Luca Soldaini

luca@soldaini.net

Education

Georgetown University — Washington, DC, USA

Doctor of Philosophy (Ph.D.) in Computer Science — Aug 2013 – May 2018 *(expected)*

- Research interests: information retrieval, machine learning, and natural language processing.
- Dissertation: "*The Knowledge and Language Gap in Medical Information Seeking.*"
- Adviser: Dr. Nazli Goharian.

Master of Science (M.S.) in Computer Science — Aug 2013 – May 2015

- GPA: 4/4

Università degli Studi di Firenze — Florence, Italy

Bachelor of Engineering (B.Eng.) in Computer Engineering — Sep 2009 – Apr 2013

- GPA: 27.7/30; final mark: 109/110
- Thesis: "*Particle Swarm Algorithm for Sphere Packing Problems*."
- Adviser: Prof. Fabio Schoen.

Employment

Amazon – Alexa Machine Learning Team — Cambridge, MA, USA

Applied research intern — Jun 2017 - Aug 2017

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

Microsoft Research – Advanced Technology Labs Israel — Herzliya, Israel

Research intern — Sep 2015 - Dec 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 - Aug 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).

Peer Reviewed Publications

1. 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.

2. 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.
3. Luca Soldaini and Elad Yom-Tov. “Inferring Individual Attributes from Search Engine Queries and Auxiliary Information.” Wide World Web conference (WWW). 2017.
4. Luca Soldaini and Nazli Goharian. “Learning to Rank for Consumer Health Search: a Semantic Approach.” European Conference on Information Retrieval (ECIR). 2017.
5. 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.
6. 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.
7. 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.
8. 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.
9. 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.
10. 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.

Technical Reports

11. Luca Soldaini, Will Edman, Nazli Goharian. “Team GU-IRLAB at CLEF eHealth 2016: Task 3.” Conference and Labs of the Evaluation Forum (CLEF). 2016. (best submission out of 10 participants)
12. Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder. “Query Reformulation for Clinical Decision Support Search.” Text REtrieval Conference (TREC). 2014.
13. Arman Cohan, Luca Soldaini, Saket S.R. Mengle, and Nazli Goharian. “Towards Citation-Based Summarization of Biomedical Literature.” Text Analysis Conference (TAC). 2014.

Teaching Experience

Georgetown University Washington, DC, USA

Teaching Assistantship (TA) August 2013 - Present

- Information Retrieval (undergraduate & graduate courses) – Fall 2013, 2014, 2016.
- Information Systems (undergraduate course) – Spring 2014.
- Data Mining (undergraduate course) – Spring 2014, 2015, 2016, 2017; Fall 2017.
- Introduction to Database (undergraduate course) – Spring 2015.

Co-instructor January 2017 - Present

- Health Search and Mining (graduate course) – Spring 2017.
- Text Mining (graduate course) – Fall 2017.

Professional Activities

- **Reviewer**, Journal of the American Medical Informatics Association (JAMIA). 2017-current.
- **Program Committee Member**, CoNLL 2017, main track.

- **Program Committee Member**, CIKM 2017, short papers.
- **Program Committee Member**, computational health track. WWW 2017, WWW 2018.
- **Subreviewer**, AAAI 2017, AAAI 2018.

Patents

- Ratwani, Raj, Allan Fong, Ross Filice, Arman Cohan, Luca Soldaini, Nazli Goharian, and Ophir Frieder. "Systems and methods for targeted radiology resident training." U.S. Patent Application 15/410,850, filed January 20, 2017.

Awards

- **SIGIR Student Travel Grant**, CIKM 2017.
- **Graduate School Conference Travel Award**. Georgetown University. 2017.
- **Student Travel Grant**. MedIR workshop. SIGIR 2016.
- **Second Place at Best Poster Award** (2 out of 40). "On Clinical Decision Support". Informatics Symposium at Georgetown University 2014.

Technical Skills

- **Programming languages**: Python (advanced), Bash (competent), Javascript, Java, C# (some exposure).
- **Frameworks**: web server (Flask), databases (MySQL, MongoDB), search engines (Elasticsearch, Terrier), scientific computation and machine learning (Keras, Tensorflow, NumPy, scikit-learn, SciPy, Vopal Wabbit, Weka), natural language processing (spaCy, Stanford CoreNLP).
- **Platforms**: UNIX (OS X, RedHat Linux, Debian), Microsoft Windows.