A Cross-Lingual Question Answering Architecture Based on Pre-trained Token-Free Model with Instruction Finetuning

Yi Dai, Franklin Kong, Chenshun Ni, Xiyu Tian {yid, ftk, nichensh, xiyutian}@umich.edu
Department of Electrical Engineering and Computer Science, Chemical Engineering, Statistics
University of Michigan, Ann Arbor, 48109

Motivation

A mature cross-lingual question answering (CLQA) system, which allows people to ask questions in mother language and to acquire accurate answers in also mother language, can break down the knowledge barrier and enhance the communication efficiency between different countries and cultures.
However, there are some challenges:

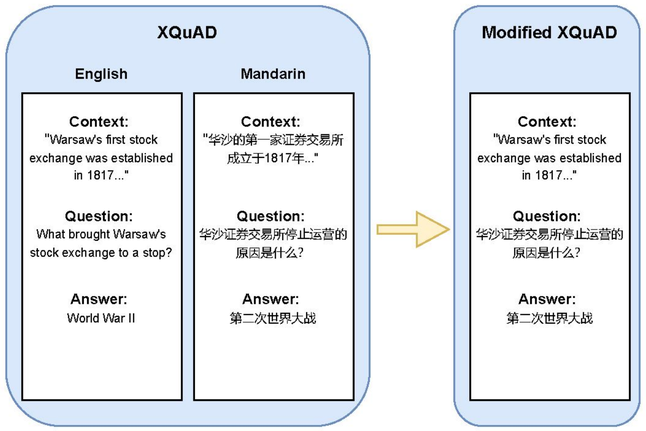
- Lack of relevant tools and resources
- Vocabulary gap between source and target languages.

In this project, we tackle the CLQA problem by fine-tuned state-of-the-art ByT5 in 3 ways:

- Directly fine-tuning on the CLQA dataset
- Instruction tuning on FLAN dataset, then fine-tuning on the CLQA dataset
- Inserting machine translation modules into the CLQA system with directly fine-tuned ByT5 and instruction tuned ByT5

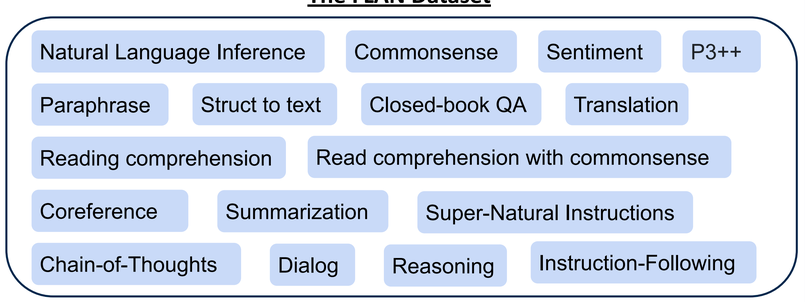
Dataset

Modified XQuAD



A modified XQuAD[1] dataset has been proposed, which consists of 240 paragraphs and 1190 question-answer pairs. Context and QA pairs are in different languages such that our model's information extraction and ability to answer questions are not dependent on the source paragraph's language.

The FLAN Dataset



The FLAN Dataset contains the FLAN 2022 Collection[2] and Instruction-Following Data[3]. The FLAN 2022 Collection offers the most extensive publicly available set of tasks for instruction tuning, which have been compiled in one place. This dataset has been supplemented with hundreds more of Google's high-quality templates, richer formatting patterns, and data augmentations, which consists of 16 tasks and more than 70 text datasets, including both language understanding and language generation tasks, into a single mixture. The Instruction-Following Data in both English and Chinese are generated by GPT-4.
The training and test datasets for fine-tuning language models are comprised of the aforementioned datasets.

Language Model – ByT5

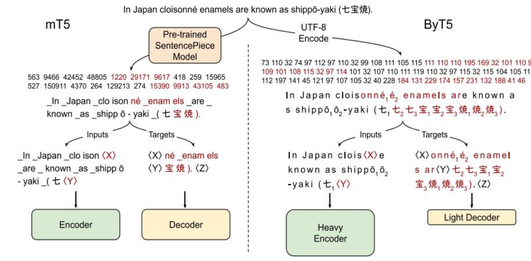
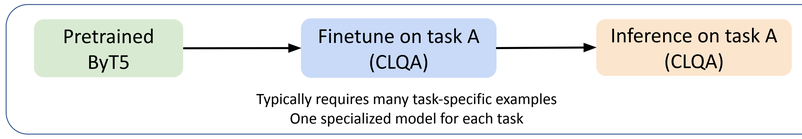


Figure 1: Pre-training example creation and network architecture of mT5 vs. ByT5. **mT5**: Text is split into SentencePiece tokens, spans of ~3 tokens are masked (red), and the encoder/decoder transformer stacks have equal depth. **ByT5**: Text is processed UTF-8 bytes, spans of ~20 bytes are masked, and the encoder is 3 times deeper than the decoder. <X>, <Y>, and <Z> represent sentinel tokens.

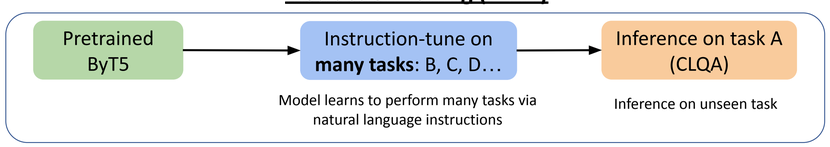
ByT5[4], a token-free variant of multilingual T5[5] that simplifies the NLP pipeline by doing away with vocabulary building, text preprocessing and tokenization. ByT5 outperforms mT5 in any of these five scenarios:
(1) at model sizes under 1 billion parameters;
(2) on generative tasks;
(3) on multilingual tasks with in -language labels
(4) on word-level tasks sensitive to spelling and/or pronunciation
(5) in the presence of various types of noise.

Instruction Tuning[6]

Pretrain-finetune (ByT5)

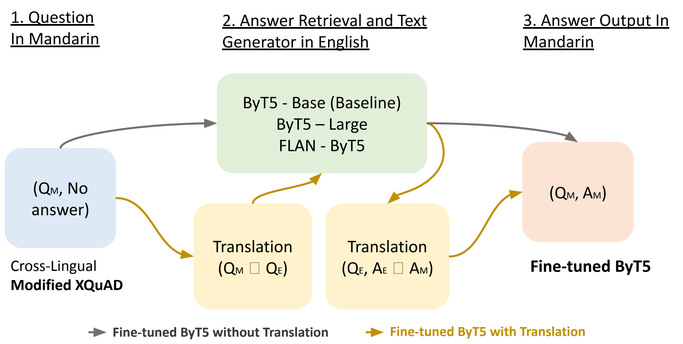


Instruction Tuning (FLAN)



An illustration of how FLAN works: the model is fine-tuned on disparate sets of instructions and generalizes to unseen instructions. As more types of tasks are added to the fine-tuning data, the model performance improves.

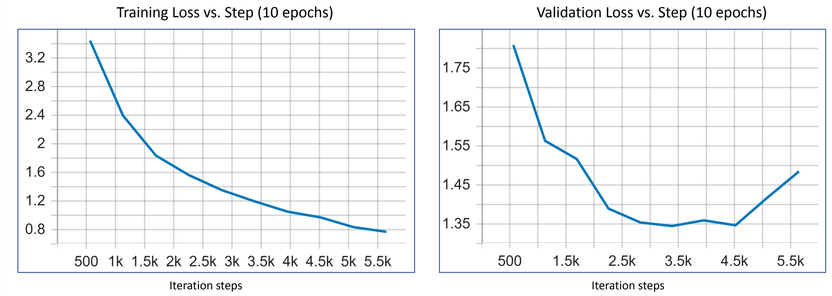
QA Pipeline in Fine-tuned ByT5



On the foundation of baseline, we fine-tuned ByT5 by using instruction fine-tuning[6] which performs well on monolingual LLMs (FLAN-ByT5). The fine-tuning methods consist of all different variants of ByT5 shown in the green box of pipeline schematic. To compare with the CLQA task performance of proposed FLAN-ByT5, an architecture with machine translation will be tested within the same training dataset.

Results and Discussion

ByT5-Base LM with Fine-Tuning on the modified XQuaD Dataset (Baseline)



Decreasing training loss and validation loss demonstrate fine-tuned ByT5 successfully.

WORKING EXAMPLES:

CONTEXT:
In the early 1970s, ABC completed its transition to color; the decade as a whole would mark a turning point for ABC, as it began to pass CBS and NBC in the ratings to become the first place network. It also began to use behavioral and demographic data to better determine what types of sponsors to sell advertising slots to and provide programming that would appeal towards certain audiences….
QUESTION: 美国广播公司在什么年代完成了向彩色电视的过渡?
(In what decade did ABC finish transitioning to color?)
ANSWER: 20世纪70年代
(1970s, lit. "20th century, 70s")

PREDICTION: 1970年代
(The 1970s)

CONTEXT:
The Broncos defeated the Pittsburgh Steelers in the divisional round, 23–16, by scoring 11 points in the final three minutes of the game. They then beat the defending Super Bowl XLIX champion New England Patriots in the AFC Championship Game, 20–18, by intercepting a pass on New England's 2-point conversion attempt with 17 seconds left on the clock…
QUESTION: 谁在分区轮输给了野马队?
(Who lost to the Broncos in the divisional round?)
ANSWER: 匹兹堡钢人队
(Pittsburgh Steelers)

PREDICTION: 匹兹堡队
(Team Pittsburgh)

FAILING EXAMPLES:

CONTEXT:
Frederick William, Elector of Brandenburg, invited Huguenots to settle in his realms, and a number of their descendants rose to positions of prominence in Prussia. Several prominent German military, cultural, and political figures were ethnic Huguenot, including poet Theodor Fontane, General Hermann von François, …
QUESTIONS: 哪位德国诗人是胡格诺派后裔?
(What German poet was descended from Huguenots?)
ANSWER: 狄奥多·冯塔纳
(Theodor Fontane)

PREDICTION: 德国诗人
(German Poet)

CONTEXT:
A piece of paper was later found on which Luther had written his last statement. The statement was in Latin, apart from "We are beggars," which was in German.

QUESTION: 声明的大部分是用什么语言写的?
(In what language was most of the statement written?)
ANSWER: 拉丁语
(Latin)

PREDICTION: 国语
(The national language)

Fine-tuned ByT5 with XQuaD in Chinese performs better than the modified XQuaD.

Conclusion and Future Work

Conclusion:

- The pre-trained ByT5-base LM has been successfully fine-tuned on the modified XQuAD dataset.
- The fine-tuned ByT5-base LM has better performance on CLQA task than without fine-tuning one.
- The prediction of the test dataset demonstrates that the model is learning such a structure by successfully identifying sorts of answers.

Future Work:

- Instruction tuning ByT5-base LM on the FLAN dataset and the modified XQuAD dataset (FLAN-ByT5).
- Using 'Rouge' and 'Bleu' metrics to measure the performance of the baseline and FLAN-ByT5 model more explicitly.
- Performing machine translation modules into the QA task before feeding the data set into variant ByT5 models to compare with the aforementioned fine-tuned ByT5 models on the CLQA task.

Reference

[1] M. Artetxe, S. Ruder, and D. Yogatama, "On the cross-lingual transferability of monolingual representations," *arXiv preprint arXiv:1910.11856*, 2019
[2] S. Longpre, L. Hou, T. Vu, A. Webson, H. W. Chung, Y. Tay, D. Zhou, Q. V. Le, B. Zoph, J. Wei et al., "The flan collection: Designing data and methods for effective instruction tuning," *arXiv preprint arXiv:2301.13688*, 2023. 8
[3] B. Peng, C. Li, P. He, M. Galley, and J. Gao, "Instruction tuning with gpt-4," *arXiv preprint arXiv:2304.03277*, 2023. 8
[4] L. Xue, A. Barua, N. Constant, R. Al-Rfou, S. Narang, M. Kale, A. Roberts, and C. Raffel, "Byt5: Towards a token-free future with pre-trained byte-to-byte models," *Transactions of the Association for Computational Linguistics, vol. 10, pp. 291–306*, 2022
[5] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, and C. Raffel, "mt5: A massively multilingual pre-trained text-to-text transformer," *arXiv preprint arXiv:2010.11934*, 2020
[6] J. Wei, M. Bosma, V. Y. Zhao, K. Guu, A. W. Yu, B. Lester, N. Du, A. M. Dai, and Q. V. Le, "Finetuned language models are zero-shot learners," *arXiv preprint arXiv:2109.01652*, 2021