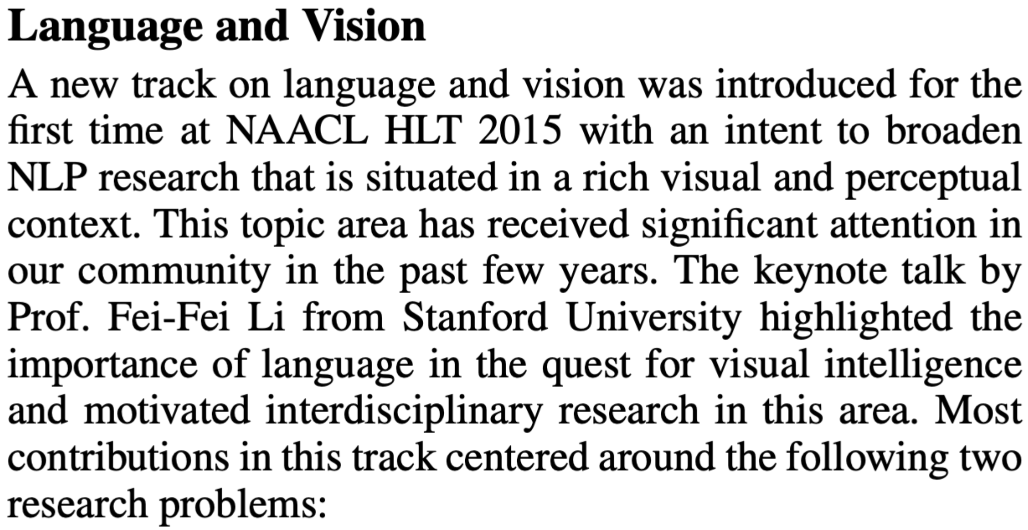
1st Workshop on Advances in Language and Vision Research (ALVR)

ACL 2020

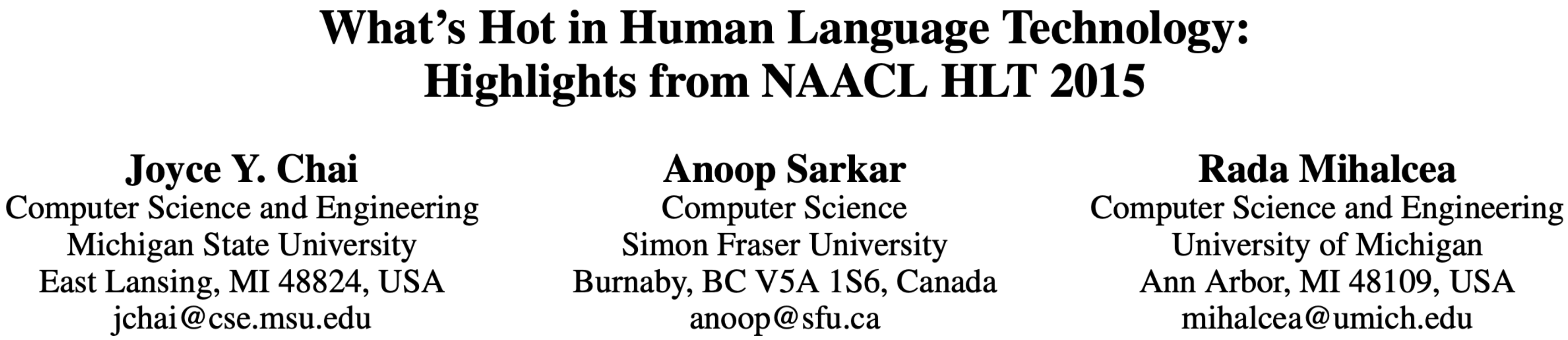
July 9, 2020

The First *Language and Vision* Track @NAACL 2015



Language and Vision

A new track on language and vision was introduced for the first time at NAACL HLT 2015 with an intent to broaden NLP research that is situated in a rich visual and perceptual context. This topic area has received significant attention in our community in the past few years. The keynote talk by Prof. Fei-Fei Li from Stanford University highlighted the importance of language in the quest for visual intelligence and motivated interdisciplinary research in this area. Most contributions in this track centered around the following two research problems:



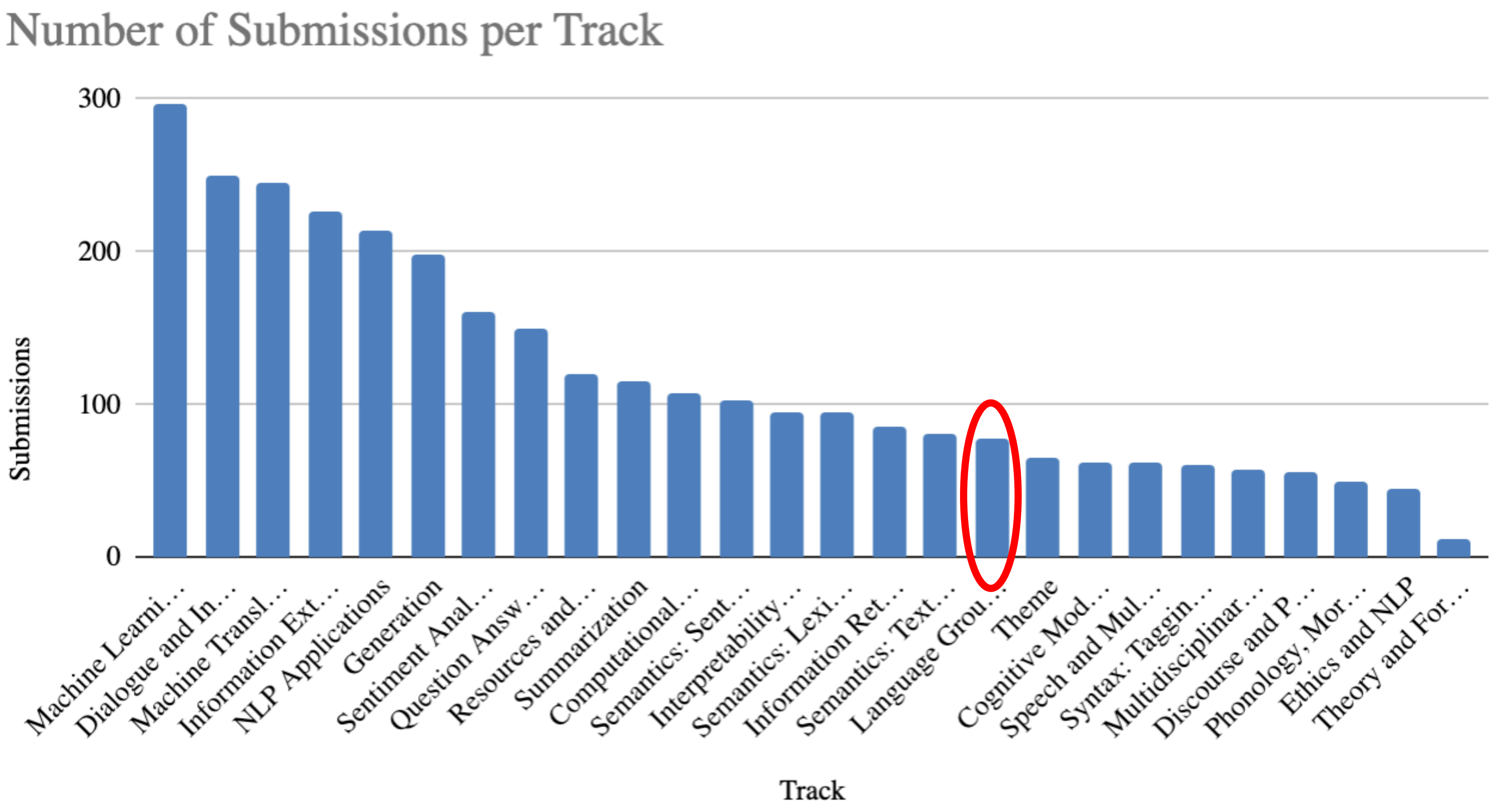
What's Hot in Human Language Technology: Highlights from NAACL HLT 2015

Joyce Y. Chai
Computer Science and Engineering
Michigan State University
East Lansing, MI 48824, USA
jchai@cse.msu.edu

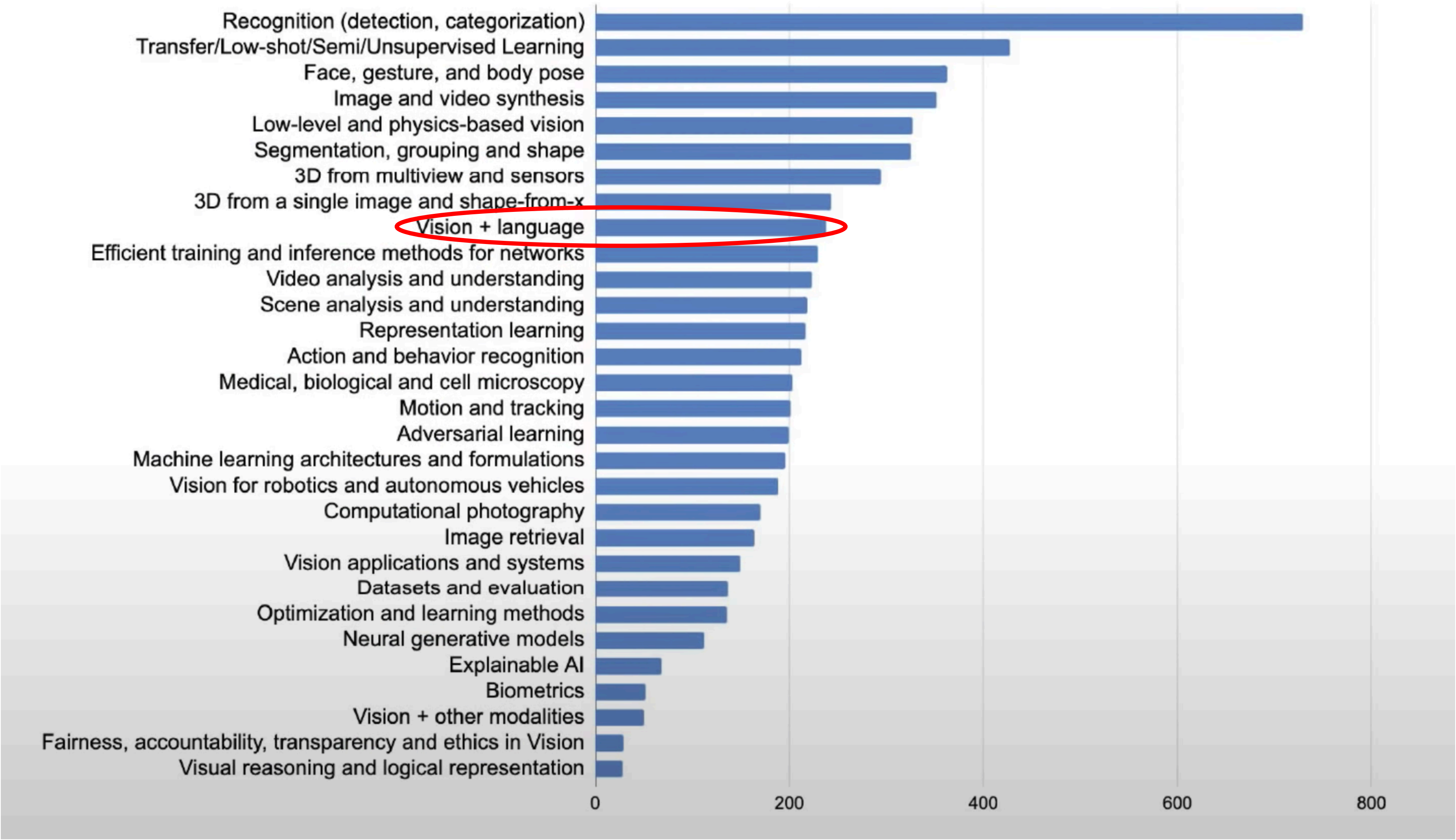
Anoop Sarkar
Computer Science
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
anoop@sfu.ca

Rada Mihalcea
Computer Science and Engineering
University of Michigan
Ann Arbor, MI 48109, USA
mihalcea@umich.edu

Language and Vision @ ACL 2020



Language and Vision @ CVPR 2020

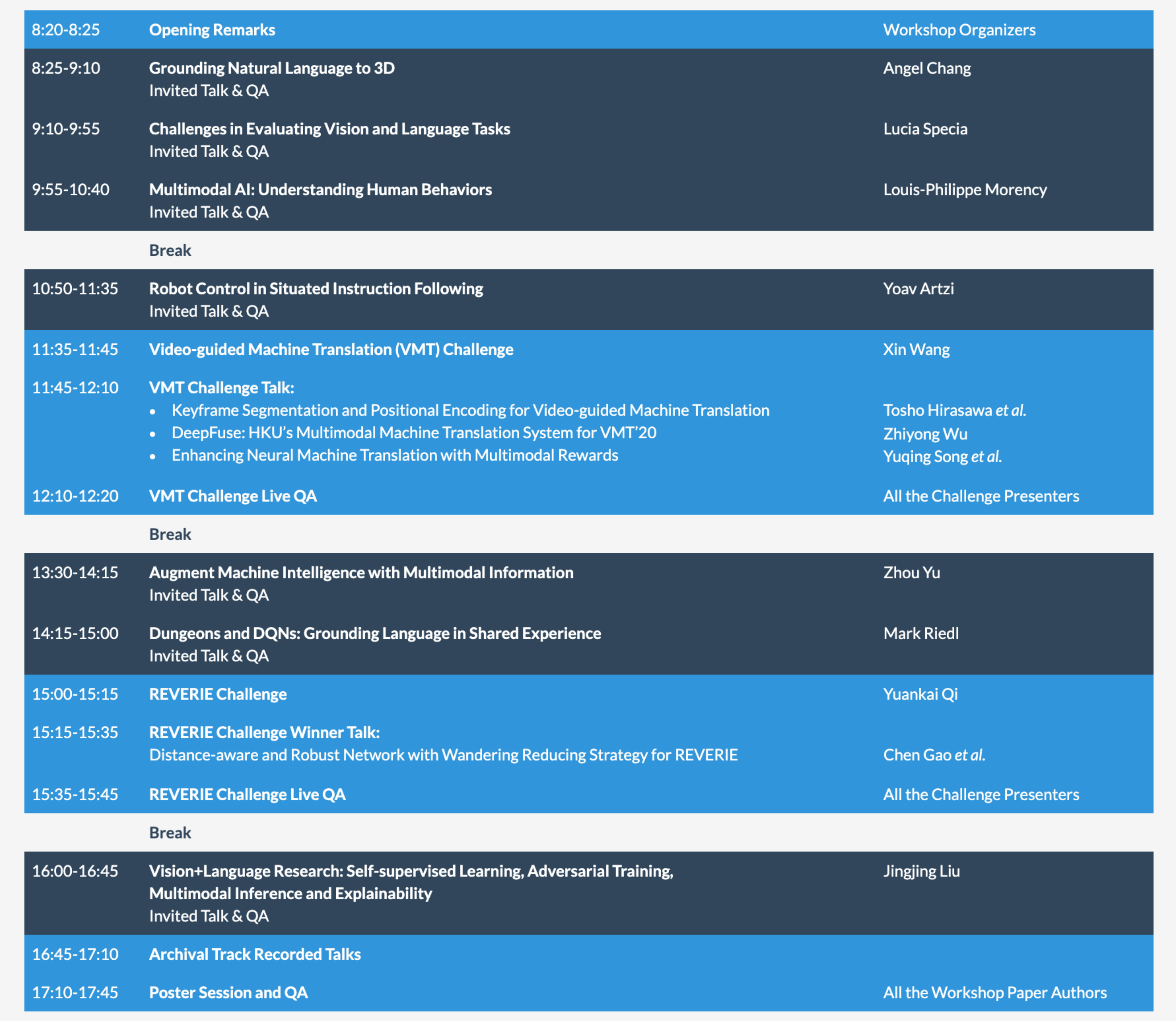


Advances in Language and Vision Research

- New tasks and datasets that provide real-world solutions in the intersection of NLP and CV;
- Language-guided interaction with the real world, such as navigation via instruction following or dialogue;
- External knowledge integration in visual and language understanding;
- Visually grounded multilingual study, for example multimodal machine translation;
- Shortcoming of existing language and vision tasks and datasets;
- Benefits of using multimodal learning in downstream NLP tasks;
- Self-supervised representation learning in language and vision;
- Transfer learning (including few/zero-shot learning) and domain adaptation;
- Cross-modal learning beyond image understanding, such as videos and audios;
- Multidisciplinary study that may involve linguistics, cognitive science, robotics, etc.

Program

- 7 Invited Talks
- 2 New Challenges + 4 Challenge Talks
- 5 Archival-track Recorded Talks
- Parallel Poster Session for All Accepted Papers



8:20-8:25	**Opening Remarks**	Workshop Organizers
8:25-9:10	**Grounding Natural Language to 3D** Invited Talk & QA	Angel Chang
9:10-9:55	**Challenges in Evaluating Vision and Language Tasks** Invited Talk & QA	Lucia Specia
9:55-10:40	**Multimodal AI: Understanding Human Behaviors** Invited Talk & QA	Louis-Philippe Morency
	Break	
10:50-11:35	**Robot Control in Situated Instruction Following** Invited Talk & QA	Yoav Artzi
11:35-11:45	**Video-guided Machine Translation (VMT) Challenge**	Xin Wang
11:45-12:10	**VMT Challenge Talk:** • Keyframe Segmentation and Positional Encoding for Video-guided Machine Translation • DeepFuse: HKU's Multimodal Machine Translation System for VMT'20 • Enhancing Neural Machine Translation with Multimodal Rewards	 Tosho Hirasawa *et al.* Zhiyong Wu Yuqing Song *et al.*
12:10-12:20	**VMT Challenge Live QA**	All the Challenge Presenters
	Break	
13:30-14:15	**Augment Machine Intelligence with Multimodal Information** Invited Talk & QA	Zhou Yu
14:15-15:00	**Dungeons and DQNs: Grounding Language in Shared Experience** Invited Talk & QA	Mark Riedl
15:00-15:15	**REVERIE Challenge**	Yuankai Qi
15:15-15:35	**REVERIE Challenge Winner Talk:** Distance-aware and Robust Network with Wandering Reducing Strategy for REVERIE	 Chen Gao *et al.*
15:35-15:45	**REVERIE Challenge Live QA**	All the Challenge Presenters
	Break	
16:00-16:45	**Vision+Language Research: Self-supervised Learning, Adversarial Training, Multimodal Inference and Explainability** Invited Talk & QA	Jingjing Liu
16:45-17:10	**Archival Track Recorded Talks**	
17:10-17:45	**Poster Session and QA**	All the Workshop Paper Authors

Invited Speakers (presentation order)

Angel Chang
Simon Fraser University

Lucia Specia
Imperial College London

Louis-Philippe Morency
CMU

Yoav Artzi
Cornell

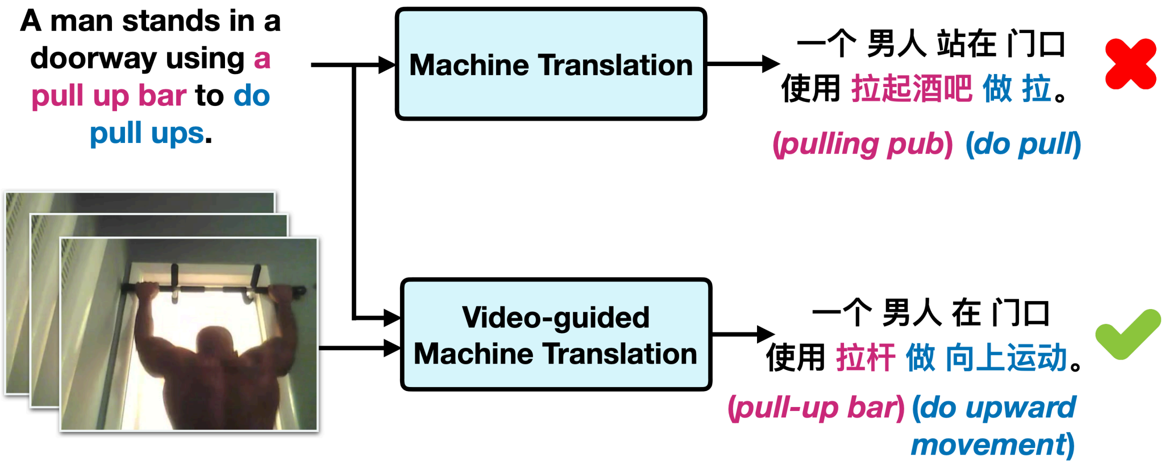
Zhou Yu
UC Davis

Mark Riedl
Georgia Tech

Jingjing (JJ) Liu
Microsoft

Two New Challenges

Video-guided Machine Translation (VMT) Challenge 2020



A man stands in a doorway using a pull up bar to do pull ups.

Machine Translation

一个 男人 站在 门口 使用 拉起酒吧 做 拉。 ✗

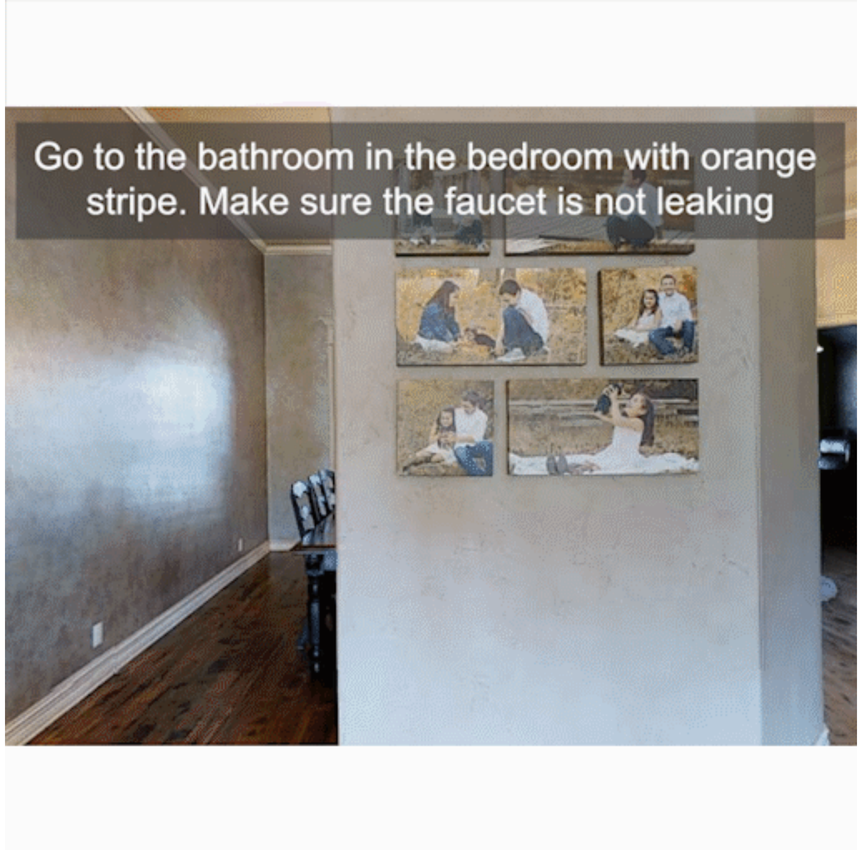
(pulling pub) (do pull)

Video-guided Machine Translation

一个 男人 在 门口 使用 拉杆 做 向上运动。 ✓

(pull-up bar) (do upward movement)

Remote Embodied Visual Referring Expression (REVERIE) Challenge 2020



Go to the bathroom in the bedroom with orange stripe. Make sure the faucet is not leaking

Accepted Papers

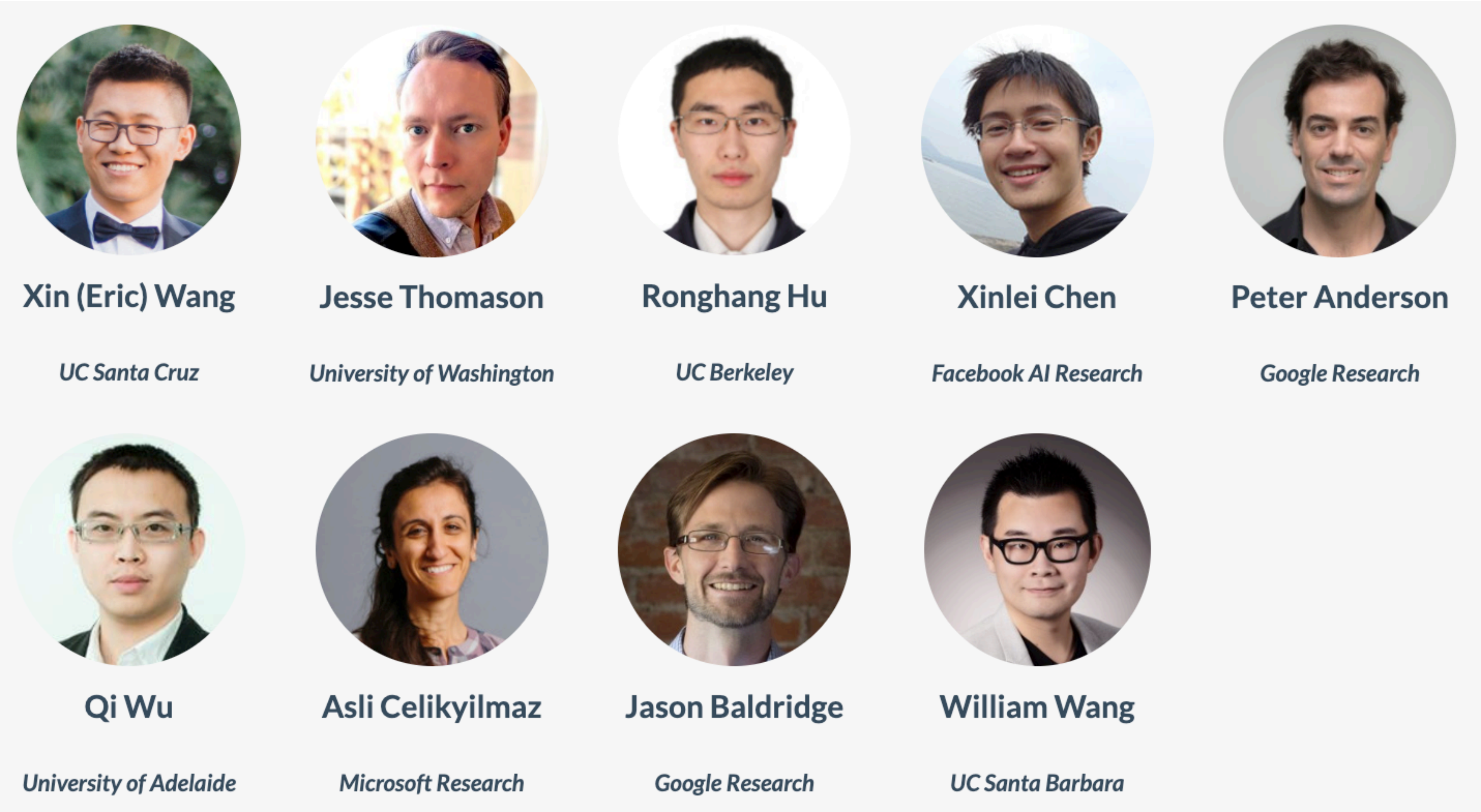
5 Papers accepted to the archival track:

- Extending ImageNet to Arabic using Arabic WordNet - *Abdulkareem Alsudais*
- Toward General Scene Graph: Integration of Visual Semantic Knowledge with Entity Synset Alignment - *Woo Suk Choi, Kyoung-Woon On, Yu-Jung Heo and Byoung-Tak Zhang*
- Visual Question Generation from Radiology Images - *Mourad Sarrouti, Asma Ben Abacha and Dina Demner-Fushman*
- On the role of effective and referring questions in GuessWhat?! - *Mauricio Mazuecos, Alberto Testoni, Raffaella Bernardi and Luciana Benotti*
- Latent Alignment of Procedural Concepts in Multimodal Recipes - *Hossein Rajaby Faghihi, Roshanak Mirzaee, Sudarshan Paliwal and Parisa Kordjamshidi*

15 Papers accepted to the non-archival track (including 4 Challenge Papers):

- Pix2R: Guiding Reinforcement Learning using Natural Language by Mapping Pixels to Rewards - *Prasoon Goyal, Scott Niekum and Raymond Mooney*
- TextCaps: a Dataset for Image Captioning with Reading Comprehension - *Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach and Amanpreet Singh*
- Improving VQA and its Explanations by Comparing Competing Explanations - *Jialin Wu, Liyan Chen and Raymond Mooney*
- Bridging Languages through Images with Deep Partial Canonical Correlation Analysis - *Guy Rotman, Ivan Vulić and Roi Reichart*
- Counterfactual Vision-and-Language Navigation via Adversarial Path Sampling - *Tsu-Jui Fu, Xin Wang, Matthew Peterson, Scott Grafton, Miguel Eckstein and William Yang Wang*
- Measuring Social Biases in Grounded Vision and Language Embeddings - *Candace Ross, Boris Katz and Andrei Barbu*
- Exploring Phrase Grounding without Training: Contextualisation and Extension to Text-Based Image Retrieval - *Letitia Parcalabescu and Anette Frank*
- What is Learned in Visually Grounded Neural Syntax Acquisition - *Noriyuki Kojima, Hadar Averbuch-Elor, Alexander Rush and Yoav Artzi*
- Learning to Map Natural Language Instructions to Physical Quadcopter Control Using Simulated Flight - *Valts Blukis, Yannick Terme, Eyvind Niklasson, Ross Knepper and Yoav Artzi*
- Learning Latent Graph Representations for Relational VQA - *Liyan Chen and Raymond Mooney*
- Entity Skeletons for Visual Storytelling - *Khyathi Raghavi Chandu, Ruo-Ping Dong and Alan W Black*

Organizers



Xin (Eric) Wang

UC Santa Cruz

Jesse Thomason

University of Washington

Ronghang Hu

UC Berkeley

Xinlei Chen

Facebook AI Research

Peter Anderson

Google Research

Qi Wu

University of Adelaide

Asli Celikyilmaz

Microsoft Research

Jason Baldridge

Google Research

William Wang

UC Santa Barbara