Antoine Simoulin, PhD

antoine.simoulin@gmail.com

I am a recently graduated PhD with both academic and industrial experiences. Quantmetry sponsored my PhD research at the *Laboratoire de Linguistique Formelle* from *Paris Cité University*. My research focuses on analyzing the role of structure in building sentence embeddings. In that regard, I implemented innovative neural network architectures and published several articles in top peer-reviewed NLP conferences. Parallel to that, I held the position of senior data scientist at *Quantmetry*, a leading AI consulting firm in France. Over the last five years, I developed 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.

Education

University of Paris Cité — Paris, France
PhD, computer science — 2019 –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 Sciences — 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
I implemented solutions for real-world problems in many large French industries. I also facilitated a connection between the research and application teams by presenting new updates in NLP research and identifying potential use cases and technologies.

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.

Columbia Consortium for Risk Management — New York, USA
Research intern — May 2015 – July 2015
I applied probability measures for the prediction of extremely rare events in a behavior model for decision under risk and uncertainty.

Research Interests

My PhD, entitled *Sentence embeddings and their relation with sentence structures*, studies how neural networks compose text units to build sentence embeddings. In my research, 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: North American Chapter of the Association for Computational Linguistics, Student Research Workshop
Antoine Simoulin, Benoit Crabbé

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

Contrasting Distinct Structured Views to Learn Sentence Embeddings 2021
EACL 2021: European Chapter of the Association for Computational Linguistics, 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**, F. Talaouit-Mockli, N. Bousquet

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**, K. Neuberger, J. Jégu

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.

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.

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.

Sentence embedding pre-trained model trained on 1B sentence pairs during the Community week using JAX/Flax for NLP & CV, organized by Hugging Face.

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

Awards and Services

- Reviewer ACL'20, EMNLP'20
- Hackathon Winner. PyTree, a PyTorch package for recursive neural networks. **PyTorch Annual Hackathon 2021**
- Hackathon Winner. Train the Best Sentence Embedding Model Ever with 1B Training Pairs. **Hugging Face Community week using JAX/Flax for NLP & CV 2021**