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
- Multi-view Learning
- T5 model
- Passage Ranking Process
- Example

Methodology
- Train with two views
- Mixing

Empirical Results
- MS MARCO Passage Ranking
- Effectiveness different depth $k$

Future Work
Introduction
**Introduction: Multi-view Learning**

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

```
<table>
<thead>
<tr>
<th>3D shape model rendered with different virtual cameras</th>
</tr>
</thead>
</table>
```

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

```
<table>
<thead>
<tr>
<th>2D rendered images</th>
<th>our multi-view CNN architecture</th>
</tr>
</thead>
<tbody>
<tr>
<td>CNN_1</td>
<td>View pooling</td>
</tr>
<tr>
<td>CNN_2</td>
<td>CNN_3</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
</tbody>
</table>
```

```
<table>
<thead>
<tr>
<th>output class predictions</th>
</tr>
</thead>
</table>
| bathtub | bed | chair | desk | dresser | toilet | ...
```
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").

Are these two pictures paired?

I am not sure, but I guess "YES"

Good Job! There are the pairs.

Oh I see, Some places are similar. I think "YES"

Great!

NO IDEA how to draw!
Example: Generative method

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

Draw this image and memorize in 10 sec

Not even close, you lost something.

Oh I see, this (window) is important!

Much better!

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)

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

### Multi-view learning with mixing rate \( \eta \)

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

- Mixing losses by proportion of training instances.

---

\( ^{1}X \sim \text{Bernoulli}(\eta) \): Note that the parameter \( \eta \) 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)

<table>
<thead>
<tr>
<th>#</th>
<th>Condition</th>
<th>Model</th>
<th># Param (M)</th>
<th>Dev</th>
<th>Dev-Rest</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Baselines</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>1</td>
<td></td>
<td>BM25</td>
<td>-</td>
<td>0.187</td>
<td>0.191</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Best non-BERT [1]</td>
<td>-</td>
<td>0.290</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>BM25 + BERT-large [3]</td>
<td>340</td>
<td>0.372</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Single-view</td>
<td>BM25 + T5-base</td>
<td>220</td>
<td>0.384</td>
<td>0.380</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>BM25 + T5-large</td>
<td>770</td>
<td>0.395</td>
<td>0.390</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>BM25 + T5-3B</td>
<td>2,800</td>
<td>0.398</td>
<td>0.395</td>
</tr>
<tr>
<td>5</td>
<td>Multi-view</td>
<td>BM25 + T5-base</td>
<td>220</td>
<td>0.385</td>
<td>0.382₁</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>BM25 + T5-large</td>
<td>770</td>
<td>0.401²</td>
<td>0.393²</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BM25 + T5-3B</td>
<td>2,800</td>
<td>0.402</td>
<td>0.396</td>
</tr>
</tbody>
</table>

**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 $p \leq 0.05$. 
Effectiveness at different depth $k$ (candidates)

- Improvement is noted as $\frac{MRR_{10\text{multi}} - MRR_{10\text{single}}}{MRR_{10\text{single}}}$ (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.


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