Arman Cohan

✉ armanc@allenai.org 🌐 http://www.ArmanCohan.com ☎ +1 (202)-509-3830

Current

- **Allen Institute for AI (AI2), Seattle WA** June 2018 - Present
  Research Scientist
  - Natural Language Processing for addressing information overload.
  - Research interests: Representation learning, language modeling, self-supervised/unsupervised learning, summarization, NLP in specialized domains (e.g., science and medicine)

Education

- **Doctor of Philosophy in Computer Science** *Aug. 2013 - May 2018*
  Georgetown University, Washington DC, USA
  - *Dissertation: Text Summarization and Categorization for Scientific and Health-related data*
    2019 Harold N. Glassman Distinguished Doctoral Dissertation Award in the Sciences
  - *Advisor: Dr. Nazli Goharian*
- **Master of Science in Computer Science** *Aug. 2013 – May. 2015*
  Georgetown University, Washington DC, USA
- **Master of Science in Information Engineering** *Oct. 2010 – Feb. 2013*
  Amirkabir University of Technology, Tehran, Iran
- **Bachelor of Science in Information Engineering** *2006 – 2010*
  Amirkabir University of Technology, Tehran, Iran

Awards

- Harold N. Glassman Distinguished Doctoral Dissertation Award in the Sciences *2019*
- COLING 2018 conference “Area Chair Favorite (outstanding) Paper” Award *2018*
- Dr. Karen Gale Exceptional PhD Student Award in Science *2018*
- EMNLP 2017 “Best Paper Award” *2017*
- ICBI (Innovation Center for Biomedical Informatics) best poster award *2017, 2018*
- ACM-BCB 2017 NSF Award *2017*
- Georgetown University’s merit-based fellowship award *2013, 2014, 2015, 2016, 2017*
- Best poster award (second place) - Innovation Center for Biomedical Informatics (ICBI) *2014*
- Ranked in the top 1% of Iranian National universities Entrance Exam *2006*
- Certificate of Distinction, University of Waterloo’s Euclid International Mathematics Contest *2005*
- Qualification to second round of 3 National Olympiads (Mathematics, Computer Science, and Physics). *2005*

Publications

Conference papers

- Fact or Fiction: Verifying Scientific Claims
  David Wadden, Kyle Lo, Lucy Lu Wang, Shanchuan Lin, Madeleine van Zuylen, Arman Cohan, Hannaneh Hajishirzi
  EMNLP 2020: Empirical Methods for Natural Language Processing (*Acceptance rate: 22.4%*)

- SLEDGE-Z: A Zero-Shot Baseline for COVID-19 Literature Search
  Sean MacAvaney, Arman Cohan, Nazli Goharian
  EMNLP 2020: Empirical Methods for Natural Language Processing (*Acceptance rate: 22.4%*)

- TLDR: Extreme Summarization of Scientific Documents Search
  Isabel Cachola, Kyle Lo, Arman Cohan, Daniel S. Weld
  EMNLP 2020 (Findings): Empirical Methods for Natural Language Processing (*Acceptance rate: 37.9%*)

- SPECTER: Document-level Representation Learning using Citation-informed Transformers
  Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, Daniel S. Weld
  ACL 2020: Association for Computational Linguistics (*Acceptance rate: 25.2%*)

- Ranking Significant Discrepancies in Clinical Reports
  Sean MacAvaney, Arman Cohan, Nazli Goharian, Ross Filice
  ECIR 2020: European Conference on Information Retrieval (*Acceptance rate: 26%*)

- Pretrained Language Models for Sequential Sentence Classification
  Arman Cohan, Iz Beltagy, Daniel King, Bhavana Dalvi, Daniel S. Weld
  EMNLP 2019: Empirical Methods for Natural Language Processing (*Acceptance rate: 20.5%*)

- SciBERT: A Pre-trained Language Model for Scientific Text
  Iz Beltagy, Kyle Lo, Arman Cohan
  EMNLP 2019: Empirical Methods for Natural Language Processing (*Acceptance rate: 20.5%*)

- Ontology-Aware Clinical Abstractive Summarization
  Sean MacAvaney, Sajad Sotudeh, Arman Cohan, Nazli Goharian, I. Talati, R. Filice
  SIGIR 2019: ACM SIGIR Conference on Research and Development in IR (*Acceptance rate: 19.7%*)

- CEDR: Contextualized Embeddings for Document Ranking
  Sean MacAvaney, Andrew Yates, Arman Cohan, Nazli Goharian
  SIGIR 2019: ACM SIGIR Conference on Research and Development in IR (*Acceptance rate: 19.7%*)

- Structural Scaffolds for Citation Intent Classification in Scientific Publications
  Arman Cohan, Waleed Ammar, Madeleine van Zuylen, Field Cady
  NAACL 2019: North America chapter of Association for Comp. Linguistics (*Acceptance rate: 22.6%*)

- Relation Extraction for Protein-protein Interactions Affected by Mutations
  Ziling Fan, Luca Soldaini, Arman Cohan, Nazli Goharian
  ACM-BCB 2019: Bioinformatics, Computational Biology, and Health Informatics (*Acceptance rate: 27%*)

- SMHD: a Large-Scale Resource for Exploring Online Language Usage for
  Multiple Mental Health Conditions
  Arman Cohan*, Bart Desmet*, Andrew Yates*, Luca Soldaini, Sean MacAvaney, and Nazli Goharian
  COLING 2018: Conference on Computational Linguistics (*Acceptance rate: 37.4%*)
  Equal contribution*, ***Area Chair Favorite Paper Award*

- A Discourse-Aware Attention Model for Abstractive Summarization of Long Documents
  Arman Cohan, Franck Dernoncourt, Doo S. Kim, Trung Bui, Seokhwan Kim, Walter Chang, Nazli Goharian
  NAACL 2018: North American Chapter of the Association for Computational Linguistics (*Acceptance rate*: 29.5%)

- Characterizing Question Facets for Complex Answer Retrieval.
  Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian,and Ophir Frieder
  SIGIR 2018: ACM SIGIR Conference on Research and Development in IR (*Acceptance rate*: 21%)

- Depression and Self-Harm Risk Assessment in Online Forums
Andrew Yates*, Arman Cohan*, and Nazli Goharian
EMNLP 2017: Empirical Methods for Natural Language Processing (*Acceptance rate*: 29.5%)
Equal contribution*, ***Best Paper Award*

- Contextualizing Citations for Scientific Summarization using Word Embeddings and Domain Knowledge
Arman Cohan and Nazli Goharian
SIGIR 2017: ACM SIGIR Conference on Research and Development in IR (*Acceptance rate*: 30%)

- Identifying Harm Events in Clinical Care through Medical Narratives
Arman Cohan, Allan Fong, Raj Ratwani, and Nazli Goharian
ACM-BCB 2017: Bioinformatics and Health informatics (*Acceptance rate*: 32%).

- A Neural Attention Model for Categorizing Patient Safety Events
Arman Cohan, Allan Fong, Nazli Goharian, and Raj Ratwani
ECIR 2017 European Conference on Information Retrieval (*Acceptance rate*: 27%).

- Revisiting Summarization Evaluation for Scientific Articles
Arman Cohan and Nazli Goharian
LREC 2016: Language Resources and Evaluation (*Acceptance rate*: 60%)

- Scientific Article Summarization Using Citation-Context and Article's Discourse Structure
Arman Cohan and Nazli Goharian
EMNLP 2015: Empirical Methods for Natural Language Processing *Acceptance rate*: 26%)

- Matching Citation Text and Cited Spans in Biomedical Literature: a Search-Oriented Approach
Arman Cohan, Luca Soldaini, and Nazli Goharian
NAACL 2015: North America chapter of Association for Comp. Linguistics (*Acceptance rate*: 22.1%).

- Retrieving Medical Literature for Clinical Decision Support
Luca Soldaini, Arman Cohan, Andrew Yates, Nazli Goharian, and Ophir Frieder
ECIR 2015: European Conference on Information Retrieval (*Acceptance rate*: 23%).

- On Clinical Decision Support
Arman Cohan, Luca Soldaini, Andrew Yates, Nazli Goharian, and Ophir Frieder.
ACM-BCB: Bioinformatics, Computational Biology, and Health Informatics (*Acceptance rate: 34%*).

Pre-prints

- Simplified Data Wrangling with ir_datasets
Sean MacAvaney, Andrew Yates, Sergey Feldman, Doug Downey, Arman Cohan, Nazli Goharian
ArXiv pre-print (in submission), 2021.

- CDLM: Cross Document Language Modeling
Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E. Peters, Arie Cattan, Ido Dagan
ArXiv pre-print (in submission), 2021.

- ABNIRML: Analyzing the Behavior of Neural IR Models
Sean MacAvaney, Sergey Feldman, Nazli Goharian, Doug Downey, Arman Cohan
ArXiv pre-print (in submission), 2020.

- ParsiNLU: A Suite of Language Understanding Challenges for Persian
Daniel Khashabi, Arman Cohan, Siamak Shakeri, Pedram Hosseini, Pouya Pezeshkpour, +22 Authors
ArXiv pre-print (in submission), 2020.

- Longformer: The Long-Document Transformer
  Iz Beltagy*, Matthew E. Peters*, Arman Cohan*
  ArXiv pre-print (in submission), 2020. *Equal contribution*

- SLEDGE: A Simple Yet Effective Baseline for COVID-19 Scientific Knowledge Search
  Sean MacAvaney, Arman Cohan, Nazli Goharian
  ArXiv pre-print, 2020

Journal papers

- Scientific Document Summarization via Citation Contextualization and Scientific Discourse
  Arman Cohan and Nazli Goharian
  International Journal on Digital Libraries (IJDL), 2018.

- Overcoming Low-utility Facets for Complex Answer Retrieval
  Sean MacAvaney, Andrew Yates, Arman Cohan, Luca Soldaini, Kai Hui, Nazli Goharian, Ophir Frieder
  Information Retrieval Journal, 2018.

- Triaging Content Severity in Online Mental-Health Forums
  Arman Cohan, Sydney Young, Andrew Yates, Nazli Goharian
  Journal of the Association for Information Science and Technology (JASIST), 2017.

Workshop and Demo papers

- On Generating Extended Summaries of Long Documents
  Sajad Sotudeh Gharebagh, Arman Cohan, Nazli Goharian
  AAAI 2021 Scientific Document Understanding workshop

- SUPP.AI: finding evidence for supplement-drug interactions
  Lucy Lu Wang, Oyvind Tafjord, Arman Cohan, Sarthak Jain, Sam Skjonsberg, Carissa Schoenick, Nick Botner, Waleed Ammar
  ACL 2020 Demo

- Learning to Generate Long Summaries from Scientific Documents
  Sajad Sotudeh Gharebagh, Arman Cohan, Nazli Goharian
  EMNLP 2020 SDP Workshop on Scholarly Document Processing

- Extracting evidence of supplement-drug interactions from literature
  Lucy Lu Wang, Oyvind Tafjord, Sarthak Jain, Arman Cohan, Sam Skjonsberg, Carissa Schoenick, Nick Botner, Waleed Ammar less
  NeurIPS 2019 ML for Health Workshop (ML4H)

- Helping or Hurting? Predicting Changes in Users' Risk of Self-Harm Through Online Community Interactions. Luca Soldaini, Timothy Walsh, Arman Cohan, Julien Han, and Nazli Goharian.
  NAACL 2018 Workshop of Computational Linguistics and Clinical Psychology Workshop (CLPsych)

- RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses
  Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian
  NAACL 2018 Workshop of Computational Linguistics and Clinical Psychology Workshop (CLPsych)

- Tree-LSTMs for Scientific Relation Classification
  Sean MacAvaney, Luca Soldaini, Arman Cohan, and Nazli Goharian
  SemEval 2018: Workshop on Semantic Evaluation

- A Framework for Cross-Domain Clinical Temporal Information Extraction
  Sean MacAvaney, Arman Cohan and Nazli Goharian
  SemEval 2017: Workshop on Semantic Evaluation

- Triaging Mental Health Forum Posts
  Arman Cohan, Sydney Young, and Nazli Goharian
  NAACL 2016 Workshop of Computational Linguistics and Clinical Psychology Workshop (CLPsych)

- Temporal Information Processing in Clinical Narratives
  Arman Cohan, Kevin Meurer, and Nazli Goharian
  SemEval 2016: Workshop on Semantic Evaluation

- Identifying Significance of Discrepancies in Radiology Reports
  Arman Cohan, Luca Soldaini, and Nazli Goharian, Allan Fong, Ross Filice, Raj Ratwani
  SDM 2016 Workshop on data Mining for Medicine and Healthcare (SDM-DMMH)

Patent applications

- Abstractive Summarization of Long Documents using Deep Learning
  Patent Application, 2018, U.S. Patent Application No. 15/915,775 – Published
  Arman Cohan, Walter W. Chang, Trung Huu Bui, Franck Dernoncourt and Doo Soon Kim
- Systems and Methods for Targeted Radiology Resident Training"
  Assignee: Medstar Institute for Innovation (MI2),
  Patent Application, 2017, PCT No. 15/410,850. – Published
  Raj Ratwani, Allan Fong, Ross Filice, Arman Cohan, Luca Soldaini, Nazli Goharian, and Ophir Frieder

Research Experience

Research Scientist

- **Allen Institute for Artificial Intelligence, Seattle, WA** *June 2018 - Present*
  Developing Natural Language Processing capabilities for addressing information overload

Doctoral Student

- **Georgetown University, Washington DC, USA**
  Computer Science *2013 - 2018*
  Dissertation: Text Summarization and Categorization for Scientific and Health-related Data
  Advisor: Dr. Nazli Goharian

Research Internships

- **Adobe Research**, San Jose, CA *Summer 2017*
  Mentor: Walter Chang
  Summarization of Long and Structured Documents

- **Medstar Health**, Washington, DC *Summer 2016*
  Mentor: Raj Ratwani
  Identifying Harm in Patient Safety Reports

- **Medstar Health**, Washington, DC *Summer 2015*
  Mentor: Raj Ratwani
  Identifying Critical Discrepancies in Medical Notes

Teaching, Mentoring and Invited talks

Professional Development

- Completed the Apprenticeship in Teaching (AT) Program *2015-2018*
  Center for New Designs in Learning and Scholarship (CNDLS)
  Georgetown University, Washington, D.C.

  Workshops completed:

  - Introduction to Teaching Resources
  - Syllabus Design
  - Assessment and Grading
  - Teaching Portfolio
  - Effective Classroom Interaction
  - Building Intellectual Communities in Large Classes
  - Inclusive Pedagogies: Designing to Engage Diversity

Instructor

- Text Mining & Analysis, Georgetown University *Fall 2017*
  Co-taught graduate-level course – prepared and gave lectures, managed TAs, and prepared exams
- Health Search and Mining, Georgetown University *Spring 2017*
  Co-taught graduate-level course – prepared and gave lectures, project prepration and discussions
- Database Systems Practicals, Amirkabir University of Technology *Fall 2012*
  Instructor of the course

Teaching Assistant

- Data Mining, Georgetown University *Spring 2015, 2016, 2017, 2018*
- Intro. to Information Retrieval, Georgetown University *Fall 2014, 2015, 2016, Spring 2018*
- Database Systems, Georgetown University *Spring 2015*
- Intro. to Information Systems, Georgetown University *Spring 2014*
- Intro. to Information Systems, Georgetown University *Spring 2014*
- Intro. to Information Systems, Georgetown University *Spring 2014*
- Intro. to e-Learning Technologies, Amirkabir University of Technology *Spring 2012*

Invited Talks

- Adapting Transformer models for Document-level Natural Language Tasks *Oct. 2020*
  Georgetown University, Washington DC
- Adapting Transformer models for Document-level Natural Language Tasks *Jun. 2020*
  Naverlabs Europe, France
- Towards Better Scientific Language Understanding *Mar. 2020*
  Ubiquitous Knowledge Processing (UKP), Germany
- Representation Learning of Scientific Papers from Citations *Oct. 2019*
  AI2, Seattle, WA
- Summarization of Long Documents using Deep Learning *Aug. 2017*
  Adobe, San Jose, CA

- Scientific Document Summarization *Oct. 2015*
  Instituto Gulbenkian de Ciencia (IGC), Portugal

Student mentoring

- Sean MacAvaney (PhD student), Research Intern at AI2 *2020*
- Anne Lauscher (PhD student), Research Intern at AI2 *2020*
- Isabel Cachola (Undergraduate), Predoctoral Young Investigator at AI2 *2019,2020*
- Kevin Henner (Masters student), Masters Thesis Supervision, University of Washington *2019*
- Tim Walsh (Masters Student), Georgetown *2018*
- Meng Han (Masters Student), Georgetown *2018*
- Sydney Young (Undergraduate student), Project Supervision, Georgetown *2016*
- Kevin Meurer, (Undergraduate student) Project Supervision, Georgetown *2016*

Professional Leadership & Services

Workshop organization

- SDP: Scholarly Document Processing at NAACL 2021 *2021*
- SciNLP: Scientific NLP workshop at AKBC 2020 *2020*
- MASC: Mid-Atlantic Student Colloquium on Speech, Language and Learning (MASC-SLL) *2017*

Tutorials

- NAACL 2021: NLP for Long Sequences *2021*

Area Chair

- NAACL: North American Chapter of Association for Computational Linguistics *2021*
- ICLR: International Conference on Learning Representations *2021*
- ACL: Assocication for Computational Linguistics *2020*

Thesis committee

- Sajad Sotoudeh (PhD), Georgeotown University *2021*
- Kevin Henner (Masters), University of Washington *2019,2020*

Journal Reviewer

- TACL: Transactions of Association for Computational Linguistics *2020*
- LREV: Language Resources and Evaluation *2020*
- NLE: Natural Language Engineering *2016-2019*

Program Committee - Conferences

- ACL: Association for Computational Linguistics *2018,2019,2020*
- EMNLP: Empirical Methods for Natural Language Processing *2018,2019,2020*
- NAACL: North American Chapter of ACL *2019*

- AAAI: Association for the Advancement of Artificial Intelligence *2017,2019*
- CoNLL: Conference on Computational Natural Language Learning *2017*
- IJCAI: International Joint Conference on Artificial Intelligence *2019*
- COLING: International Conference on Computational Linguistics *2018*
- CIKM: Conference on Information and Knowledge Management *2019*
- SIGIR: ACM Conference on Research and Development of IR *2018*

Program Committee - Workshops

- CLPsych: Computational Linguistics and Clinical Psychology Workshop, @NAACL *2019,2021*
- ML4H: Machine Learning for Healthcare @NeurIPS *2020*
- W-NUT: Noisy User-generated Text @EMNLP *2018,2019*
- BIRNDL: Bibliometric-enhanced IR and NLP @SIGIR *2018, 2019*

Technical Reviewer

- Technical Book Reviewer: Natural Language Processing with TensorFlow 2 *2021*