SLWWW at the NTCIR-16 WWW-4 Task

Yuya Ubukata, Masaki Muraoka, Sijie Tao, Tetsuya Sakai
Waseda University

Table of Contents

- Introduction
- Methodology
  - COIL
  - PARADE
  - Run details
- Results
  - NEW runs
  - REP runs
- Discussion
  - per-topic analysis
- Conclusion

Introduction

SLWWW team participated in the NTCIR-16 WWW-4 task

What we have done for the task

- Took 2 different approaches to generate NEW runs
  - COIL [1]
  - PARADE [2]
- Reproduced the KASYS run at the NTCIR-15 WWW-3 task [3]
- Performed per-topic analyses for further discussion
  - Poorly performing topics overall
  - Effect of document length
  - Comparison of COIL and BM25

[1] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List.
[2] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2021. PARADE: Passage Representation Aggregation for Document Reranking
[3] Kohei Shinden, Atsuki Maruta, and Makoto P. Kato. 2020. KASYS at the NTCIR-15 WWW-3 Task. In Proceedings of NTCIR-15. 235–238.

Methodology

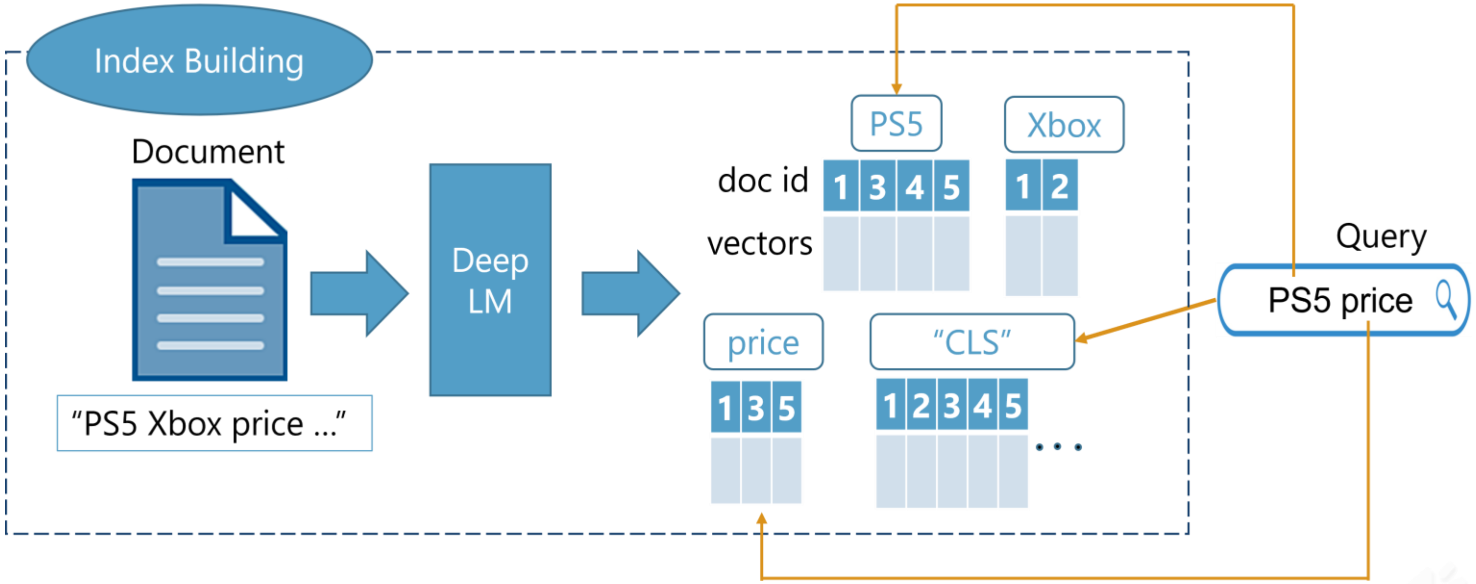
COIL (Contextualized Inverted List)

Exact lexical matching → Cannot handle vocabulary mismatch problems

Soft matching → Loses computational efficiency

COIL introduces contextualized vector representations into the exact matching framework to incorporate the best of two systems

Retrieval architecture that stores representation vectors into inverted lists to perform contextualized exact matching



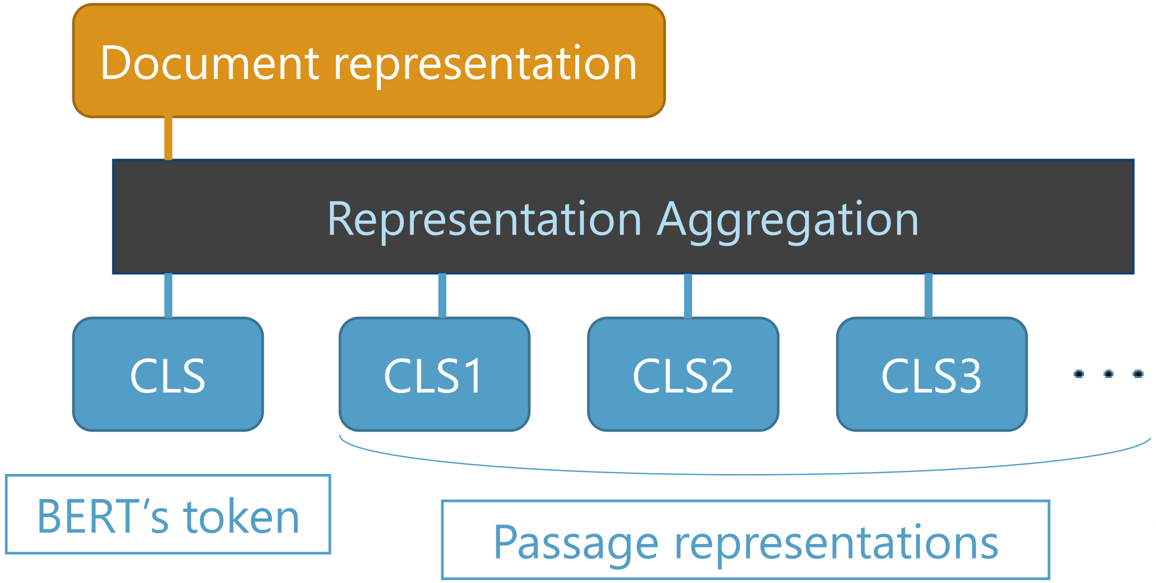
[2] Luyu Gao, Zhuyun Dai, and Jamie Callan. 2021. COIL: Revisit Exact Lexical Match in Information Retrieval with Contextualized Inverted List.

PARADE (Passage Representation Aggregation for Document Reranking)

- A Problem in using BERT for document ranking task
  - The input length limit of 512 token
    → Unable to handle long documents
  - Many approaches have been proposed to overcome this problem

PARADE aggregates passage representations to gain overall document representation

- Documents are split into a fixed number of passages
- Passage representations are computed for each passage-query pair by a pretrained transformer encoder



Run Details

Reproduced run

SLWWW-CO-REP-1

PARADE

SLWWW-CO-NEW-5

COIL

Run name	Divided into chunks of 510 tokens	Corpus type
SLWWW-CO-NEW-2	✓	A
SLWWW-CO-NEW-3	✓	B
SLWWW-CO-NEW-4		A

Corpus A: compilation of the top 1,000 most relevant documents for each topic, extracted by BM25

Corpus B: compilation of the top 10,000 most relevant documents for each topic, extracted by BM25

Results

Results (NEW runs)

- No statistically significant differences were observed
- SLWWW-CO-NEW-4 performed well in terms of nDCG and Q
- Splitting documents and using a larger corpus did not contribute to the search effectiveness …
- NEW runs based on COIL outperform the baseline

  → shows the effectiveness of contextualized representations

Results of NEW runs based on the Gold file

Run	nDCG	Q	nERR	iRBU
SLWWW-CO-NEW-2	0.3398	0.2718	0.5129	0.7358
SLWWW-CO-NEW-3	0.3388	0.2670	**0.5248**	0.7368
SLWWW-CO-NEW-4	**0.3650**	**0.2891**	0.5052	**0.7986**
SLWWW-CO-NEW-5	0.3193	0.2538	0.4288	0.7133
baseline	0.3205	0.2473	0.4541	0.7327

Results of NEW runs based on the Bronze-All file

Run	nDCG	Q	nERR	iRBU
SLWWW-CO-NEW-2	0.5600	0.5316	**0.7330**	**0.9244**
SLWWW-CO-NEW-3	0.5464	0.5137	0.7242	0.9192
SLWWW-CO-NEW-4	**0.5750**	**0.5397**	0.7209	0.9213
SLWWW-CO-NEW-5	0.5410	0.5113	0.6939	0.8888
baseline	0.5170	0.4806	0.6711	0.8920

Results (REP run)

Results of REP run based on the Gold file

Run	nDCG	Q	nERR	iRBU
SLWWW-CO-REP-1	0.3686	0.2886	0.5098	0.7840
KASYS-CO-REV-6	0.3682	0.2890	0.5098	0.7811

Results of REP run based on the Bronze-All file

Run	nDCG	Q	nERR	iRBU
SLWWW-CO-REP-1	0.5846	0.5629	0.7537	0.9397
KASYS-CO-REV-6	0.5931	0.5743	0.7634	0.9424

- Our REP run and the KASYS team’s REV run performed very similarly
  → Succeeded in reproducing the target run to some degree
  → Although we used the provided fine-tuned model …

Topic Analysis

Poorly performing topics

Topic ID	Content	Description	Mean nDCG
203	idf inventor	I want to know if my search engine can find who invented inverse document frequency.	0.0000
220	half life	I'm looking for information about Half-Life, the story, and the characters	0.1491
234	Warriors v.s. NETS 2021	You want to find the NBA match information between Warriors team and NETS team in 2021	0.1517
228	block chain crypto	You want to know the relationship between block chain and crypto currencies	0.2392

Highly rated documents by our runs included …

"idf inventor"

- Different "idf"
  - India Design Forum, Intel Developer Forum, Israeli Defense Forces

"half life"

- Radioactive half-life, Biological half-life
- Download page, news article

Topics are ambiguous and have multiple intents

COIL vs. BM25

- COIL (Run 4): Lexical matching framework with contextualized representations
- BM25 (baseline): Conventional lexical matching methods

  - Documents rated higher in BM25 are those that contain the words contained in the topic as they are
  - Topics with poor COIL results are cases where contextual information is taken into account, which in turn leads to a discrepancy with the intent of the topic.

nDCG of the topics where Run 4 significantly outperformed the baseline

Topic	Content	Run 4	baseline	difference
214	inventor of the Web	0.7461	0.0568	0.6893
234	Warriors v.s. NETS 2021	0.5273	0.0000	0.5273

nDCG of the topics where the baseline significantly outperformed Run 4

Topic	Content	Run 4	baseline	difference
210	hypothermia treatment	0.1357	0.6303	0.4946
240	what is clickbait	0.3748	0.8249	0.4501

Conclusions

Conclusions

- Our NEW runs outperformed the BM25 baseline
- COIL showed the effectiveness of introducing contextualized vector representations
- Splitting input documents and using a larger corpus did not improve the results
- Successfully reproduced the KASYS team’s run

Future Work

- Reproduce the KASYS team’s run using our own fine-tuned model
- Create a system that also uses the description field

Thank you for your attention