Introduction

Our workshop focuses on machine reading for question answering (MRQA), which has become an important testbed for evaluating how computer systems understand natural language, as well as a crucial technology for applications such as search engines and dialog systems. In recent years, research community has showed rapid progress on both datasets and models. Many large-scale datasets are proposed and the development of more accurate and more efficient question answering systems followed. Despite recent progress, yet there is much to be desired about these datasets and systems, such as model interpretability, ability to abstain from answering when there is no adequate answer, and adequate modeling of inference (e.g., entailment and multi-sentence reasoning).

This year, we focus on generalization of QA systems and present a new shared task on the topic. Our shared task addresses the following research question: how can one build a robust question answering system that can perform well questions from unseen domains? Train and test datasets may differ in passage distribution (from different sources (e.g., science, news, novels, medical abstracts, etc) with pronounced syntactic and lexical differences), question distribution (different styles (e.g., entity-centric, relational, other tasks reformulated as QA, etc) from different sources (e.g., crowdworkers, domain experts, exam writers, etc.)), as well as joint question-answering distribution (e.g., question collected independent vs. dependent of evidence).

For this task, we adapted and unified 18 distinct question answering datasets into the same format. We focus on extractive question answering. That is, given a question and context passage, systems must find a segment of text, or span in the document that best answers the question. While this format is somewhat restrictive, it allows us to leverage many existing datasets, and its simplicity helps us focus on out-of-domain generalization, instead of other important but orthogonal challenges. We released six larger datasets as training, and another six datasets for development. The rest six datasets were hidden from shared task participants until the final evaluation. Nine teams submitted to our shared task and the winning system achieved an average F1 score of 72.5 on the held-out datasets, 10.7 absolute points higher than our initial baseline based on BERT large.

This proceeding includes our report on the findings from this shared task as well as six system description papers from the shared task participants.

Similar to last year, we also sought research track submissions. We have received 39 paper submissions to the research track after the withdrawls, almost double the submission from last year. Out of this, twenty two papers are accepted and presented in this proceedings, and two papers are selected for the best paper award.

In the workshop program, we also include four cross submissions of work presented in other venues already.

The program features 22 new research track papers, six shared track papers and four cross-submissions from related areas, to be presented as either posters and talks. We are also excited to host remarkable invited speakers, including Mohit Bansal, Antoine Bordes, Jordan Boyd-Graber and Matt Gardner.

We thank the program committee, the EMNLP workshop chairs, the invited speakers, our sponsors Baidu, Facebook and NAVER and our steering committee: Jonathan Berant, Percy Liang, Luke Zettlemoyer.
Steering Committee:
Jonathan Berant, Tel Aviv University
Percy Liang, Stanford University
Luke Zettlemoyer, University of Washington

Organizers:
Adam Fisch, MIT
Alon Talmor, Tel Aviv University
Robin Jia, Stanford University
Minjoon Seo, NAVER & University of Washington
Eunsol Choi, University of Washington & Google
Danqi Chen, Princeton University

Program Committee:
Alane Suhr, Cornell University
Chenguang Zhu, Microsoft
Christopher Clark, University of Washington
Daniel Khashabi, Allen Institute for AI
Danish Contractor, IBM
Dheeru Dua, UC Irvine
Gabriel Stanovsky, University of Washington/Allen Institute for AI
Hoifung Poon, Microsoft
Huan Sun, Ohio State University
Jiahua Liu, Tsinghua University
Jing Liu, Baidu
Jinhyuk Lee, Korea University
Johannes Welbl, University College London
Jonathan Herzig, Tel Aviv University
Kai Sun, Cornell University
Karthik Narasimhan, Princeton University
Kenton Lee, Google
Kevin Gimpel, TTIC
Li Dong, Microsoft
Luheng He, Google
Mandar Joshi, University of Washington
Matt Gardner, Allen Institute for AI
Matthew Richardson, Microsoft
Mausam, IIT
Minghao Hu, NUDT
Mohit Iyyer, UMass
Mor Geva, Tel Aviv University
Mrinmaya Sachan, Carnegie Mellon University
Nan Duan, Microsoft
Ni Lao, Mosaix
Nitish Gupta, University of Pennsylvania
Omer Levy, Facebook
Oyvind Tafjord, Allen Institute for AI
Panupong Pasupat, Google
Patrick Lewis, University College London
Peng Qi, Stanford University
Pradeep Dasigi, Allen Institute for AI
Quan Wang, Baidu
Rajarshi Das, UMass
Saizheng Zhang, University of Montreal
Sameer Singh, UC Irvine
Scott Wen-tau Yih, Facebook
Sebastian Riedel, University College London/Facebook
Semih Yagcioglu, Hacettepe University
Sewon Min, University of Washington
Siva Reddy, Stanford University
Todor Mihaylov, Heidelberg University
Tom Kwiatkowski, Google
Tushar Khot, Allen Institute for AI
Adams Wei Yu, Carnegie Mellon University
Wenhan Xiong, UC Santa Barbara
Xiaodong Liu, Microsoft
Yichen Jiang, UNC Chapel Hill
Zhilin Yang, Carnegie Mellon University

Invited Speaker:

Mohit Bansal, UNC Chapel Hill
Antoine Bordes, Facebook AI Research
Jordan Boyd-Graber, University of Maryland
Matt Gardner, Allen Institute for AI
## Table of Contents

**MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension**  
Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi and Danqi Chen  
1

**Inspecting Unification of Encoding and Matching with Transformer: A Case Study of Machine Reading Comprehension**  
Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Lei Cui, Songhao Piao and Ming Zhou  
14

**CALOR-QUEST: generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations**  
FREDERIC BECHET, Cindy Aloui, Delphine Charlet, Geraldine Damnati, Johannes Heinecke, Alexis Nasr and Frederic Herledan  
19

**Improving Subject-Area Question Answering with External Knowledge**  
Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie and Dong Yu  
27

**Answer-Supervised Question Reformulation for Enhancing Conversational Machine Comprehension**  
Qian Li, Hui Su, CHENG NIU, Daling Wang, Zekang Li, Shi Feng and yifei zhang  
38

**Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering**  
Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Hong Wang, Shiyu Chang, Murray Campbell and William Yang Wang  
48

**Improving the Robustness of Deep Reading Comprehension Models by Leveraging Syntax Prior**  
Bowen Wu, Haoyang Huang, Zongsheng Wang, Qihang Feng, Jingsong Yu and Baoxun Wang  
53

**Reasoning Over Paragraph Effects in Situations**  
Kevin Lin, Oyvind Tafjord, Peter Clark and Matt Gardner  
58

**Towards Answer-unaware Conversational Question Generation**  
Mao Nakanishi, Tetsunori Kobayashi and Yoshihiko Hayashi  
63

**Cross-Task Knowledge Transfer for Query-Based Text Summarization**  
Elozino Egonmwan, Vittorio Castelli and Md Arafat Sultan  
72

**Book QA: Stories of Challenges and Opportunities**  
Stefanos Angelidis, Lea Frermann, Diego Marcheggiani, Roi Blanco and Lluís Márquez  
78

**FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension**  
Yi-Ting Yeh and Yun-Nung Chen  
86

**Do Multi-hop Readers Dream of Reasoning Chains?**  
Haoyu Wang, Mo Yu, Xiaoxiao Guo, Rajarshi Das, Wenhan Xiong and Tian Gao  
91

**Machine Comprehension Improves Domain-Specific Japanese Predicate-Argument Structure Analysis**  
Norio Takahashi, Tomohide Shibata, Daisuke Kawahara and Sadao Kurohashi  
98

**On Making Reading Comprehension More Comprehensive**  
Matt Gardner, Jonathan Berant, Hannaneh Hajishirzi, Alon Talmor and Sewon Min  
105
Multi-step Entity-centric Information Retrieval for Multi-Hop Question Answering  
Rajarshi Das, Ameya Godbole, Dilip Kavarthapu, Zhiyu Gong, Abhishek Singhal, Mo Yu, Xiaoxiao Guo, Tian Gao, Hamed Zamani, Manzil Zaheer and Andrew McCallum ........................................... 113

Evaluating Question Answering Evaluation  
Anthony Chen, Gabriel Stanovsky, Sameer Singh and Matt Gardner ......................... 119

Bend but Don’t Break? Multi-Challenge Stress Test for QA Models  
Hemant Pugaliya, James Route, Kaixin Ma, Yixuan Geng and Eric Nyberg ................. 125

ReQA: An Evaluation for End-to-End Answer Retrieval Models  
Amin Ahmad, Noah Constant, Yinfei Yang and Daniel Cer ................................. 137

Comprehensive Multi-Dataset Evaluation of Reading Comprehension  
Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner and Sameer Singh .... 147

A Recurrent BERT-based Model for Question Generation  
Ying-Hong Chan and Yao-Chung Fan ...................................................................... 154

Let Me Know What to Ask: Interrogative-Word-Aware Question Generation  
Junmo Kang, Haritz Puerto San Roman and sung-hyon myaeng ............................ 163

Extractive NarrativeQA with Heuristic Pre-Training  
Lea Frermann ......................................................................................................... 172

CLER: Cross-task Learning with Expert Representation to Generalize Reading and Understanding  
Takumi Takahashi, Motoki Taniguchi, Tomoki Taniguchi and Tomoko Ohkuma ....... 183

Question Answering Using Hierarchical Attention on Top of BERT Features  
Reham Osama, Nagwa El-Makky and Marwan Torki ............................................. 191

Domain-agnostic Question-Answering with Adversarial Training  
Seanie Lee, Dongyu Kim and Jangwon Park ....................................................... 196

Generalizing Question Answering System with Pre-trained Language Model Fine-tuning  
Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu and Pascale Fung . 203

D-NET: A Pre-Training and Fine-Tuning Framework for Improving the Generalization of Machine Reading Comprehension  
Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang, Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu and Haifeng Wang ................................................................. 212

An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering  
Shayne Longpre, Yi Lu, Zhucheng Tu and Chris DuBois ......................................... 220
Conference Program

Monday, November 4, 2019

9:00–9:35  Invited talk I: Antoine Bordes

9:35–10:10 Invited talk II: Matt Gardner

10:10–10:30 Best paper session I: Multi-step Entity-centric Information Retrieval for Multi-Hop Question Answering

10:30–11:00 Coffee break

11:00–11:35 Invited talk III: Jordan Boyd-Graber

Shared task

11:35–12:10 MRQA 2019 Shared Task: Evaluating Generalization in Reading Comprehension
Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi and Danqi Chen

12:10–12:30 Shared task best system session: D-NET: A Pre-Training and Fine-Tuning Framework for Improving the Generalization of Machine Reading Comprehension

12:30–14:00 Lunch break

14:00-14:20 Best paper session II: Evaluating Question Answering Evaluation

14:20–14:55 Invited talk IV: Mohit Bansal
Poster session

14:55–16:30 *Inspecting Unification of Encoding and Matching with Transformer: A Case Study of Machine Reading Comprehension*
Hangbo Bao, Li Dong, Furu Wei, Wenhui Wang, Nan Yang, Lei Cui, Songhao Piao and Ming Zhou

14:55–16:30 *CALOR-QUEST : generating a training corpus for Machine Reading Comprehension models from shallow semantic annotations*
FREDERIC BECHET, Cindy Aloui, Delphine Charlet, Geraldine Damnati, Johannes Heinecke, Alexis Nasr and Frederic Herledan

14:55–16:30 *Improving Subject-Area Question Answering with External Knowledge*
Xiaoman Pan, Kai Sun, Dian Yu, Jianshu Chen, Heng Ji, Claire Cardie and Dong Yu

14:55–16:30 *Answer-Supervised Question Reformulation for Enhancing Conversational Machine Comprehension*
Qian Li, Hui Su, CHENG NIU, Daling Wang, Zekang Li, Shi Feng and yifei zhang

14:55–16:30 *Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering*
Wenhan Xiong, Mo Yu, Xiaoxiao Guo, Hong Wang, Shiyu Chang, Murray Campbell and William Yang Wang

14:55–16:30 *Improving the Robustness of Deep Reading Comprehension Models by Leveraging Syntax Prior*
Bowen Wu, Haoyang Huang, Zongsheng Wang, Qihang Feng, Jingsong Yu and Baoxun Wang

14:55–16:30 *Reasoning Over Paragraph Effects in Situations*
Kevin Lin, Oyvind Tafjord, Peter Clark and Matt Gardner

14:55–16:30 *Towards Answer-unaware Conversational Question Generation*
Mao Nakanishi, Tetsunori Kobayashi and Yoshihiko Hayashi

14:55–16:30 *Cross-Task Knowledge Transfer for Query-Based Text Summarization*
Elozino Egonmwan, Vittorio Castelli and Md Arafat Sultan

14:55–16:30 *Book QA: Stories of Challenges and Opportunities*
Stefanos Angelidis, Lea Frermann, Diego Marcheggiani, Roi Blanco and Lluís Márquez

14:55–16:30 *FlowDelta: Modeling Flow Information Gain in Reasoning for Conversational Machine Comprehension*
Yi-Ting Yeh and Yun-Nung Chen
<table>
<thead>
<tr>
<th>Time</th>
<th>Title</th>
<th>Authors</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:55–16:30</td>
<td>Do Multi-hop Readers Dream of Reasoning Chains?</td>
<td>Haoyu Wang, Mo Yu, Xiaoxiao Guo, RajarshiDas, Wenhan Xiong and Tian Gao</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>Evaluating Question Answering Evaluation</td>
<td>Anthony Chen, Gabriel Stanovsky, Sameer Singh and Matt Gardner</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>Bend but Don’t Break? Multi-Challenge Stress Test for QA Models</td>
<td>Hemant Pugaliya, James Route, Kaixin Ma, Yixuan Geng and Eric Nyberg</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>ReQA: An Evaluation for End-to-End Answer Retrieval Models</td>
<td>Amin Ahmad, Noah Constant, Yinfei Yang and Daniel Cer</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>Comprehensive Multi-Dataset Evaluation of Reading Comprehension</td>
<td>Dheeru Dua, Ananth Gottumukkala, Alon Talmor, Matt Gardner and Sameer Singh</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>A Recurrent BERT-based Model for Question Generation</td>
<td>Ying-Hong Chan and Yao-Chung Fan</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>Extractive NarrativeQA with Heuristic Pre-Training</td>
<td>Lea Frermann</td>
</tr>
<tr>
<td>14:55–16:30</td>
<td>CLER: Cross-task Learning with Expert Representation to Generalize Reading and Understanding</td>
<td>Takumi Takahashi, Motoki Taniguchi, Tomoki Taniguchi and Tomoko Ohkuma</td>
</tr>
</tbody>
</table>
Monday, November 4, 2019 (continued)

14:55–16:30  Question Answering Using Hierarchical Attention on Top of BERT Features
Reham Osama, Nagwa El-Makky and Marwan Torki

14:55–16:30  Domain-agnostic Question-Answering with Adversarial Training
Seanie Lee, Donggyu Kim and Jangwon Park

14:55–16:30  Generalizing Question Answering System with Pre-trained Language Model Fine-tuning
Dan Su, Yan Xu, Genta Indra Winata, Peng Xu, Hyeondey Kim, Zihan Liu and Pascale Fung

14:55–16:30  D-NET: A Pre-Training and Fine-Tuning Framework for Improving the Generalization of Machine Reading Comprehension
Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang, Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu and Haifeng Wang

14:55–16:30  An Exploration of Data Augmentation and Sampling Techniques for Domain-Agnostic Question Answering
Shayne Longpre, Yi Lu, Zhucheng Tu and Chris DuBois

16:30–17:30  Panel discussion