Mahsa Yarmohammadi

CONTACT	3400 North Charles Street Baltimore, MD 21218 USA	*phone:* +1 (503) 737 5075 *e-mail:* mahsa.yarmohamadi@gmail.com

RESEARCH INTERESTS
Natural Language Processing (NLP), Automatic Speech Recognition (ASR), Machine Learning, Deep Learning, Information Extraction, Information Retrieval, Machine Translation, Parsing.

CURRENT POSITION

Assistant Research Scientist 02/2021-present

Johns Hopkins University
Center for Language and Speech Processing

EDUCATION

Postdoctoral Fellow 12/2017-01/2021

Johns Hopkins University
Center for Language and Speech Processing
Mentors: Benjamin Van Durme, Sanjeev Khudanpour, Daniel Povey.

Ph.D. in Computer Science & Engineering 09/2009-09/2016

Oregon Health & Science University
Department of Computer Science and Electrical Engineering
Center for Spoken Language Understanding

Thesis: *Incremental Segmentation and Annotation Strategies for Real-time Natural Language Processing Applications*
Advisor: Brian Roark

M.Sc. in Computer (Software) Engineering 09/2004-09/2007

Shahid Beheshti University
Department of Computer Engineering

Thesis: *Extracting the Final Answer from Retrieved Documents in a QA System*
Advisor: Mehrnoush Shamsfard

B.Sc. in Computer (Software) Engineering 09/1999-09/2004

Amirkabir University of Technology (Tehran Polytechnic)
Department of Computer Engineering and Information Technology

Thesis: *Designing and Implementing an Expert Tool for Verifying and Validating the Knowledge Represented in OAV Triples*
Advisors: Ahmad Abdollahzadeh Barforoush, Chitra Dadkhah

EXPERIENCE

Johns Hopkins University 12/2017-01/2021
Postdoctoral Fellow. Center for Language and Speech Processing

- *Researcher / R&D Manager* – 10 to 20-member teams including senior faculty, staff, and students.
  - Developing a schema-based AI system to identify complex events described in multimedia inputs.

- Deep learning methods for extracting semantic information, in the form of who-did-what-to-whom-when-where, across multiple languages.
- Docker packaging the systems to be run on the Test & Evaluation platform.
- Developing an NLP-backed informational bot (ChatBot) for comprehensive COVID-19 information and misinformation.
- Developing, integrating, end-to-end testing and analysis of an "English-in, English-out" information retrieval system that, given an English query, retrieves relevant data from a multilingual speech and text repository.
- Building robust neural net models for domain classification in low-resource settings.
- Synthetic methods for training data generation for domain classifiers and crowdsourced curation.

- *Researcher in the Speech Recognition team.*
  - Designing and developing graph algorithms for Automatic Speech Recognition.
  - Developing and code-reviewing for the Kaldi speech recognition toolkit.

- *Senior Affiliate.* Frederick Jelinek Memorial Summer Workshop on Human Language Technology
  - 2019: Making neural machine translation systems robust for translation of informal text.
  - 2018: Lecturing a tutorial on "Building a State-of-the-Art ASR System with Kaldi".

- *Senior Researcher.* Human Language Technology Center of Excellence
  - Leading a team of students to develop speech translation systems for high-resource languages using extant broadcast news and conversations data.
  - Packaging and deployment of the speech translation system to the sponsor.
  - Leading student interns in the project of handling label sparsity and inconsistency in neural Named Entity Recognition.
  - Leading a team to build a TensorFlow-based framework for transfer and multi-task deep learning.

Intel Corporation – Hillsboro, OR, USA   08/2015-12/2017

Software Architect
- Graph and neural net based decoding algorithms for hardware accelerators for speech and NLP.
- Optimized implementation of dependency parsing using Intel deep neural net accelerator.
- Designing the architecture of a multi-modal language understanding system for future IP blocks.
- Implementing a Python wrapper for Intel deep neural net accelerator library.
- Leading the effort to capture industry trends.

Oregon Health & Science University   09/2009-05/2016

Graduate Research Assistant. Center for Spoken Language Understanding
- Optimal (syntax-aware) input segmentation for high-accuracy/low-latency real-time processing.
- Real-time speech-to-speech translation (in collaboration with AT&T Labs-Research).
- Applications of lexicographic semirings in speech and NLP.
- Regular approximation of Context-Free Grammars.
- Discriminative training of the language model for ASR.
- Creating NLP tools: MEMM tagger, Viterbi decoding, syntax-tree transformer, shift-reduce parser, modules for OpenGrm library.

AT&T Labs, Inc. - Research   07/2012-09/2012

Summer Intern. Speech and NLP group
- Real-time (incremental) speech-to-speech translation.
- Harvesting parallel text in multiple languages from web for machine translation.

GAM Electronics Company – Tehran, Iran   08/2004-12/2008

Java/J2EE developer

- Developing large-scale enterprise office automation applications.
- Tools: J2EE, Oracle 10g, Socket Programming.

Shahid Beheshti University   01/2006-12/2009

Graduate Research Assistant. Natural Language Processing Lab
- SBUQA question answering system.
- Building a WordNet for Persian verbs.
- Designing and building the database and interface of Persian WordNet (FarsNet 1).
- Database Systems Lab course Teaching Assistant, Oracle Database.

Amirkabir University of Technology   09/2002-09/2004

Undergraduate Research Assistant. Intelligent Systems Laboratory
- Implementing knowledge representation tools for expert systems.

JOURNAL ARTICLES

H. Lv, D. Povey, **M. Yarmohammadi**, K. Li, Y. Wang, L. Xie, and S. Khudanpur, "LET-Decoder: A WFST-based Lazy-evaluation Token-group Decoder with Exact Lattice Generation." *IEEE Signal Processing Letters*, 2021.

R. Sproat, **M. Yarmohammadi**, I. Shafran and B. Roark, "Applications of Lexicographic Semirings to Problems in Speech and Language Processing," *Computational Linguistics*, 2014.

M. Shamsfard and **M. Yarmohammadi**, "A Semantic Approach to Extract the Final Answer in SBUQA Question Answering System," *International Journal of Digital Content Technology and its Applications*, 2010.

A. Hesabi, M. Assi, M. Shamsfard and **M. Yarmohammadi**, "Designing and Developing a Word net for Persian Nouns," *International Journal of Information & Communication Technology*, Iran Telecom Research Center, 2009.

CONFERENCE PAPERS

H. Xu, S. Ebner, **M. Yarmohammadi**, A. S. White, B. Van Durme and K. Murray, "Gradual Fine-Tuning for Low-Resource Domain Adaptation." In *Proceedings of EACL, The 2nd Workshop on Domain Adaptation for NLP*, 2021.

A. Singh, P. Xia, G. Qin, **M. Yarmohammadi** and B. Van Durme, "CopyNext : Explicit Span Copying and Alignment in Sequence to Sequence Models." In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP), The 4th Workshop on Structured Prediction for NLP*, 2020. *Contributed (spotlight) Talk.*

A. Poliak et. al., "Collecting Verified COVID-19 Question Answer Pairs." In *Proceedings of Empirical Methods in Natural Language Processing (EMNLP), NLP COVID-19 Workshop*, 2020.

Z. Chen, **M. Yarmohammadi**, H. Xu, H. Lv, L. Xie, D. Povey and S. Khudanpur, "Incremental Lattice Determinization for WFST Decoders." In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, Sentosa, Singapore, 2019. *Best paper award nominee.*

M. Yarmohammadi, X. Ma, S. Hisamoto,Y. Wang, H. Xu, D. Povey, P. Koehn and K. Duh. "Robust Document Representations for Cross-Lingual Information Retrieval in Low-Resource Settings." In *Proceedings of the 17th Machine Translation Summit*, Dublin, Ireland, 2019.

D. Povey, G. Cheng, Y. Wang, K. Li, H. Xu, **M. Yarmohammadi** and S. Khudanpur. "Semi-Orthogonal Low-Rank Matrix Factorization for Deep Neural Networks." *Conference of the Interna-*

tional Speech Communication Association (Interspeech), Hyderabad, India, 2018.

M. Yarmohammadi, A. Dunlop and B. Roark. "Transforming Trees into Hedges and Parsing with "Hedgebank" Grammars." In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, Baltimore, USA, 2014.

M. Yarmohammadi, V. K. R. Sridhar, S. Bangalore and B. Sankaran. "Incremental Segmentation and Decoding Strategies for Simultaneous Translation." In *Proceedings of the 6th International Joint Conference on Natural Language Processing (IJCNLP)*, Nagoya, Japan, 2013.

L. Barbosa, V. K. R. Sridhar, **M. Yarmohammadi** and S. Bangalore. "Harvesting Parallel Text in Multiple Languages with Limited Supervision." In *Proceedings of the 24th International Conference on Computational Linguistics (COLING)*, Mumbai, India, 2012.

I. Shafran, R. Sproat, **M. Yarmohammadi** and B. Roark. "Efficient Determinization of Tagged Word Lattices using Categorial and Lexicographic Semirings." In *Proceedings of IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU)*, Hawaii, USA, 2011.

M. Rouhizadeh, **M. Yarmohammadi** and M. Shamsfard. "Developing the Persian WordNet of Verbs; Issues of Compound Verbs and the Editor." In *Proceedings of the 5th Global WordNet Conference (GWC)*, Mumbai, India, 2010.

M. Yarmohammadi, M, Shamsfard, Mahshid Yarmohammadi, M. Rouhizadeh. "SBUQA Question Answering System", *Advances in Computer Science and Engineering*, Volume 6, 2009.

M. Yarmohammadi, M. Shamsfard, M. Rouhizadeh and Mahshid Yarmohammadi, "Using Word-Net in Extracting the Final Answer from Retrieved Documents in a Question Answering System." In *Proceedings of 4th Global WordNet conference (GWC)*, Szeged, Hungary, 2008.

M. Rouhizadeh, M. Shamsfard, **M. Yarmohammadi**. "Building a WordNet for Persian Verbs." In *Proceedings of the 4th Global WordNet conference (GWC)*, Szeged, Hungary, 2008.

M. Rouhizadeh, M. Assi, **M. Yarmohammadi**. "Designing Persian Verbs WordNet." In *Proceedings of the 2007 Iranian Conference on Linguistics*, Tehran, Iran, 2007.

C. Dadkhah, A. A. Barforoush and **M. Yarmohammadi**. "Design a Tool for Construction of Unique Knowledge Base for Expert Systems." In *Proceedings of the 18th International Conference on Systems Engineering (ICSENG)*, Las Vegas, USA, 2005.

OTHER PAPERS **M. Yarmohammadi**. "Incremental Hedge Parsing and Segmentation and its Evaluation on Simultaneous Machine Translation." The 4th Pacific Northwest Regional NLP Workshop (NW-NLP), Seattle, USA, 2016.

M. Yarmohammadi. "Discriminative Training with Perceptron Algorithm for POS Tagging Task." Technical Report #CSLU-2014-001, Center for Spoken Language Understanding, Oregon Health & Science University, USA, 2014.

M. Yarmohammadi. "Approximate Parsing with "Hedge Grammars"." The 3rd Pacific Northwest Regional NLP Workshop (NW-NLP), Redmond, USA, 2014.

M. Yarmohammadi, B. Roark, I. Shafran and R. Sproat. "A Summary and Comparison of Two Approaches for Determinization of Lattices." The 2nd Pacific Northwest Regional NLP Workshop (NW-NLP), Redmond, USA, 2012.

M. Yarmohammadi. "Semi-supervised Discriminative N-gram Language Modeling in ASR." CSLU Qualifying Exam Written Paper, Oregon Health & Science University, USA, 2011.

PROFESSIONAL
ACTIVITIES

Reviewing and Program Committee

Co-chair of the 4th Pacific Northwest Regional NLP Workshop: NW-NLP 2016.
AAAI Conference 2020, Program Committee.
ACL Conference 2019, 2020, Review Committee.
EACL Conference 2021, Review Committee.
EMNLP Conference 2019, 2020, Primary Reviewer.
ICASSP Conference 2017, 2018, 2019, 2020, Scientific Review Committee.
Interspeech Conference 2016, 2017, 2018, 2019, Scientific Review Committee.
Intel Spoken Language Technologies Summit 2017, Program Committee.
NAACL-HLT Conference 2013, 2015, 2019, Primary Reviewer.
NW-NLP Workshop 2014, Scientific Review Committee.

Organizing Committee

Interspeech Conference, 2012.
ACL-HLT Conference, 2011.

TECHNICAL
SKILLS

Programming: strong programming and debugging skills: C++, Python, Java, R, SQL, etc.

Language and Speech Processing Tools: OpenFST, Kaldi ASR toolkit, Moses SMT, CoreNLP, AllenNLP, NLTK, Lucene, Elasticsearch.

Distributed Computing: Condor, Hadoop, AWS, Docker.

Deep Learning Skills:
Word representations: Word2vec, fastText, GloVe.
Contextual word representations: ELMO, BERT, mBERT, XLM.
DL frameworks: TensorFlow, PyTorch, Fairseq, scikit-learn, spaCy, pandas.