Open-Domain Question Answering

Danqi Chen
Princeton University
Princeton, NJ
danqic@cs.princeton.edu

Wen-tau Yih
Facebook AI Research
Seattle, WA
scottyih@fb.com

1 Description

Open-domain question answering (QA), the task of answering questions using a large collection of documents of diversified topics, has been a long-standing problem in NLP, information retrieval (IR) and related fields (Voorhees et al., 1999; Moldovan et al., 2000; Brill et al., 2002; Ferrucci et al., 2010). Traditional QA systems were usually constructed as a pipeline, consisting of many different components such as question processing, document/passage retrieval, and answer processing. With the rapid development of neural reading comprehension (Chen, 2018), modern open-domain QA systems have been restructured by combining traditional IR techniques and neural reading comprehension models (Chen et al., 2017; Yang et al., 2019; Min et al., 2019a) or even implemented in a fully end-to-end fashion (Lee et al., 2019; Seo et al., 2019; Guu et al., 2020; Roberts et al., 2020). In this tutorial, we aim to provide a comprehensive and coherent overview of *cutting-edge* research in this direction.[1]

We will start by first giving a brief background of open-domain question answering, discussing the basic setup and core technical challenges of the research problem. We aim to give the audience a historical view of how the field has advanced in the past several decades, from highly-modulated pipeline systems in the early days, to modern end-to-end training of deep neural networks in the present.

We will then discuss modern datasets proposed for open-domain QA (Voorhees et al., 1999; Berant et al., 2013; Rajpurkar et al., 2016; Joshi et al., 2017; Dhingra et al., 2017; Dunn et al., 2017; Kwiatkowski et al., 2019), as well as common evaluation metrics and benchmarks. We plan to provide a detailed discussion on available datasets — their collection methodology and properties — as well as insights on how these datasets should be viewed in the context of open-domain QA.

Next, the focus will shift to cutting-edge models proposed for open-domain QA, which is also the central part of this tutorial. We divide existing models into three main categories: *Two-stage retriever-reader approaches*, *Dense retriever and end-to-end training*, and *Retriever-free approaches*. We will present the logical elements behind different sorts of models and discuss their pros and cons.

Two-stage retriever-reader approaches. We will start by discussing two-stage retriever-reader frameworks for open-domain QA, pioneered by Chen et al. (2017): a *retriever* component finding documents that (might) contain an answer from a large collection of documents, followed by a *reader* component finding the answer in a given paragraph or a document. In this category, the retriever component is usually implemented by traditional sparse vector space methods, such as TF-IDF or BM25 and the reader is implemented by neural reading comprehension models. We will further discuss several challenges and techniques arising in this area, including multi-passage training (Clark and Gardner, 2018; Wang et al., 2019), passage reranking (Wang et al., 2018; Nogueira and Cho, 2019), and denoising distantly-supervised data (Lin et al., 2018).

Dense retriever and end-to-end training. The first category mainly employs a non-machine learning model for the retrieval stage. The second category will focus on how to *learn* the retriever component by replacing traditional IR methods with dense representations, as well as joint training of both components. Learning and searching in dense vector space is challenging, as it usually involves

[1] All the tutorial materials will be released at https://github.com/danqi/acl2020-openqa-tutorial.

an enormous search space (easily ranging from millions to billions of documents). We will discuss in depth how this was achieved by existing models, including novel pre-training methods (Lee et al., 2019; Guu et al., 2020), carefully-designed learning algorithms (Karpukhin et al., 2020) or a hybrid approach using both dense and sparse representations (Seo et al., 2019).

Retriever-free approaches. The third category, which is a recent emerging trend, only relies on large-scale pre-trained models (Radford et al., 2018; Devlin et al., 2018; Liu et al., 2019) as implicit knowledge bases and doesn't require access to text data during inference time. These pre-trained models will be used directly to answer questions, in a zero-shot manner (Radford et al., 2019; Raffel et al., 2019) or fine-tuned using question-answer pairs (Roberts et al., 2020). As these methods don't need a retriever component, we call them *Retriever-free approaches*.

Up to this point, our tutorial has mainly focused on textual question answering. At the end, we also plan to discuss some hybrid approaches for answering open-domain questions using both text and large knowledge bases, such as Freebase (Bollacker et al., 2008) and Wikidata (Vrandečić and Krötzsch, 2014), and give a critical review on how structured data complements the information from unstructured text. The approaches include (1) how to leverage structured data to guide the retriever or reader stage of existing textual QA systems (Asai et al., 2020; Min et al., 2019b), or (2) how to synthesize information from these two heterogeneous sources and build effective QA models on the combined information (Sun et al., 2018, 2019; Xiong et al., 2019).

Finally, we will discuss some important questions, including (1) How much progress have we made compared to the QA systems developed in the last decade? (2) What are the main challenges and limitations of current approaches? (3) How to trade off the efficiency (computational time and memory requirements) and accuracy in the deep learning era? We hope our tutorial will not only serve as a useful resource for the audience to efficiently acquire up-to-date knowledge, but also provide new perspectives to stimulate the advances of open-domain QA research in the next phase.

Prerequisites The tutorial will be accessible to anyone who has the basic knowledge of machine learning and natural language processing. The tutorial will target both NLP researchers/students in academia and NLP practitioners in industry.

2 Tutorial Outline

The intended duration of this tutorial is 3.5 hours, including a half an hour break.

1. Introduction
2. Problem definition & motivation
3. A history of open-domain (textual) QA
   (a) Early QA systems
   (b) TREC QA competitions
   (c) IBM's DeepQA project
   (d) More recent developments: 2017-2020
4. Datasets & evaluation
   (a) Reading comprehension vs QA datasets
   (b) Categorization of QA datasets
   (c) Evaluation metrics
5. Two-stage retriever-reader approaches
   (a) General framework
   (b) Multi-passage training
   (c) Passage reranking
   (d) Denoising distantly supervised data
6. Dense retriever and end-to-end training
   (a) Dense passage retrieval
   (b) Joint training of retriever and reader
   (c) Dense-sparse phrase indexing
7. Retriever-free approaches
8. Open-domain QA using KBs and text
   (a) Improving retriever and reader using structured KBs
   (b) Answering questions over combined KBs and text
9. Open problems and future directions

3 Presenters

Danqi Chen Danqi Chen is an Assistant Professor of Computer Science at Princeton University and co-directs the Princeton NLP Group. Danqi's research interests lie within deep learning for natural language processing, with an emphasis on the intersection between text understanding and knowledge representation/reasoning and applications

such as question answering and information extraction. Before joining Princeton University, Danqi worked as a visiting scientist at Facebook AI Research (FAIR). She received her PhD from Stanford University (advised by Christopher Manning) in 2018 and B.Eng from Tsinghua University in 2012. Website: https://www.cs.princeton.edu/∼danqic/.

Scott Wen-tau Yih Scott Wen-tau Yih is a Research Scientist at Facebook AI Research (FAIR), and his recent research focuses on continuous representations and neural network models, with applications in knowledge base embedding, semantic parsing and question answering. Yih received the best paper award from CoNLL'11, an outstanding paper award from ACL'15 and has served as an area co-chair and a program co-chair for several top conferences. He is also a co-presenter for several popular tutorials on topics including Semantic Role Labeling, Deep Learning for NLP, Question Answering with Knowledge Base, Web and Beyond and NLP for Precision Medicine. Website: http://scottyih.org/.

References

Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong. 2020. Learning to retrieve reasoning paths over wikipedia graph for question answering. In *International Conference on Learning Representations (ICLR)*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1533–1544.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *ACM SIGMOD international conference on Management of data*, pages 1247–1250.

Eric Brill, Susan Dumais, and Michele Banko. 2002. An analysis of the AskMSR question-answering system. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 257–264.

Danqi Chen. 2018. *Neural Reading Comprehension and Beyond*. Ph.D. thesis, Stanford University.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Association for Computational Linguistics (ACL)*, pages 1870–1879.

Christopher Clark and Matt Gardner. 2018. Simple and effective multi-paragraph reading comprehension. In *Association for Computational Linguistics (ACL)*, pages 845–855.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Bhuwan Dhingra, Kathryn Mazaitis, and William W Cohen. 2017. Quasar: Datasets for question answering by search and reading. *arXiv preprint arXiv:1707.03904*.

Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. SearchQA: A new Q&A dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179*.

David Ferrucci, Eric Brown, Jennifer Chu-Carroll, James Fan, David Gondek, Aditya A Kalyanpur, Adam Lally, J William Murdock, Eric Nyberg, John Prager, et al. 2010. Building Watson: An overview of the DeepQA project. *AI magazine*, 31(3):59–79.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. *arXiv preprint arXiv:2002.08909*.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Association for Computational Linguistics (ACL)*, pages 1601–1611.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Change, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics (TACL)*.

Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. 2019. Latent retrieval for weakly supervised open domain question answering. In *Association for Computational Linguistics (ACL)*, pages 6086–6096.

Yankai Lin, Haozhe Ji, Zhiyuan Liu, and Maosong Sun. 2018. Denoising distantly supervised open-domain question answering. In *Association for Computational Linguistics (ACL)*, pages 1736–1745.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Sewon Min, Danqi Chen, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2019a. A discrete hard EM approach for weakly supervised question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2019b. Knowledge guided text retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*.

Dan Moldovan, Sanda Harabagiu, Marius Pasca, Rada Mihalcea, Roxana Girju, Richard Goodrum, and Vasile Rus. 2000. The structure and performance of an open-domain question answering system. In *Association for Computational Linguistics (ACL)*, pages 563–570.

Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage re-ranking with BERT. *arXiv preprint arXiv:1901.04085*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language models are unsupervised multitask learners. Technical report, OpenAI.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? *arXiv preprint arXiv:2002.08910*.

Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. In *Association for Computational Linguistics (ACL)*, pages 4430–4441.

Haitian Sun, Tania Bedrax-Weiss, and William W Cohen. 2019. PullNet: Open domain question answering with iterative retrieval on knowledge bases and text. In *Empirical Methods in Natural Language Processing (EMNLP)*.

Haitian Sun, Bhuwan Dhingra, Manzil Zaheer, Kathryn Mazaitis, Ruslan Salakhutdinov, and William Cohen. 2018. Open domain question answering using early fusion of knowledge bases and text. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 4231–4242.

Ellen M Voorhees et al. 1999. The TREC-8 question answering track report. In *Trec*, volume 99, pages 77–82. Citeseer.

Denny Vrandečić and Markus Krötzsch. 2014. Wikidata: a free collaborative knowledge base.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerald Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced reader-ranker for open-domain question answering. In *Conference on Artificial Intelligence (AAAI)*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A globally normalized BERT model for open-domain question answering. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 5881–5885.

Wenhan Xiong, Mo Yu, Shiyu Chang, Xiaoxiao Guo, and William Yang Wang. 2019. Improving question answering over incomplete kbs with knowledge-aware reader. In *Association for Computational Linguistics (ACL)*, pages 4258–4264.

Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin. 2019. End-to-end open-domain question answering with BERTserini. In *North American Association for Computational Linguistics (NAACL)*, pages 72–77.