Gabriele Sarti

PhD Student in Interpretability and Natural Language Processing

gabriele.sarti996@gmail.com | gsarti.com | gsarti | gabrielesarti | @gsarti_ | Gabriele Sarti

Education

University of Groningen, Center for Language and Cognition *Groningen, Netherlands*
Doctorate (PhD) in Natural Language Processing *September 2021 - Present*

- Development of interpretability approaches for the study of generative language models.
- Supervised by Arianna Bisazza, Malvina Nissim and Grzegorz Chrupała with funding from the NWO InDeep consortium.
- Main instructor of the 2024 BSc course “Fundamentals of Machine Learning: Theory and Practice”. Co-instructor of multiple editions of the Advanced Natural Language Processing MSc course and co-supervisor of several BSc and MSc thesis projects.

University of Trieste & International School for Advanced Studies (SISSA) *Trieste, Italy*
Master’s of Science (MSc) in Data Science and Scientific Computing *October 2018 - December 2020*

- Data Science for Social Sciences track, all courses taught and assessed in English. Graduation score: 110 Cum Laude.
- Thesis: Interpreting Neural Language Models for Linguistic Complexity Assessment. Supervisors: Felice Dell’Orletta, Davide Crepaldi.

Cégep of Saint-Hyacinthe *Saint-Hyacinthe, Canada*
Collegial Studies Degree (DEC) in Management Informatics *August 2015 - June 2018*

- All courses taught and assessed in French. Final R Score: 34.13.

Work Experience

Amazon Web Services *New York, United States*
Applied Scientist Intern, AWS AI Labs *June 2022 - September 2022*

- Research project with Amazon Translate on prompting large multilingual language models for attribute-controlled machine translation.

Aindo *Trieste, Italy*
Research Scientist *November 2020 - August 2021*

- Generative AI for question answering applied to structured prediction tasks in the clinical domain.
- Experimental work in cross-lingual transfer to Italian for generative language models pre-trained in English.
- Creation of temporally-aware neural recommender system based on graph neural networks for online shopping recommendations.
- Controllable sketch-to-image synthesis for the fashion industry with multimodal pre-trained neural networks (StyleGAN & CLIP).

Institute of Computational Linguistics “Antonio Zampolli” (ILC-CNR) *Pisa, Italy*
Research Assistant in Natural Language Processing *October 2019 - December 2019*

- Developed a multitask training pipeline for reading times predictions using neural language models. Models were evaluated on linguistic complexity assessment and analyzed using probing and representational similarity methods.

Skytech Communications Inc. *Montréal, Canada*
Machine Learning Engineer Intern *February 2018 - June 2018*

- Augmented a convolutional neural networks system for automatic record keeping with handwritten characters recognition capabilities.
- Development of a sentiment analysis platform for highlighting inappropriate student comments in professors’ evaluation forms.

Talks and Publications

Selected Publications

Model Internals-based Answer Attribution for Trustworthy Retrieval-Augmented Generation
Jirui Qi*, **Gabriele Sarti***, Raquel Fernández, Arianna Bisazza
Conference on Empirical Methods for Natural Language Processing (EMNLP 2024)

A Primer on the Inner Workings of Transformer-based Language Models
Javier Ferrando, **Gabriele Sarti**, Arianna Bisazza, Marta R. Costa-jussà
Computational Linguistics (Under review)

🔗 **Quantifying the Plausibility of Context Reliance in Neural Machine Translation**
Gabriele Sarti, Grzegorz Chrupała, Malvina Nissim, Arianna Bisazza
International Conference on Learning Representations (ICLR 2024)

🔗 **IT5: Text-to-text Pretraining for Italian Language Understanding and Generation**
Gabriele Sarti, Malvina Nissim
Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)

🔗 **Inseq: An Interpretability Toolkit for Sequence Generation Models**
Gabriele Sarti, Nils Feldhus, Ludwig Sickert, Oskar van der Wal, Malvina Nissim, Arianna Bisazza
Annual Meeting of the Association for Computational Linguistics: System Demonstrations (ACL Demo 2023)

🔗 **RAMP: Retrieval and Attribute-Marking Enhanced Prompting for Attribute-Controlled Translation**
Gabriele Sarti, Phu Mon Htut, Xing Niu, Benjamin Hsu, Anna Currey, Georgiana Dinu, Maria Nadejde
Annual Meeting of the Association for Computational Linguistics (ACL 2023)

🔗 **DivEMT: Neural Machine Translation Post-Editing Effort Across Typologically Diverse Languages**
Gabriele Sarti, Arianna Bisazza, Ana Guerberof-Arenas, Antonio Toral
Conference on Empirical Methods for Natural Language Processing (EMNLP 2022)

🔗 **That Looks Hard: Characterizing Linguistic Complexity in Humans and Language Models**
Gabriele Sarti, Dominique Brunato, Felice Dell'Orletta
***Hon. mention, Best student paper**, Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2021)*

→ 🔗 *Full list of publications*

Selected Talks

Explaining Neural Language Models from Internal Representations to Model Predictions
Gabriele Sarti, Alessio Miaschi — Laboratory at AILC Lectures on Computational Linguistics 2023

Empowering Human Translators with Interpretable Interactive Neural Machine Translation
Gabriele Sarti — Oral presentation, A glimpse of the future track, XAI4Debugging Workshop at NeurIPS 2021

Neural Language Models: The New Frontier of Natural Language Understanding
Gabriele Sarti — Young Chapter of Italian Statistical Society (ySIS), StaTalk 2019

The Educational Impact of Artificial Intelligence
Gabriele Sarti, Alexandre Brunet — 38th Symposium of Québec Association for Collegial Pedagogy (AQPC), 2018

→ 🔗 *Full list of talks*

Extracurricular Activity

Research Projects

2021-Present	Core developer, 🔗 Inseq interpretability toolkit
March 2020	Creator of the 🔗 COVID-19 Semantic Browser, in collaboration with Area Science Park and AILC
2019	System Demonstration Organizer, 🔗 Trieste Next Science Festival, AI-talo Svevo

Social Engagements

2022-Present	Research and Upskilling Advisor, 🔗 AI Safety Initiative Groningen (AISIG)
2021-2022	Team Mentor, 🔗 PiCampus School of AI W21 Program
2020	President and Founder, 🔗 Artificial Intelligence Student Society (AI2S)
2020-2021	Communication Team Member, 🔗 Italian Association for Computational Linguistics (AILC)
2015-2020	Volounteer, AFS Italy and AFS Canada

Honors & Awards

2023	20K€ Grant for Human-AI Interaction Project, Imminent Research Center	*Rome (IT)*
2023	2K€ eScience Center Fellowship, Amsterdam eScience Center	*Amsterdam (NL)*
2018	Excellence Scholarship for Data Science MSc Students, SISSA	*Trieste (IT)*
2018	Valedictorian of the Management Informatics Class of 2018, Cégep of Saint-Hyacinthe	*Montréal (CA)*
2016	Perseverance Scolarship for Informatics Students, Hydro-Québec	*Montréal (CA)*
2013	Full merit scolarship towards an AFS exchange year in Canada, Telecom Italia Mobile	*Rome (IT)*