQL**, and working knowledge of **Hadoop, Spark**
	In-depth knowledge of **Linux/Unix/Shell environments**
Office & Web	In-depth knowledge of **Microsoft Office**. Basic understanding of web design: **HTML, CSS**

Research Interests

My PhD, entitled *Sentence embeddings and their relation with sentence structures*, studies how neural networks compose text units to build sentence embeddings. I apply linguistic insights to neural network architectures. I design and implement dynamic architectures following tree or graph syntactic patterns. I aim to quantify the impact of linguistic bias on neural network architectures and how compositionality might be leveraged through the network structure. Along with linguistics, my work involves implementing complex structured neural networks as well as pre-training large language models at scale such as a version of GPT-2 for French with over a billion parameters.

Publications

Unifying Parsing and Tree-Structured Models for Generating Sentence Semantic Representations 2022
NAACL 2022, Student Research Workshop
Antoine Simoulin, Benoit Crabbé

How Many Layers and Why? An Analysis of the Model Depth in Transformers 2021
ACL 2021, Student Research Workshop
Antoine Simoulin, Benoit Crabbé

Contrasting Distinct Structured Views to Learn Sentence Embeddings 2021
EACL 2021, Student Research Workshop
Antoine Simoulin, Benoit Crabbé

Generative Pre-trained Transformer in ______ (French) 2021
TALN 2021: Traitement Automatique des Langues Naturelles
Antoine Simoulin, Benoit Crabbé

Unifying Parsing and Tree-Structured Models for Generating Sentence Semantic Representations 2021
In submission
Antoine Simoulin, Benoit Crabbé

Deep Learning : des usages contrastés dans le monde socio-économique 2021
Statistique et Société, 8: 55-108
R. Adon, F. Arthur, G. Baquiast, G. Hochard, A. Kaid Gherbi, A. Nègre, **A. Simoulin** et al.

An innovative solution for breast cancer textual big data analysis 2020
In submission
N. Thiebaut, **A. Simoulin**, K. Neuberger, I. Ibnouhsein, N. Bousquet, N. Reix, S. Molière, C. Mathelin

Impact du dépistage : une expérience française 2017
Mise à jour du Collège National des Gynécologues et Obstétriciens Français
C. Mathelin, J. Colin, S. Molière, A. Fleury, C. Linck, M. Paté, C. Guldenfels, **A. Simoulin**, et al.

Teaching

Natural language processing (2020 – 2022)
Graduate level course in natural language processing (NLP) at Paris Cité University. The course includes 7 sessions (course and lab) and introduces statistical models (TF-IDF, Bag-of-Words, LDA, Embeddings, language models) for NLP. Around 25 students from the mathematics department followed the course each year.

Talks and Presentations

Pre-trained neural networks for text generation and their implications Apr. 2021
Machine Learning Meetup, Epitech engineering school, Nantes France
Around 30 students and professionals in the field of data science attended the talk. I presented my paper about the first large pre-trained generative model in French.

Implementing and deploying natural language processing projects Dec. 2019
AI Paris, France
Around 800 professionals in the field of data science attended the presentation. We presented the project of emails classification at MAIF and the challenges to deploy a project in production.

Melusine open-source release Dec. 2019
BigData Paris, France
Open source release of Melusine, a library for emails processing. Around 80 professionals in the field of data science attended the presentation.

Senometry project: analysis of textual medical records for structured data extraction May 2018
NLP Meetup, Paris, France
Presentation to around 40 professionals in the field of data science. The research project consists in using NLP methods to automatically analyze data from medical records.

Open Source Contributions

GPT-fr is a French large pre-trained language model for French. The base version, equivalent to OpenAI GPT in English, includes above 1B parameters.

PyTree implements tree-structured neural networks in PyTorch. The package provides highly generic implementations as well as efficient batching methods. The project was listed among the winners of the PyTorch Annual Hackathon 2021.

Sentence embedding pre-trained model trained on 1B sentence pairs. The project was listed among the winners during the Hugging Face Community week using JAX/Flax for NLP & CV 2021.

Melusine is a high-level Python library for email classification and feature extraction developed by Quantmetry and MAIF.