SLWWW at the NTCIR-16 WWW-4 Task

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

INTRODUCTION

Our work for the WWW-4 task

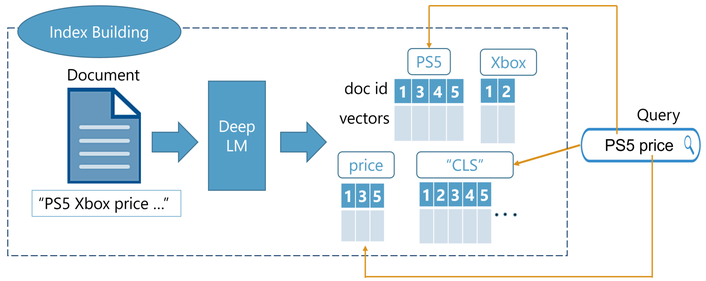
- 2 different approaches to generate NEW runs
  - COIL
  - PARADE
- Reproduced the KASYS run at the NTCIR-15 WWW-3 task
- Performed per-topic analyses for further discussion

COIL

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 the two systems



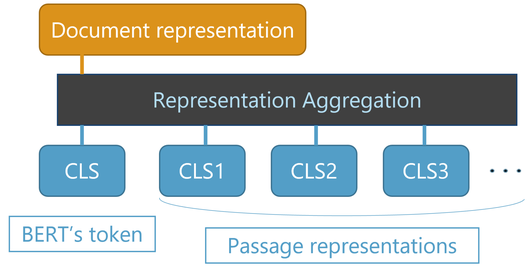
PARADE

A problem in using BERT for document ranking task

The input length limit of 512 token

→ Unable to handle long documents like web documents

PARADE aggregates passage representations to gain overall document representation



RUN DETAILS

Run name	Method	Divided into chunks of 510 tokens	Corpus type
SLWWW-CO-REP-1	Birch	–	–
SLWWW-CO-NEW-2	COIL	✓	A
SLWWW-CO-NEW-3	COIL	✓	B
SLWWW-CO-NEW-4	COIL		A
SLWWW-CO-NEW-5	PARADE	–	–

RESULTS

- SLWWW-CO-NEW-4 performed well in terms of nDCG and Q
- NEW runs based on COIL outperform the baseline
- Our REP run and the KASYS team's REV run performed very similarly

Results of NEW runs based on the Gold file

Run	nDCG	Q	nERR	iRBU
NEW-2	0.3398	0.2718	0.5129	0.7358
NEW-3	0.3388	0.2670	**0.5248**	0.7368
NEW-4	**0.3650**	**0.2891**	0.5052	**0.7986**
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
NEW-2	0.5600	0.5316	**0.7330**	**0.9244**
NEW-3	0.5464	0.5137	0.7242	0.9192
NEW-4	**0.5750**	**0.5397**	0.7209	0.9213
NEW-5	0.5410	0.5113	0.6939	0.8888
baseline	0.5170	0.4806	0.6711	0.8920

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

TOPIC ANALYSIS

Poorly performing topics

- Topics that are ambiguous or have multiple intents performed poorly
  - idf inventor → India Design Forum, Intel Developer Forum, Israeli Defense Forces
  - half life → Radioactive half-life, Biological half-life

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

COIL vs. BM25

- 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.

Topics where Run 4 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

Topics where the baseline 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 & FUTURE WORK

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