Text-to-text Multi-view Learning for Passage Re-ranking

Jia-Huei Ju,[†] Jheng-Hong Yang,[‡] and Chuan-Ju Wang[†]

June 11, 2021

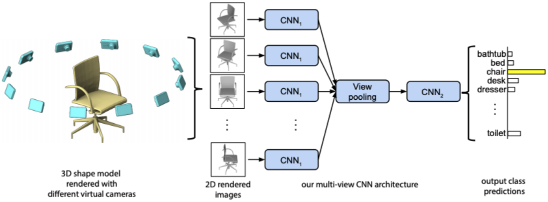
[†] Research Center for Information Technology Innovation, Academia Sinica
[‡] David R. Cheriton School of Computer Science, University of Waterloo

Table of Contents

Introduction

Introduction: Multi-view Learning

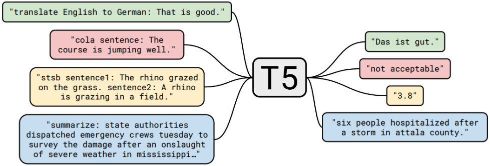
- Better representation by leveraging multiple views.
  - More **generalized** and less overfitting result.
  - For example on CV, the 3D object recongnition [5]:



- How to apply this idea on text (NLP)?
  - Backbone: Text-to-text Transfer Transformer [4] aka T5

Introduction: T5 model

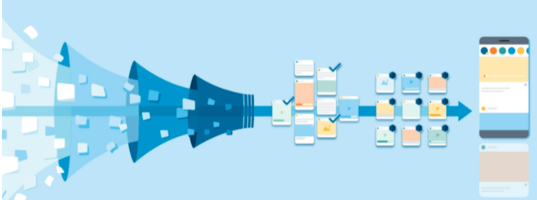
- How T5 works?
  - Train with different NLP tasks



  - Formulate each with "text-to-text" format
  - And also well-adapted to the pre-training technique.

Introduction: Document Ranking process

- Common two-stage IR architectures[1]



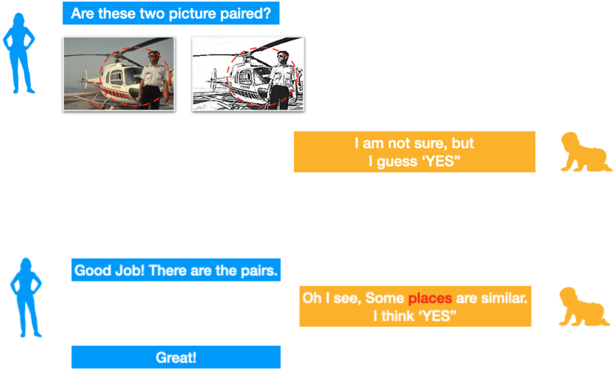
1. **Retrieve** from large collections: Using term-matching model BM25.
2. **Rank** on smaller subset: Using neural ranking model, such as BERT.

- BUT, there is still a potential issue: overfitting.
  - Model only learns to **discriminate** from shallow associations.
- Multi-view learning with additional "**generative** view" may be a solution to alleviate the shortcoming of the existing approach.

[1] Photo credit: Post by Akos Lada, Meihong Wang, Tak Yan

Example: Discriminative method

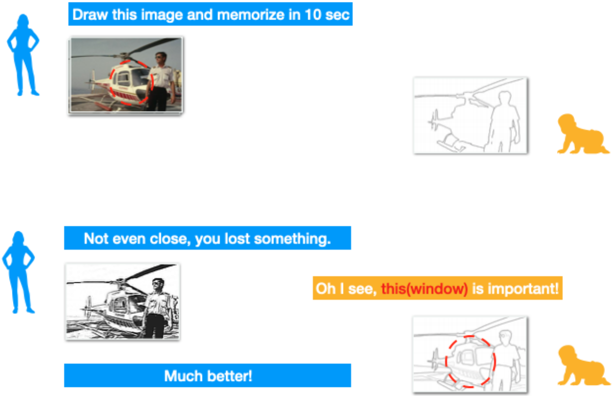
Teach a kid to classify the relevance (by "difference").



NO IDEA how to draw!

Example: Generative method

Teach a kid to copy the image. (memorize then draw).



Learned the representative part!

Methodology

Methodology: Train with two views

- Passage ranking task aka Rank (Discriminative)
- Query generation task[2] aka P2Q (Generative)

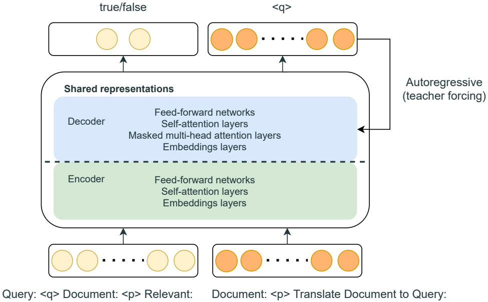


Figure 1: Text-to-text multi-view learning for the shared representations using the two objectives of passage ranking (left half) and text generation (right half).

Methodology: Mixing

Rank view & P2Q view (CE loss & NLL loss)

- $\mathcal{L}_{\text{Rank}}(q, p^+, p^-) = -\log P(\texttt{true} \mid q, p^+) - \log P(\texttt{false} \mid q, p^-)$
- $\mathcal{L}_{\text{P2Q}}(q, p) = -\sum_{t=1}^{|q|} \log P(q_{(t:t)} \mid q_{(1:t-1)}, p)$

Multi-view learning with mixing rate η[1]

$$\mathcal{L}_{\text{multi-view}} = (1 - X) \times \mathcal{L}_{\text{Rank}}(q, p^+, p^-) + X \times \mathcal{L}_{\text{P2Q}}(q, p)$$

- Mixing losses by proportion of training instances.

[1] $X \sim \mathrm{Bernoulli}(\eta)$: Note that the parameter η controls the sampling views, which is identical to the example proportional sampling.

Empirical Results

Effectiveness on MS MARCO Passage Ranking task

- Evaluated by official MRR@10 on 2 validation data (last 2 column)

#	Condition	Model	# Param (M)	Dev	Dev-Rest
	Baselines	BM25	-	0.187	0.191
		Best non-BERT [1]	-	0.290	-
		BM25 + BERT-large [3]	340	0.372	-
1	**Single-view**	BM25 +T5-base	220	0.384	0.380
2		BM25 +T5-large	770	0.395	0.390
3		BM25 +T5-3B	2,800	0.398	0.395
4	**Multi-view**	BM25 +T5-base	220	0.385	**0.382**[1]
5		BM25 +T5-large	770	**0.401**[2]	**0.393**[2]
6		BM25 +T5-3B	2,800	0.402	0.396

Table 1: Comparison on overall ranking effectiveness (MRR@10). The scores are in boldface if they are significantly better than the compared condition (see the superscript) under a paired t-test with $p \leq 0.05$.

Effectivenss at different depth k (candidates)

- Improvement is noted as $\frac{\text{MRR@10}_{\text{multi}} - \text{MRR@10}_{\text{single}}}{\text{MRR@10}_{\text{single}}}$ (growth)

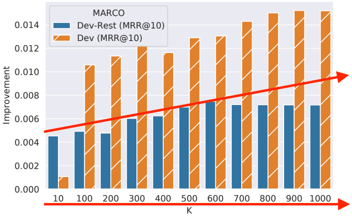


Figure 2: Improvement of MRR@10 with top-K candidates based on the BM25. The re-ranking model is T5-large (multi-view versus single-view).

- Performance improved more even in the noisy environment (more candidates.)

Future Work

Future Work

Fuse more views:

- (P2Q-) Negative P2Q view: Try to generate the irrelevant passage.
- (P2W) Term generative view: Try to extract the keywords of the passage.

Improve the primary task (Rank view):

- Fusing BM25 score: Consider relative scores between candidates, since our reranker is only based on pointwise approach.

References i

[1] S. Hofstätter, N. Rekabsaz, C. Eickhoff, and A. Hanbury. On the effect of low-frequency terms on neural-ir models. In *Proc. of SIGIR*, page 1137–1140, 2019.

[2] R. Nogueira, J. Lin, and A. Epistemic. From doc2query to docttttquery. *Online preprint*, 2019.

[3] R. Nogueira, W. Yang, K. Cho, and J. Lin. Multi-stage document ranking with bert. *arXiv:1910.14424*, 2019.

[4] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(140):1–67, 2020.

[5] H. Su, S. Maji, E. Kalogerakis, and E. Learned-Miller. Multi-view convolutional neural networks for 3d shape recognition. In *Proc. of ICCV*, pages 945–953, 2015.

Thank You!

Are there any questions you'd like to ask?

Jia-Huei Ju	dylanjootw@gmail.com
Jheng-Hong Yang	j587@uwaterloo.ca
Chuan-Ju Wang.	cjwang@citi.sinica.edu.tw