PANEL DISCUSSION ON MRQA 2021 MULTILINGUAL TRACK

Yiming Cui
Research Center for SCIR, Harbin Institute of Technology, Harbin, China
Joint Laboratory of HIT and iFLYTEK Research (HFL), Beijing, China
ymcui@ir.hit.edu.cn

ABSTRACT

Due to the limited time of panel discussion in the Third Workshop on Machine Reading for Question Answering (MRQA 2021) workshop[1], there are many interesting topics left undiscussed. This report presents several personal opinions behind these discussion topics, mainly focusing on multilingual question answering.

1 EVALUATION

Discussion 1-1: Our evaluation metrics (and often, the test sets) have mostly been developed with English-based question answering in mind. (1) Are any of these metrics particularly not transferable across languages, (2) Is it even appropriate to be able to expect to quantitatively compare QA performance across languages, and (3) if so, how should we go about it?

For example, the Exact Match (EM) score in SQuAD (Rajpurkar et al., 2016) is quite universal for different languages, as it is a string match between the ground truth and predicted answer. But for the F1 score, it is not directly applicable for some languages. For example, our CMRC 2018 (Cui et al., 2019b) dataset does not directly adopt the original F1-score metrics in SQuAD. Because Chinese text is usually tokenized as characters in BERT, but some of you may know that the Chinese character sometimes does not have an actual meaning, so it must be combined with other characters to form a word. And this is not the case for English. In our dataset, we use the longest common string (LCS) to detect the maximum overlap to calculate F1-score.

For the second question, I think it is not quite possible, at least in the current stage. The performance is subjected to the type of domain, and language-specific characteristics, etc. Considering SQuAD and CMRC 2018 has the same type of domain (they are all based on Wikipedia), I will compare them here. For example, there are many systems that achieved over-human performance on SQuAD with a relatively weak pre-trained language model. But for Chinese CMRC 2018, while we have used a much stronger pre-trained language model (such as MacBERT (Cui et al., 2021)), the top-system has not surpassed human performance yet.

Discussion 1-2: (Related to the previous question) Cultural factors matter for whether an answer is acceptable. Answering with a short span-like answer may be OK in English but might be strange, incomplete, or even considered rude in other languages. How should we balance exact task transfer (of models, dataset formats, evaluation, etc) and the cultural aspects in which languages are rooted?

Yes, this might be true for some languages (but I did not find such cases). I think for the general study of multilingual QA systems, it is unavoidable because we cannot understand all languages that we are investigating. In this case, I think this might be a good research topic for the researchers on those languages, to make a localized multilingual QA system that is suitable for relevant languages.

Discussion 1-3: What will it take to achieve QA 'parity' across languages, and is 'parity' even a realistic, appropriate, or well-dened goal? If it is, how should it be measured?

[1] https://mrqa.github.io

'Parity' across languages is somewhat not realistic, at least for now. Different languages have different backgrounds, even the datasets for different languages significantly differs, and this is the diversity of the languages.

Most of the multilingual QA systems are based on multilingual pre-trained language models, and its pre-training data is already unbalanced. Resource-rich languages are subject to better performance, while for resource-scarce languages are quite worse. Maybe we should take these factors into account when we design a new benchmark, for example, weighting the scarcity of the languages when calculating the evaluation metric.

Discussion 1-4: Cross-lingual, aggregated benchmarks are beginning to shape our directions on developing multilingual systemswill we select for the right things?

I think the main goal is OK. At least they are promoting multilingual systems for better accuracies. These research can be seen as establishing a better baseline for building multilingual systems. With these research, we can have an initial understanding on how our systems performs for different languages in a rough view. However, we also need some research on how to handle these multilingual systems to usable multilingual systems with additional efforts on developing language-specific and task-specific components. And these efforts require to involve linguistic knowledge.

2 TECHNIQUES

Thank you for your constructive comments on our work, and we are grateful for your positive comments. We are delighted to see that our previous manuscript has addressed your concerns.

Discussion 2-1: What are the major limiting factors for multilingual QA systems? Is it a lack of unsupervised data in certain languages (e.g., for pre-training)? Supervised data? Or are our models fundamentally ill-equipped for certain languages?

- **Pre-training:** A lack of long-tail languages that has few pre-training data available. In this case, it is not enough to learn basic embeddings for that language. Also, balancing ratios between different languages is still a hard question.
- **Fine-tuning:** At least, for now, multilingual QA systems are not quite different from other multilingual NLU systems. So the major problem may come from the pre-training stage. Also, language-specific treatment for different languages is very important, such as using different answer decoding techniques for morphologically different languages.

Discussion 2-2: A natural baseline for work on cross-lingual tasks is to translate the target language into a high-resource language (e.g., English), and use a model trained for this high-resource language. Obviously, this approach may be suboptimal due to errors in machine translation. To what extent is improving MT systems for use cases like these a fruitful research direction, as opposed to tackling the multilingual problem directly?

If the multilingual pre-trained language model does not have good support for that language, using a machine translation (MT) system might be a better solution. Another possible scenario is for the data in a specialized area with many professional words. In that case, a simple specialized SMT system may give a good translation for those terms. However, these terms might not be well-represented with a general multilingual PLM.

Discussion 2-3: Simply thinking about the linguistic challenges of multi-lingual QA can ignore some of the additional, practical challenges real users of those systems face. For example, speakers of under-represented languages might also live in areas for which expensive (in compute and memory) models cannot be feasible deployed. To what extent can (or should) modeling of multilingual QA be tightly coupled with work in efcient QA, or is it sufcient to tackle these challenges in parallel?

Creating an efficient QA system is quite important to those who have limited resources. Current multilingual QA systems are mostly relying on multilingual pre-trained language models with hundreds of million parameters or more.

- **Distillations**: We can use the knowledge distillation method to transfer knowledge from a big model to a small model. Last year, we released TextBrewer toolkit (Yang et al., 2020), which is a knowledge distillation tool for NLP, and it can be used for this purpose. But the drawback is that the distillation process still requires a lot of compute.
- **Model Pruning**: Multi-lingual QA system usually has a large vocabulary for covering many languages. In this context, we can simply discard some entries that we are not interested in. Currently, we are working on a new toolkit that supports model pruning in NLP. Stay tuned!

Discussion 2-4: Can cross-lingual transfer ever really replace in-language annotated training data?

For resource-scarce languages, it might be possible, especially for those who have no or few task-specific data available. For resource-rich languages, I personally dont think it is pretty promising.

Discussion 2-5: What can strong monolingual QA systems in resource-rich domains (like those in English) potentially gain from cross-lingual capabilities?

It might be helpful in a code-switch scenario. For example, there are several Chinese words mixed in an English MRC example that might prevent us from understanding the example properly. With cross-lingual capabilities, such as completely translating the example into Chinese or fusing the representation from a Chinese pre-trained model (Cui et al., 2019a), that might be helpful in QA performance.

Discussion 2-6: Transformers are fairly free of inductive bias and assumptions when it comes to input and output sequences. Tokenizers, not so much. What are your thoughts on how to make more robust tokenizers? Is a universal input processor even possible? What do we lose or gain?

Sentence-piece or word-piece tokenizers are already somewhat universal but not that accurate for many languages. Developing a universal input processor is promising but requires linguistic knowledge for different languages. Maybe we can train an end-to-end universal tokenizer by using raw text and language-specifically tokenized text for different languages. But I dont know if that could result in a better performance.

3 LANGUAGES

Discussion 3-1: What are languages that you would be excited to look at further (for some typological or cultural reasons)? How should data curators select languages to include in multilingual datasets?

For developing multilingual datasets for domestic research, I think it is interesting to incorporate dialectal and languages for different minorities. For example, in China, there are many different types of dialects and minorities. Many minorities have their own languages. So it is useful to incorporate these languages when building multilingual datasets to accelerate the research in these endangered languages. Recently, we have released the first multilingual pre-trained language model for Chinese minority languages, named CINO (Chinese mINOrity PLM). If you are interested, please check our GitHub repository[2].

For the second question, I think that it is up to the exact motivation of the created dataset. I think the multilingual dataset should maximize its diversity for general purposes. For example,

- Including different word orders, Chinese, English (SVO), Japanese (SOV), etc.
- Including different writing systems, such as left-to-right and right-to-left (Arabic), etc.
- Including different tokenization styles, English (word), Chinese (char), Korean, Arabic, etc.
- (Optional) Some long-tail languages that have few data available.

[2] https://github.com/ymcui/Chinese-Minority-PLM

Discussion 3-2: To make progress on creating multilingual systems, a researcher could either attempt to create a method that applies simultaneously across many languages, or they could focus on a particular language, with the hope that enough such researchers would work together on many languages in aggregate. What are the advantages or disadvantages or these approaches? Given that the former seems to be more in vogue right now, are there things we need to do to encourage the latter as well?

Multilingual systems enable us to understand many languages at scale. Most of the works are trying to develop a universal approach for many languages, which is a good starting point. For example, building multilingual pre-trained language models, multilingual NLU benchmarks, etc. In this case, we dont have to understand each language that we are investigating. But on the contrary, as we dont know these languages, whether it is really applicable is questionable.

Applying these systems in real-world applications still has a long way to go. We have to invite linguistics to further check that these systems are really applicable in a realistic scenario. In this context, I think it is quite hard to gather many researchers with different language backgrounds to achieve this goal. But maybe we can start by building a bilingual or multilingual system that contains linguistically/regionally similar languages, such as building a Chinese/Japanese/Korean (CJK) NLU system, etc.

Overall:

- Develop multilingual systems for many languages. This is a must because it is the fundamental part. And many researchers may be interested in this kind of research.
- Further fine-tuning multilingual systems to adjust specific language(s). This might be of interest to a particular research community.
- The third is to develop and perform fine-tuning on several similar languages. For example, building an Asian multilingual NLU system, where it can be much easier to gather the researcher in the same region and collaborate with each other.

4 Conclusion

In this report, I have presented personal opinions on the current and future prospects of multilingual question answering systems. We have seen the rapid progress of current multilingual (cross-lingual) QA benchmarks with the help of powerful multilingual pre-trained language models. However, there are also many points that should be discussed and investigated in future work, including reconsidering the appropriateness of current evaluation on multilingual QA, what makes the multilingual QA system different from general multilingual NLU, etc.

For workshop proceedings, please refer to Fisch et al. (2021).

Acknowledgment

I would like to thank Danqi Chen for inviting me as an invited speaker and panelist at MRQA 2021. I also thank the MRQA 2021 organization committee for their hard work in preparation for MRQA 2021 workshop.

Disclaimer

The idea presented in this report should not be regarded as the official opinion on behalf of HIT and iFLYTEK. The material should only be treated for archival purposes.

References

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, Shijin Wang, and Guoping Hu. Cross-lingual machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pp. 1586–1595, Hong Kong, China, November 2019a. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-1169.

Yiming Cui, Ting Liu, Wanxiang Che, Li Xiao, Zhipeng Chen, Wentao Ma, Shijin Wang, and Guoping Hu. A span-extraction dataset for Chinese machine reading comprehension. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp. 5883–5889, Hong Kong, China, November 2019b. Association for Computational Linguistics. doi: 10.18653/v1/D19-1600. URL https://aclanthology.org/D19-1600.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. Pre-training with whole word masking for chinese bert, 2021.

Adam Fisch, Alon Talmor, Danqi Chen, Eunsol Choi, Minjoon Seo, Patrick Lewis, Robin Jia, and Sewon Min (eds.). *Proceedings of the 3rd Workshop on Machine Reading for Question Answering*, Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.mrqa-1.0.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 2383–2392, Austin, Texas, November 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1264. URL https://aclanthology.org/D16-1264.

Ziqing Yang, Yiming Cui, Zhipeng Chen, Wanxiang Che, Ting Liu, Shijin Wang, and Guoping Hu. TextBrewer: An Open-Source Knowledge Distillation Toolkit for Natural Language Processing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pp. 9–16, Online, July 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-demos.2. URL https://aclanthology.org/2020.acl-demos.2.