Roman Castagné

Location	Paris, France	**Phone**	+33 6 31 52 80 26
Nationality	French	**LinkedIn**	linkedin.com/in/roman-castagné
Website	romancast.github.io	**Email**	roman.castagne@gmail.com

Work History

Oct 2021 - **ALMAnaCH, Inria**, Paris, Fr
PhD student, directed by Benoît Sagot and Eric de la Clergerie

- Worked on MANTa, a module combining the **robustness** of tokenizer-free models with the **speed** of subword tokenizers.
- Built **BLOOM tokenizer** as part of the BigScience Tokenizer Working Group.
- Investigated the use of artificial data to unlock **pretraining for low resource languages**.
- Explored **model-based Curriculum Learning** as a way to accelerate Language Model pretraining.

Publications :

- MANTa: Efficient Gradient-Based Tokenization for Robust End-to-End Language Modeling, *EMNLP 2022 (Findings)*
- BLOOM: A 176B-Parameter Open-Access Multilingual Language Model

April 2021 - Oct 2021 **ALMAnaCH, Inria**, Paris, Fr
Research Intern, under the supervision of Benoît Sagot

- Working on **character-level language modelling** to improve multilingual models and the transfer of performance for out-of-domain data.

Feb 2020 - Jul 2020 **Naver Labs Europe**, Grenoble, Fr
NLP (Natural Language Processing) Research Intern

- Worked on a more informative modelisation of **Aspect Based Sentiment Analysis** using SemEval datasets.
- Used HuggingFace Transformers and PyTorch to design a **multitask model** with balanced losses, that led to a **patent**.
- Completed extensive evaluation of the system in a **zero-shot multilingual setting** using XLM-RoBERTa.

Jul 2019 - Jan 2020 **Reacfin**, Louvain-la-Neuve, Be
Data Science Research Intern

- Improved accuracy of an **insurance text classifier** by benchmarking models (embeddings, RNN, CNN...).
- Implemented a **neural Named Entity Recognition system** robust to errors in the data (typos, OCR errors...).

Education

2020 - 2021 **ENS Paris-Saclay**, Paris, Fr
MVA (Mathematics, Vision and Learning) master's degree

Convex Optimisation (A. D'Aspremont), Optimal Transport (G. Peyré), Reinforcement Learning (A. Lazaric, M. Pirotta), Deep Learning (V. Lepetit), Computer Vision (I. Laptev, J. Ponce), NLP (B. Sagot, E. Dupoux), Kernel Methods (J. Mairal, JP. Vert), Bayesian ML (R. Bardenet), Graphs in ML (D. Calandriello)

Deep Learning project: Studied performance drops in disaster tweet classification under different domain shifts, and used Domain Adaptation to design robust classifiers.

RL project: Designed an algorithm for Reinforcement Learning from imperfect demonstrations.

Computer Vision project: Adapted the SinGAN architecture for image inpainting.

2017 - 2021 **Ecole des Ponts ParisTech**, Paris, Fr
Mathematics Engineering and Computer Science

2015 - 2017 **Lycée Thiers**, Marseille, Fr
Preparatory Classes for Engineering Schools

Teaching, Volunteering and Presentations

Oct 2022 **RJMI**, Inria Paris
Animated a workshop to solve an olympiad math problem as part of the RJMI, a two-day event for high school students eager to pursue a career in STEM.

Jun 2022 **Deep Voice**, SCAI (Sorbonne Université AI lab)
Prepared and gave a 3 hours-long tutorial to learn about and train Language Models, now available as a blogpost.

Jun 2022 **Franco-German Workshop on AI**, Inria Headquarters
Talked about the challenges posed by tokenizers for multilingual Language Models, slides available here.

Mar 2022 **Machine Learning for NLP**, ENSAE ParisTech
Taught the course labs to last year students from ENSAE's Data Science master's degree.

2018-2019 **Sports Association**, ECOLE DES PONTS
Logistics Manager for infrastructures and equipments.

Skills

- **Programming Languages**
  Python (advanced), Latex (advanced),
  C++ (confirmed), R (confirmed), Julia (basics)
- **Languages**
  - French (mother tongue)
  - English (Professional, TOEIC: 980/990)
- **Libraries and Frameworks**
  - *Deep Learning and ML:* PyTorch, Keras, Scikit-learn
  - *Natural Language Processing:* HuggingFace Transformers and Tokenizers, Megatron-LM, SentencePiece, NLTK, Gensim, Spacy
  - *Cluster workload manager:* Slurm, OAR

Interests

Climbing (lead and bouldering), **Trail Running**, **Music** (bass played in a Contemporary Music conservatory).