# MIA 2022 Shared Task: Instructions for System Descriptions

**MIA Organizers**
Multilingual Information Access
mia.nlp.workshop@gmail.com

## Abstract

This document present instructions for participants of the MIA 2022 shared task on multilingual open question answering. The document itself conforms to its own specifications, and is therefore an example of what your manuscript should look like. The style generally follows the ACL proceedings template.

## 1 Introduction

These instructions are for authors submitting papers for system descriptions. They are not self-contained. All authors must follow the general instructions for *ACL proceedings,[1] and this document contains additional instructions for the LaTeX style files.

The templates include the LaTeX source of this document (`acl.tex`), the LaTeX style file used to format it (`acl.sty`), an ACL bibliography style (`acl_natbib.bst`), an example bibliography (`custom.bib`), and the bibliography for the ACL Anthology (`anthology.bib`).

## 2 Recommended Content

For a system description, we recommend that authors provide information that will help other participants and the organizers understand and compare systems. Such information will also help subsequent work to replicate results and make comparisons on multilingual open question answering (QA). We also welcome discussions aligned with the goals of the MIT shared task: **what are the challenges should researchers and practitioners be aware of when we apply our (English-centric) open QA systems to practical, multilingual settings?**

Specifically, we suggest that authors include (some of) the following content:

- Data and Preprocessing
- Model Architecture and Pipeline
- Model (Pre)training and Finetuning
- Model Inference
- Development Performance and Baselines
- Memory Overhead and Inference Speed
- Language-Specific Details or Challenges (if any)

**Data and Preprocessing** What data did you use to train/finetune your model? Did you use QA datasets for finetuning? Provide citations for datasets that you used if any. Are all data publicly available? Did you apply any filtering to your data? Did you use tokenization such as BPE (Sennrich et al., 2016) or sentencepiece (Kudo and Richardson, 2018)? Did you apply language identification to select data for the target languages?

**Model Architecture and Pipeline** What architecture did you use for retrieval and answer generation? Did you develop a transformer-based model like the dense passage retriever (Karpukhin et al., 2020)? Did you base your system on a widely-used codebase, such as the HuggingFace's Transformers (Wolf et al., 2020) or fairseq (Ott et al., 2019) libraries? Did you have machine translation systems in the pipeline to deal with multilingual questions or answer generation? If so, what machine translation models did you use?

**Model (Pre)training and Finetuning** Did you use any off-the-shelf pretrained model, such as multilingual BERT (Devlin et al., 2019) or multilingual T5 (Xue et al., 2021)? What kind of finetuning did you apply on top of these off-the-shelf models? How did you train your retrieval or generation model? Did you apply *in-batch negative*

[1] http://acl-org.github.io/ACLPUB/formatting.html.

to train your retrieval model (Gillick et al., 2019; Karpukhin et al., 2020)? What hyperparameters did you use for your training?

**Model Inference** How does inference work on your system? Does your system run left-to-right autregressive answer generation or answer extraction from evidence documents?

**Development Performance and Baselines** It is useful to report development performances of your model. How does it compare to multilingual open QA models from prior work (e.g., Asai et al., 2021a,b)? Did you use these performance scores for your design choices?

**Memory Overhead and Inference Speed** How large is the memory overhead of your model? How many questions can your system process per second on what hardware? While such information is often not a focus of many open QA papers, it might be worth reporting because the shared task is highly motivated by practical applications. Did you apply any technique to scale your model to **many languages**?

**Language-Specific Details or Challenges** If you applied any processing (e.g., tokenization, data crawling, data cleaning, and language identification) that is specific to some languages, it is also worth reporting here. This is an important point of discussion that helps us understand problems of the current English-centric systems.

### 2.1 Useful Resources

You can find some useful references for system descriptions from past shared tasks. For example,

- Annual news translation task from WMT (Barrault et al., 2020; Akhbardeh et al., 2021). You can find system descriptions from past participants (e.g., Ng et al., 2019; Marchisio et al.).
- Shared task from the workshop on machine reading for question answering (MRQA; Fisch et al.). Several papers are available for system descriptions (e.g., Longpre et al., 2019; Li et al., 2019).

## References


Farhad Akhbardeh, Arkady Arkhangorodsky, Magdalena Biesialska, Ondřej Bojar, Rajen Chatterjee, Vishrav Chaudhary, Marta R. Costa-jussa, Cristina España-Bonet, Angela Fan, Christian Federmann, Markus Freitag, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Leonie Harter, Kenneth Heafield, Christopher Homan, Matthias Huck, Kwabena Amponsah-Kaakyire, Jungo Kasai, Daniel Khashabi, Kevin Knight, Tom Kocmi, Philipp Koehn, Nicholas Lourie, Christof Monz, Makoto Morishita, Masaaki Nagata, Ajay Nagesh, Toshiaki Nakazawa, Matteo Negri, Santanu Pal, Allahsera Auguste Tapo, Marco Turchi, Valentin Vydrin, and Marcos Zampieri. 2021. Findings of the 2021 conference on machine translation (WMT21). In *Proc. of WMT*.

Akari Asai, Jungo Kasai, Jonathan Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2021a. XOR QA: Cross-lingual open-retrieval question answering. In *Proc. of NAACL*.

Akari Asai, Xinyan Yu, Jungo Kasai, and Hannaneh Hajishirzi. 2021b. One question answering model for many languages with cross-lingual dense passage retrieval. In *Proc. of NeurIPS*.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proc. of WMT*.

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

Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. MRQA 2019 shared task: Evaluating generalization in reading comprehension. In *Proc. of MRQA*.

Daniel Gillick, Sayali Kulkarni, Larry Lansing, Alessandro Presta, Jason Baldridge, Eugene Ie, and Diego Garcia-Olano. 2019. Learning dense representations for entity retrieval. In *Proc. of CoNLL*.

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. In *Proc. of EMNLP*.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proc. of EMNLP*.

Hongyu Li, Xiyuan Zhang, Yibing Liu, Yiming Zhang, Quan Wang, Xiangyang Zhou, Jing Liu, Hua Wu, and Haifeng Wang. 2019. D-NET: A pre-training and fine-tuning framework for improving the generalization of machine reading comprehension. In *Proc. of MRQA*.

Shayne Longpre, Yi Lu, Zhucheng Tu, and Chris DuBois. 2019. An exploration of data augmentation and sampling techniques for domain-agnostic question answering. In *Proc. of MRQA*.

Kelly Marchisio, Yash Kumar Lal, and Philipp Koehn. Johns Hopkins University submission for WMT news translation task. In *Proc. of WMT*.

Nathan Ng, Kyra Yee, Alexei Baevski, Myle Ott, Michael Auli, and Sergey Edunov. 2019. Facebook FAIR's WMT19 news translation task submission. In *Proc. of WMT*.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proc. of NAACL: Demonstrations*.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proc. of ACL*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proc. of EMNLP: System Demonstrations*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proc. of NAACL*.

## A Example Appendix

This is an appendix.