TREC NeuCLIR 2022 - CFDA & CLIP Lab

Cross-language Passage Re-ranking

Jia-Huei Ju[†], Wei-Chih Chen[†], Heng-Ta Chang[†], Cheng-Wei Lin[†]
Ming-Feng Tsai[‡] and Chuan-Ju Wang[†]

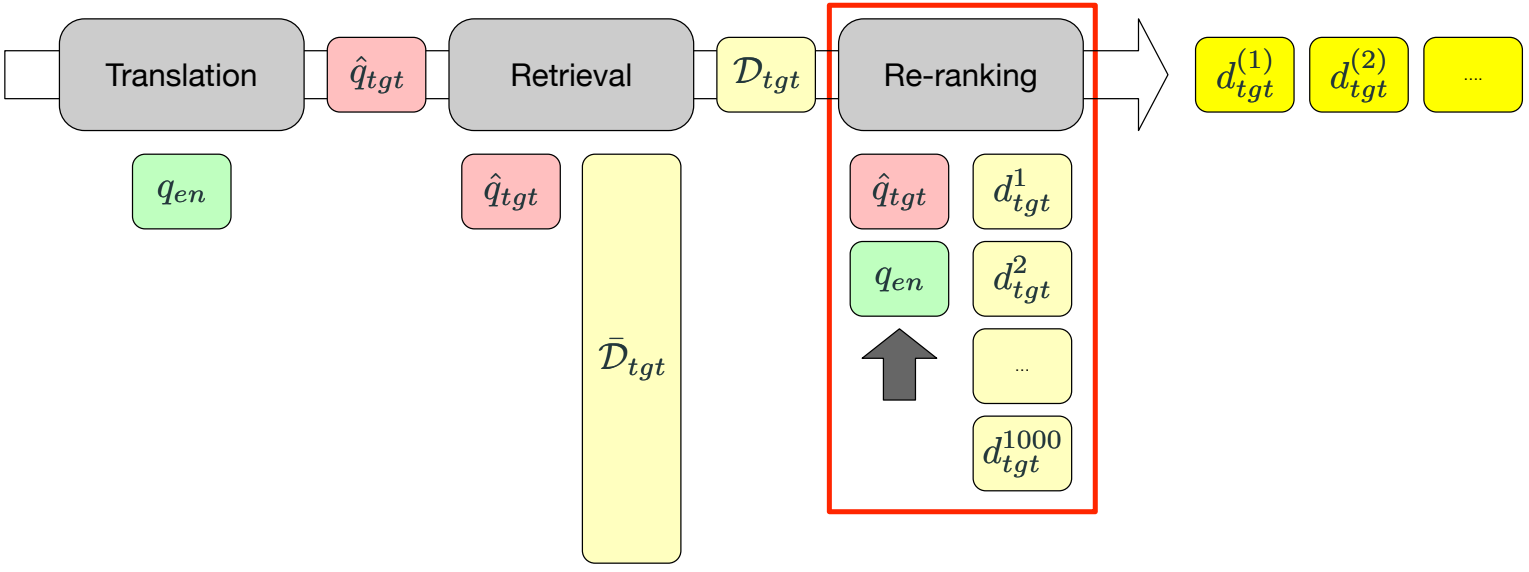
Presenter: (Dylan) Jia-Huei Ju

[†] Research Center for Information Technology Innovation, Academia Sinica,
[‡] Department of Computer Science, National Chengchi University

Table of Contents

Our multi-stage pipeline

The multi-stage pipeline for CLIR



- Query translation: Google's translator, NLLB [5].
- Candidate passage retrieval: Top-1000 passages from BM25(+PRF).
- **Cross-language passage re-ranking: using $\hat{q}_{tgt}$, and q_{en} as well.**

Cross-language passage re-ranking

Cross-language Query

We fine-tuned mT5 models [6] for passage re-ranking followed [1]:

- Dataset: mMARCO [1]
- Iteration: 100K fine-tuning steps.
- Objective: "yes" and "no" target tokens for d^{+}_{tgt} and d^{-}_{tgt} passages.

Particularly, we constructed two settings of **cross-language query**:

Settings		mT5 Text-to-text Formulation
Baseline	mT5	Query: $\hat{q}_{tgt}$ Document: d_{tgt} Relevant:
Cross-lingual query	mT5-cl	Query: q_{en} Document: d_{tgt} Relevant:
Bilingual query	mT5-bq	Query: q_{en} Query Translation: $\hat{q}_{tgt}$ Document: d_{tgt} Relevant:

Experiments & Results

Experiments & Results

Experimental setups:

- Evaluation data: HC4 testing query and collections.
  - in Persian, Russian and Chinese.
- Top-1000 candidate passages: retrieved from BM25 with Human-translated queries.

Zero-shot capability.

Rerankers	Size	nDCG@20	mAP@20	MAP@100	MAP@1K
Target language: Persian (fas)					
mT5	large	0.5488	0.3987	0.4253	0.4285
mT5-cl	large	0.5491	0.4078	0.4296	0.4330
mT5-bq	large	**0.5644**	**0.4123**	**0.4411**	**0.4442**

Experiments & Results

Inconsistent impact of cross-language query.

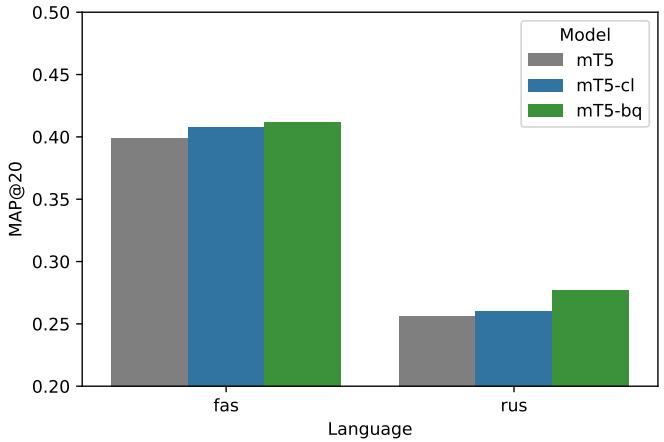
The opposite trends in Chinese, especially mT5-bq (i.e., bilingual query).

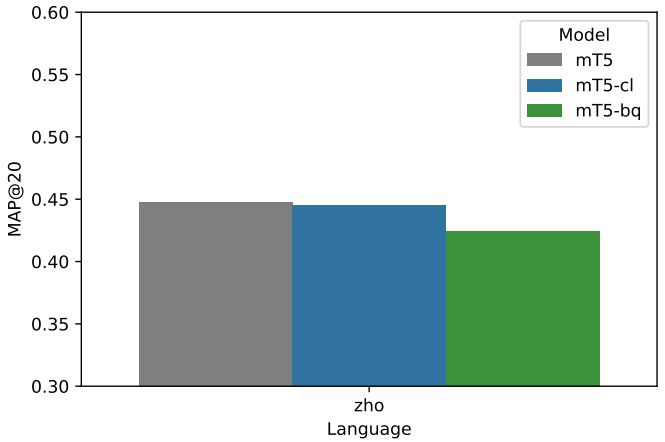
Rerankers	Size	nDCG@20	mAP@20	MAP@100	MAP@1K
Target language: Russian (rus)					
mT5-mono	large	0.3698	0.2564	0.3168	0.3243
mT5-cl	large	0.3757	0.2603	0.3172	0.3251
mT5-bq	large	**0.3822**	**0.2768**	**0.3377**	**0.3450**
Target language: Chinese (zho)					
mT5-mono	large	0.5778	**0.4473**	**0.4817**	**0.4851**
mT5-cl	large	**0.5924**	0.4450	0.4794	0.4823
mT5-bq	large	0.5743	0.4246	0.4574	0.4621

Experiments & Results

Take MAP@20 judgement as an example,

Persian and Russian (left) v.s. **Chinese** (right) showed differently.





The **linguistic English-Chinese gap** is larger than English-Russian

- Linguistic differences: grammar, tokenization, inversion, sentence
- Confused attention: results in ineffective contextualization
- More empirical evaluation[1]

[1] mT5-bq is the best one among other our submitted runs in Chinese :)

Conclusion

Conclusion

Cross-lingual IR pre-trained language models

- Cross-language pre-training. (e.g. TLM [3])
- Retrieval-oriented pre-training. (e.g. coCondenser [2], ICT [4])

Dense retriever

- Multi-tasking (e.g. query translation)

References i

[1] L. Bonifacio, V. Jeronymo, H. Q. Abonizio, I. Campiotti, M. Fadaee, R. Lotufo, and R. Nogueira. mmarco: A multilingual version of the ms marco passage ranking dataset, 2021. URL https://arxiv.org/abs/2108.13897.

[2] L. Gao and J. Callan. Unsupervised corpus aware language model pre-training for dense passage retrieval, 2021. URL https://arxiv.org/abs/2108.05540.

[3] G. Lample and A. Conneau. Cross-lingual language model pretraining, 2019. URL https://arxiv.org/abs/1901.07291.

[4] K. Lee, M.-W. Chang, and K. Toutanova. Latent retrieval for weakly supervised open domain question answering, 2019. URL https://arxiv.org/abs/1906.00300.

[5] NLLB Team, M. R. Costa-jussà, J. Cross, O. Çelebi, M. Elbayad, K. Heafield, K. Heffernan, E. Kalbassi, J. Lam, D. Licht, J. Maillard, A. Sun, S. Wang, G. Wenzek, A. Youngblood, B. Akula, L. Barrault, G. Mejia-Gonzalez, P. Hansanti, J. Hoffman, S. Jarrett, K. R. Sadagopan, D. Rowe, S. Spruit, C. Tran, P. Andrews, N. F. Ayan, S. Bhosale, S. Edunov, A. Fan, C. Gao, V. Goswami, F. Guzmán, P. Koehn, A. Mourachko, C. Ropers, S. Saleem, H. Schwenk, and J. Wang. No language left behind: Scaling human-centered machine translation. 2022.

[6] 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. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Thank You!

Are there any questions you'd like to ask?

Jia-Huei Ju	jhjoo@citi.sinica.edu.tw
Ming-Feng Tsai	mftsai@nccu.edu.tw
Chuan-Ju Wang	cjwang@citi.sinica.edu.tw