# Multi-Domain Multilingual Question Answering

**Sebastian Ruder**
DeepMind
ruder@google.com

**Avirup Sil**
IBM Research AI
avi@us.ibm.com

## Abstract

Question answering (QA) is one of the most challenging and impactful tasks in natural language processing. Most research in QA, however, has focused on the open-domain or monolingual setting while most real-world applications deal with specific domains or languages. In this tutorial, we attempt to bridge this gap. Firstly, we introduce standard benchmarks in multi-domain and multilingual QA. In both scenarios, we discuss state-of-the-art approaches that achieve impressive performance, ranging from zero-shot transfer learning to out-of-the-box training with open-domain QA systems. Finally, we will present open research problems that this new research agenda poses such as multi-task learning, cross-lingual transfer learning, domain adaptation and training large scale pre-trained multilingual language models.[1]

## 1 Overall

Question answering (QA) has emerged as one of the most popular areas in natural language processing (NLP). Established benchmarks such as the Stanford Question Answering Dataset (SQuAD; Rajpurkar et al., 2016) are used as a standard testing ground for new models while open-domain QA benchmarks such as Natural Questions (Kwiatkowski et al., 2019) represent the frontier of what is possible with current NLP technology (Zaheer et al., 2020).

In this tutorial, we will review recent advances in open-domain QA but focus on an area that has received less attention both in research and in past tutorials—multi-domain and multilingual QA. Open-domain QA is of interest for building general-purpose assistants that can answer questions about any topic (Adiwardana et al., 2019). Most real-world applications of QA, however, deal with the needs of specific domains. Multi-domain QA is particularly promising as it allows us to adapt models to new domains that are of practical importance, such as answering questions about COVID-19 (Tang et al., 2020).

At the same time, over the course of the last year we have seen the emergence of the first benchmarks for multilingual QA (Lewis et al., 2020; Artetxe et al., 2020; Clark et al., 2020). These benchmarks are a step towards enabling access to technology beyond English and building question answering systems that serve all of the world's approximately 6,900 languages. In addition to introducing standard datasets for multilingual QA, we will discuss advances in cross-lingual learning that made such benchmarks viable for the first time.

We generally aim to highlight methods and techniques that can be applied to adapt to many domains and languages in order to be helpful to the majority of the audience. While multi-domain and multilingual data differ in many ways both can be formulated as transfer learning problems and approached using a similar set of fundamental tools and principles, which we aim to convey to our audience.

As one example of such a tool, we will cover training procedures for large pre-trained language models (LMs). For multi-domain QA, we will discuss adaptation of LMs *e.g.* BERT (Devlin et al., 2019) or RoBERTa (Liu et al., 2019). For multilingual QA, we will teach the methods for training LMs from large multilingual supervised and unsupervised data *e.g.* XLM-RoBERTa (Conneau et al., 2019) and M4 (Arivazhagan et al., 2019). Notably, our tutorial will highlight the challenges of applying such methods to specific domains and languages. Overall, we will aim to provide a set of best practices that will enable researchers and practitioners to train methods for their domain and

[1]The tutorial materials are available at https://github.com/sebastianruder/emnlp2021-multiqa-tutorial.



*Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts*, pages 17–21
Punta Cana and Online, November 10–11, 2021. 

language of interest, from the nature of the training data, to model architectures and hyper-parameter settings.

**Type of the tutorial:** Cutting-edge.

**Prior QA tutorials at ACL:** The broader area of question answering has been a staple of tutorials at NLP conferences *e.g.* ACL 2018, ACL 2020. In general, we will demonstrate that techniques from open-domain QA cannot be directly applied to perform QA on unseen new domains (Tang et al., 2020; Castelli et al., 2020) and emphasize the importance of domain-specific training is necessary. This is the first tutorial to focus specifically on multi-domain and multilingual question answering, which has not been taught anywhere before.

**Breadth:** The tutorial will cover 90% of work from the QA, machine reading comprehension, domain adaptation and multilingual literature and 10% of the presenters work.

**Diversity:** The tutorial will cover multilingual work including discussions of large multilingual pre-trained language models and QA examples in different languages. We will also discuss how methods scale to different languages and domains, including how much training data is necessary to achieve a certain performance.

**Prerequisites:** Familiarity with Transformer models and pre-trained language models such as BERT.

## 2 Brief Tutorial Outline

This is a 3 hour tutorial: hence, we will divide our time between the following novel topics:

### 2.1 First half: Multi-Domain QA

1. **Open-Domain monolingual QA and its limitations [20 mins]:** We will begin our tutorial by introducing our audience to the existing work on open-domain QA (also known as reading comprehension) and its recent progress on benchmark tasks such as SQuAD (Rajpurkar et al., 2016, 2018) and Natural Questions (Kwiatkowski et al., 2019). We will then survey work on monolingual QA: giving a brief historical background, discussing the basic setup and core technical challenges of the research problem, and then describe modern datasets with the common evaluation metrics and benchmarks. Finally, we will discuss their limitations when applied to unseen closed domains e.g. movies, information technology (IT) or biomedical questions and motivate the next section.

2. **Introduce Multi-domain QA [20 mins]:** We will focus on several recent benchmark datasets *e.g.* TechQA (Castelli et al., 2020) and DoQA (Campos et al., 2020), which introduce more realistic QA scenarios. The former introduces a dataset and a leaderboard for IT that comes with only a limited amount of training data. The latter requires strong domain adaptation as QA systems are trained on the "cooking" domain and tested by answering questions about movies and travel. DoQA is rather challenging as QA systems need to take narrative context into consideration, which most reading comprehension systems do not. We will furthermore discuss recent datasets such as CovidQA (Tang et al., 2020), which focus on emerging domains that are of practical importance.

3. **Modeling and Evaluation [30 mins]:** Finally, we will focus on various initial baselines which can be adopted to achieve impressive results via transfer learning on top of large pre-trained language models such as BERT (Devlin et al., 2019). We will also discuss the evaluation methodology including the various metrics that measure document retrieval and QA performance. Finally, we give an overview of many practical ways to adapt to another domain such as via in-domain pre-training and *task-adaptive pretraining*, which improves performance by adapting to a task's unlabeled data (Gururangan et al., 2020).

### 2.2 Coffee Break: [30 mins]

### 2.3 Hour 2: Multilingual QA and open research problems

1. **From Mono to large Multilingual Language Models [15 mins]:** In this half we will first survey some of the large multilingual language models *e.g.* mBERT (Devlin et al., 2019), XLM (Conneau and Lample, 2019), XLM-R (Conneau et al., 2019), M4 (Arivazhagan et al., 2019). We will show how they have helped close the gap on cross-lingual tasks by introducing zero-shot cross-lingual learning.

2. **Multilingual QA [40 mins]:** Then we will give a comprehensive overview of several

non-English multilingual question answering datasets and systems such as DuReader (He et al., 2018) and DRCD (Shao et al., 2018) in Chinese, ARCD (Mozannar et al., 2019) in Arabic, multi-domain QA (Gupta et al., 2018) in Hindi-English, and visual QA (Gao et al., 2016) in Chinese-English. We distinguish between datasets that have been created by obtaining naturally occurring data in a language or via translations from SQuAD into Korean (Lee et al., 2018; Li et al., 2018), French and Japanese (Asai et al., 2018) and Italian (Croce et al., 2019). Recent datasets such as XQuAD (Artetxe et al., 2020) and MLQA (Lewis et al., 2020) cover more languages while the recently introduced TyDiQA (Clark et al., 2020) and MKQA (Longpre et al., 2020) can be seen as multilingual counterparts to Natural Questions. Three of these datasets are part of XTREME (Hu et al., 2020), a massively multilingual benchmark for testing the cross-lingual generalization ability of state-of-the-art methods. While state-of-the-art models have matched or surpassed human performance in general-purpose monolingual benchmarks such as GLUE (Wang et al., 2019), current methods still fall short of human performance on multilingual benchmarks, despite recent gains (Chi et al., 2020). Multilingual question answering consequently is at the frontier of such cross-lingual generalization. We will generally aim to highlight the settings where current methods fail, showing validation examples in different languages, and highlight best practices of how methods can be adapted to better deal with them.

3. **Open research problems [25 mins]:** Finally, we will discuss challenges and promising research directions for multi-domain and multilingual question answering.

## 3 Goals

### 3.1 What are the objectives of the tutorial?

Firstly, to familiarize the audience with the task of monolingual question answering and latest benchmarks on open-domain QA. We furthermore aim to raise awareness of the challenges of QA across multiple domains and languages, to demonstrate the usefulness of adapting models to such settings, and to teach best practices for different adaptation scenarios.

### 3.2 Why is this tutorial important to include at ACL?

Multi-domain and multilingual question answering is a key technology to deal with emerging topics and challenges around the world such as COVID-19 (Tang et al., 2020). We expect that being familiar and having access to the toolkit of multi-domain multilingual QA will both enable researchers to make progress on fundamental challenges and allow practitioners to leverage research advances in real-world applications. In addition, highlighting challenges and introducing the audience to techniques for adapting QA models to other languages may contribute to a broader, less English-centric research landscape.

## 4 Presenters

- **Name:** Sebastian Ruder
  **Affiliation:** DeepMind
  **Email:** sebastian@ruder.io
  **Website:** http://ruder.io
  Sebastian is a research scientist at DeepMind where he works on transfer learning and multilingual natural language processing. He has been area chair in machine learning and multilinguality for major NLP conferences including ACL and EMNLP and has published papers on multilingual question answering (Artetxe et al., 2020; Hu et al., 2020). He was the Co-Program Chair for EurNLP 2019 and has co-organized the 4th Workshop on Representation Learning for NLP at ACL 2019. He has taught tutorials on "Transfer learning in natural language processing" and "Unsupervised Cross-lingual Representation Learning" at NAACL 2019 and ACL 2019 respectively. He has also co-organized and taught at the NLP Session at the Deep Learning Indaba 2018 and 2019.

  **Section:** Sebastian will teach Multilingual QA during this tutorial (Second 1 1/2 hrs).

- **Name:** Avirup Sil
  **Affiliation:** IBM Research AI
  **Email:** avi@us.ibm.com
  **Website:** http://ibm.biz/avirupsil
  Avi is a Research Scientist and the Team Lead for Question Answering in the Multilingual NLP group at IBM Research AI. His team (comprising of research scientists and engineers) works on research on indus-

try scale NLP and Deep Learning algorithms. His team's system called 'GAAMA' has obtained the top scores in public benchmark datasets (Kwiatkowski et al., 2019) and has published several papers on question answering (Chakravarti et al., 2019; Castelli et al., 2020; Glass et al., 2020). He is also the Chair of the NLP professional community of IBM. Avi is a Senior Program Committe Member and the Area Chair in Question Answering for major NLP conferences *e.g.* ACL, EMNLP, NAACL and has published several papers on Question Answering. He has taught a tutorial at ACL 2018 on "Entity Discovery and Linking". He has also organized the workshop on the "Relevance of Linguistic Structure in Neural NLP" at ACL 2018. He is also the track coordinator for the Entity Discovery and Linking track at the Text Analysis Conference.

**Section:** Avi will teach Multi-domain QA during this tutorial (First 1 1/2 hrs).

## References


Daniel Adiwardana, Minh-thang Luong David, R So Jamie, Gaurav Nemade, Yifeng Lu, and Quoc V Le. 2019. Towards a Human-like Open-Domain Chatbot.

Naveen Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George Foster, Colin Cherry, et al. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *arXiv preprint arXiv:1907.05019*.

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2020. On the Cross-lingual Transferability of Monolingual Representations. In *Proceedings of ACL 2020*.

Akari Asai, Akiko Eriguchi, Kazuma Hashimoto, and Yoshimasa Tsuruoka. 2018. Multilingual extractive reading comprehension by runtime machine translation. *arXiv preprint arXiv:1809.03275*.

Jon Ander Campos, Arantxa Otegi, Aitor Soroa, Jan Deriu, Mark Cieliebak, and Eneko Agirre. 2020. DoQA–Accessing domain-specific FAQs via conversational QA. *ACL*.

Vittorio Castelli, Rishav Chakravarti, Saswati Dana, Anthony Ferritto, Radu Florian, Martin Franz, Dinesh Garg, Dinesh Khandelwal, Scott McCarley, Mike McCawley, et al. 2020. The TechQA Dataset. *Association for Computational Linguistics (ACL)*.

Rishav Chakravarti, Cezar Pendus, Andrzej Sakrajda, Anthony Ferritto, Lin Pan, Michael Glass, Vittorio Castelli, J William Murdock, Radu Florian, Salim Roukos, and Avirup Sil. 2019. CFO: A framework for building production nlp systems. *EMNLP-IJCNLP, Demo Track*.

Zewen Chi, Li Dong, Furu Wei, Nan Yang, Saksham Singhal, Wenhui Wang, Xia Song, Xian-Ling Mao, Heyan Huang, and Ming Zhou. 2020. InfoXLM: An Information-Theoretic Framework for Cross-Lingual Language Model Pre-Training. *arXiv preprint arXiv:2007.07834*.

Jonathan H. Clark, Eunsol Choi, Michael Collins, Dan Garrette, Tom Kwiatkowski, Vitaly Nikolaev, and Jennimaria Palomaki. 2020. TYDI QA: A Benchmark for Information-Seeking Question Answering in Typologically Diverse Languages. *Transactions of the ACL 2020*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau and Guillaume Lample. 2019. Cross-lingual language model pretraining. In *Advances in Neural Information Processing Systems*, pages 7057–7067.

Danilo Croce, Alexandra Zelenanska, and Roberto Basili. 2019. Enabling deep learning for large scale question answering in italian. *Intelligenza Artificiale*, 13(1):49–61.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2016. Multilingual image question answering. US Patent App. 15/137,179.

Michael Glass, Alfio Gliozzo, Rishav Chakravarti, Anthony Ferritto, Lin Pan, Bhargav GP Shrivatsa, Dinesh Garg, and Avirup Sil. 2020. Span selection pre-training for question answering. *Association for Computational Linguistics (ACL)*.

Deepak Gupta, Surabhi Kumari, Asif Ekbal, and Pushpak Bhattacharyya. 2018. Mmqa: A multi-domain multi-lingual question-answering framework for english and hindi. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. *ACL*.

Wei He, Kai Liu, Jing Liu, Yajuan Lyu, Shiqi Zhao, Xinyan Xiao, Yuan Liu, Yizhong Wang, Hua Wu, Qiaoqiao She, Xuan Liu, Tian Wu, and Haifeng Wang. 2018. DuReader: a Chinese machine reading comprehension dataset from real-world applications. In *Proceedings of the Workshop on Machine Reading for Question Answering*, pages 37–46, Melbourne, Australia. Association for Computational Linguistics.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. *arXiv preprint arXiv:2003.11080*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a benchmark for question answering research. *TACL*.

Kyungjae Lee, Kyoungho Yoon, Sunghyun Park, and Seung-won Hwang. 2018. Semi-supervised training data generation for multilingual question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2020. MLQA: Evaluating cross-lingual extractive question answering. *ACL*.

Jian Li, Zhaopeng Tu, Baosong Yang, Michael R Lyu, and Tong Zhang. 2018. Multi-head attention with disagreement regularization. In *EMNLP*, pages 2897–2903.

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. *CoRR*, abs/1907.11692.

Shayne Longpre, Yi Lu, and Joachim Daiber. 2020. MKQA: A Linguistically Diverse Benchmark for Multilingual Open Domain Question Answering. *arXiv preprint arXiv:2007.15207*.

Hussein Mozannar, Elie Maamary, Karl El Hajal, and Hazem Hajj. 2019. Neural Arabic question answering. In *Proceedings of the Fourth Arabic Natural Language Processing Workshop*, pages 108–118, Florence, Italy. Association for Computational Linguistics.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ questions for machine comprehension of text. *EMNLP*.

Chih Chieh Shao, Trois Liu, Yuting Lai, Yiying Tseng, and Sam Tsai. 2018. Drcd: a chinese machine reading comprehension dataset. *arXiv preprint arXiv:1806.00920*.

Raphael Tang, Rodrigo Nogueira, Edwin Zhang, Nikhil Gupta, Phuong Cam, Kyunghyun Cho, and Jimmy Lin. 2020. Rapidly Bootstrapping a Question Answering Dataset for COVID-19.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R Bowman. 2019. GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *Proceedings of ICLR 2019*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big Bird: Transformers for Longer Sequences. *arXiv preprint arxiv:2007.14062*, pages 1–51.