Luca Soldaini

pronouns: he/him or they/them
luca@soldaini.net soldaini.net

Employment

Amazon

Alexa AI, Web Info — Manhattan Beach, CA, USA
Senior Applied Scientist — *July 2021 – Present*
Applied Scientist — *August 2019 – June 2021*

- Working on ranking and generative question answering models as part of a system that uses web knowledge to answer questions from Amazon Alexa users.
- Supervised 7 research interns and published 6 peer-reviewed publications.

Alexa AI, Natural Language Understanding — Cambridge, MA, USA
Applied Scientist — *June 2018 – August 2019*

- Worked on methods for hypotheses ranking in the Alexa Natural Language Understanding (NLU) pipeline. I lead or collaborated on several initiatives that aimed at (*i*) fusing interpretations from diverse sources or (*ii*) using heterogeneous signals to improve ranking performance.
- Core member of team developing a neural modeling framework/inference engine for NLU applications built on Apache MxNet.

Applied Research Intern — *June 2017 – August 2017*

- Studied the problem of obtaining multilingual, domain-specific word embeddings that can be used to accelerate model training in low-resource languages.
- Explored methods for bootstrapping NER to new languages when no domain-specific training data is available.

Microsoft Research – Advanced Technology Labs Israel — Herzliya, Israel
Research Intern — *September 2015 – December 2015*

- Studied the problem of identifying small cohorts of search engine users who might be affected by the same disease (a publication based on this work has been accepted at WWW 2017).

MedStar Institute for Innovation (MI2) — Washington, DC, USA
Intern — *May 2015 – August 2015*

- Developed a pipeline to extract human factors concepts from patient safety events generated by care providers.
- Helped creating a system to evaluate the quality of reports produced by radiology residents (a publication based on this work has been accepted at the DMMH workshop at SDM 2016).

Education

Georgetown University — Washington, DC, USA
Doctor of Philosophy (Ph.D.) in Computer Science — *August 2013 – April 2018*

- **Dissertation**: *"The Knowledge and Language Gap in Medical Information Seeking."*
- **Adviser**: Dr. Nazli Goharian.
- **Committee**: Dr. Der-Chen Chang, Dr. Ophir Frieder, Dr. Elad Yom-Tov, Dr. Wenchao Zhou.

Georgetown University — Washington, DC, USA
Master of Science (M.S.) in Computer Science; GPA: 4/4 — *August 2013 – May 2015*

Università degli Studi di Firenze — Florence, Italy
Bachelor of Engineering (B.Eng.) in Computer Engineering; GPA: 27.7/30 — *September 2009 – April 2013*

- **Thesis**: *"Particle Swarm Algorithm for Sphere Packing Problems."*
- **Adviser**: Prof. Fabio Schoen.

Peer-Reviewed Manuscripts

- Chao-Chun Hsu, Eric Lind, Luca Soldaini, and Alessandro Moschitti. *"Rethinking Answer Sentence Selection as a Generation Task."* Findings of Annual Conference of the Association for Computational Linguistics (ACL). 2021.
- Rujun Han, Luca Soldaini, and Alessandro Moschitti. *"Modeling Context in Answer Sentence Selection Systems on a Latency Budget."* Annual Conference of the European Chapter of the Association for Computational Linguistics (EACL). 2021.
- Mingda Li, Xinyue Liu, Weitong Ruan, Luca Soldaini, Wael Hamza, and Chengwei Su. *"Multi-task Learning of Spoken Language Understanding by Integrating N-Best Hypotheses with Hierarchical Attention."* Conference on Computational Linguistics (COLING). 2020.
- Luca Soldaini and Alessandro Moschitti. *"The Cascade Transformer: Efficient Answer Sentence Selection."* Annual Conference of the Association for Computational Linguistics (ACL). 2020.
- Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. *"Don't Parse, Generate! A Sequence to Sequence Architecture for Task-Oriented Semantic Parsing."* The Web Conference (formerly WWW). 2020.
- Sean MacAvaney, Luca Soldaini, and Nazli Goharian. *"Teaching a New Dog Old Tricks: Resurrecting Multilingual Retrieval Using Zero-shot Learning."* European Conference on Information Retrieval (ECIR). 2020.
- Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian, and Ophir Frieder. *"Overcoming Low-Utility Facets for Complex Answer Retrieval."* Information Retrieval Journal, 2018.
- Ziling Fan, Luca Soldaini, Arman Cohan, and Nazli Goharian. *"Relation Extraction for Protein-Protein Interactions Affected by Mutation."* ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (ACM-BCB). 2018.
- Arman Cohan, Bart Desmet, Andrew Yates, Luca Soldaini, Sean MacAvaney, and Nazli Goharian. *"SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions."* Conference on Computational Linguistics (COLING). 2018. **Area chair favorite paper.**
- Luca Soldaini, Timothy Walsh, Arman Cohan, Julien Han, and Nazli Goharian. *"Helping or Hurting? Predicting Changes in Users' Risk of Self-Harm Through Online Community Interactions."* CLPsych Workshop, Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2018.
- Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. *"TempMH: Temporal Annotation of Self-Reported Mental Health Diagnoses."* CLPsych Workshop, Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2018.
- Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian, and Ophir Frieder. *"Characterizing Question Facets for Complex Answer Retrieval."* ACM conference on Research and Development in Information Retrieval (SIGIR). 2018.
- Sean MacAvaney, Luca Soldaini, Arman Cohan, and Nazli Goharian. *"Tree-LSTMs for Scientific Relation Classification."* SemEval Workshop, Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2018.
- Luca Soldaini, Andrew Yates, and Nazli Goharian. *"Denoising Clinical Notes for Medical Literature Retrieval with Convolutional Neural Model."* Conference on Information and Knowledge Management (CIKM). 2017.
- Luca Soldaini, Andrew Yates, and Nazli Goharian. *"Learning to Reformulate Long Queries for Clinical Decision Support."* Journal of the Association for Information Science and Technology (JASIST), Special Issue on Biomedical Information Retrieval. 2017.
- Luca Soldaini and Elad Yom-Tov. *"Inferring Individual Attributes from Search Engine Queries and Auxiliary Information."* Wide World Web conference (WWW). 2017.
- Luca Soldaini and Nazli Goharian. *"Learning to Rank for Consumer Health Search: a Semantic Approach."* European Conference on Information Retrieval (ECIR). 2017.
- Luca Soldaini and Nazli Goharian. *"QuickUMLS: a Fast, Unsupervised Approach for Medical Concept Extraction."* MedIR workshop, ACM conference on Research and Development in Information Retrieval (SIGIR). 2016.
- Arman Cohan, Luca Soldaini, and Nazli Goharian. *"Identifying Significance of Discrepancies in Radiology Reports."* Workshop on Data Mining for Medicine and Healthcare (DMMH), SIAM International Conference on Data Mining (SDM). 2016.
- Luca Soldaini, Andrew Yates, Elad Yom-Tov, Ophir Frieder, and Nazli Goharian. *"Enhancing Web Search in the Medical Domain via Query Clarification."* Information Retrieval Journal, 2016.
- Arman Cohan, Luca Soldaini, and Nazli Goharian. *"Matching Citation Text and Cited Spans in Biomedical Literature: a Search–Oriented Approach."* Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT). 2015.
- Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. *"Retrieving Medical Literature for Clinical Decision Support."* European Conference on Information Retrieval (ECIR). 2015.
- Arman Cohan, Luca Soldaini, Andrew Yates, Nazli Goharian, and Ophir Frieder. *"On Clinical Decision Support."* ACM Conference on Bioinformatics, Computational Biology, and Health Informatics (BCB). 2014.

Peer-Reviewed Abstracts

- Pranav A, MaryLena Bleile, Arjun Subramonian, Luca Soldaini, Danica J Sutherland, Sabine Weber, and Pan Xu. *"How to Make Virtual Conferences Queer-Friendly: A Guide."* Widening NLP (WiNLP) Workshop, Empirical Methods in Natural Language Processing (EMNLP). 2021.
- Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian, and Ophir Frieder. *"Overcoming Low-Utility Facets for Complex Answer Retrieval."* Persented at the KG4IR Workshop, ACM conference on Research and Development in Information Retrieval (SIGIR). 2018.
- Luca Soldaini and Nazli Goharian. *"Learning to Rank for Consumer Health Search: a Semantic Approach."* Presented at the Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL). 2017.

Other Professional Activities

- **Public speaking**:
  - **SoCal ML/NLP Symposium** (panelist). Virtual, March 23rd, 2021.
  - **Engineers for Professional Inclusion Conference** (panelist). UCLA, April 7th, 2020.
  - **Qwer Hacks** (keynote speaker) *"We Deserve More Queer Scientists."*. UCLA, January 25th, 2020.
- **Queer in AI, core organizer** (2020–present):
  - **Social events** (organizer and host). Coordinated events at EMNLP 2020, AACL 2020, COLING 2020, EACL 2021, NAACL 2021, ACL 2021.
  - **Graduate applications relief program** (program chair). reviewed applications and provided financial support for queer scientists applying to graduate school.
- **D&I committee, social chair**: NAACL 2021.
- **Action editor**: ACL Rolling Review, 2021–2022 cycle.
- **Area chair/senior program committee member**: ACL 2020 (Information Retrieval and Text Classification), NAACL 2021 (Information Retrieval and Text Classification), ACL 2021 (Question Answering), AAAI 2022.
- **Reviewer**: Journal of the American Medical Informatics Association (JAMIA), 2017-current; Journal of Artificial Intelligence (JAIR) 2020–current.
- **Program committee member**: ACL 2018, 2019 (main track); SIGIR 2018–2020 (full and short paper tracks); CoNLL 2017 (main track); CIKM 2017–2020 (short papers track); EMNLP 2018–2020 (main track); AACL-IJCNLP 2020 (main track); The Web Conference 2017–2021 (formerly WWW); EACL 2021 (main track), IJCAI 2020, 2021.
- **Core developer** of *QuickUMLS*, a toolkit for fast unsupervised biomedical concept extraction for clinical and lay text. (230+ stars on GitHub as of September 2021). Available at `github.com/Georgetown-IR-Lab/QuickUMLS`.