Retrieval for Extremely Long Queries and Documents with RPRS: a Highly Efficient and Effective Transformer-based Re-Ranker

Arian Askari, Suzan Verberne, Amin Abolghasemi, Wessel Kraaij, Gabriella Pasi
Leiden University and University of Milano-Bicocca

Study

- A method for efficient and effective retrieval with Transformers where the queries and candidates documents are extremely lengthy:
  - E.g., the candidate documents' average length in a Patent prior art retrieval dataset is 10,001 words, with outliers up to 407,308 words.
  - This type of retrieval is commonly known as **Query-by-Document (QBD)** retrieval.
- We call our method RPRS which stands for a **R**e-Ranker based on a novel **P**roportional **R**elevance **S**core

Method

- **Assumption**: a candidate document d is likely to be relevant to a query document q if a large proportion of d's sentences are similar to q's sentences, and a large proportion of q's sentences are similar to d's sentences.
- We model the above assumption below:

$$PRS(q, d, S_{TK_q}, n) = QP(S_q, S_d, S_{TK_q}, n) \times DP(S_d, S_q, S_{TK_q}, n)$$

$$DP(S_d, S_q, S_{TK_q}, n) = \frac{Fd_{R_n}(S_d, S_q, S_{TK_q})}{\text{count of } d\text{'s sentences}}$$

$$QP(S_q, S_d, S_{TK_q}, n) = \frac{Fq_{R_n}(S_q, S_d, S_{TK_q})}{\text{count of } q\text{'s sentences}}$$

$$Fq_{R_n}(S_q, S_d, S_{TK_q}) = \sum_{q_s}^{S_q} \frac{|S_d \cap r_n(q_s, S_{TK_q})|}{|S_d \cap r_n(q_s, S_{TK_q})| + k1((1-b) + \frac{b \cdot dl}{avgdl})}$$

$$Fd_{R_n}(S_d, S_q, S_{TK_q}) = \sum_{d_s}^{S_d} \frac{\sum_{r_n}^{R_n} |\{d_s\} \cap r_n|}{\sum_{r_n}^{R_n} |\{d_s\} \cap r_n| + k1((1-b) + \frac{b \cdot dl}{avgdl})}$$

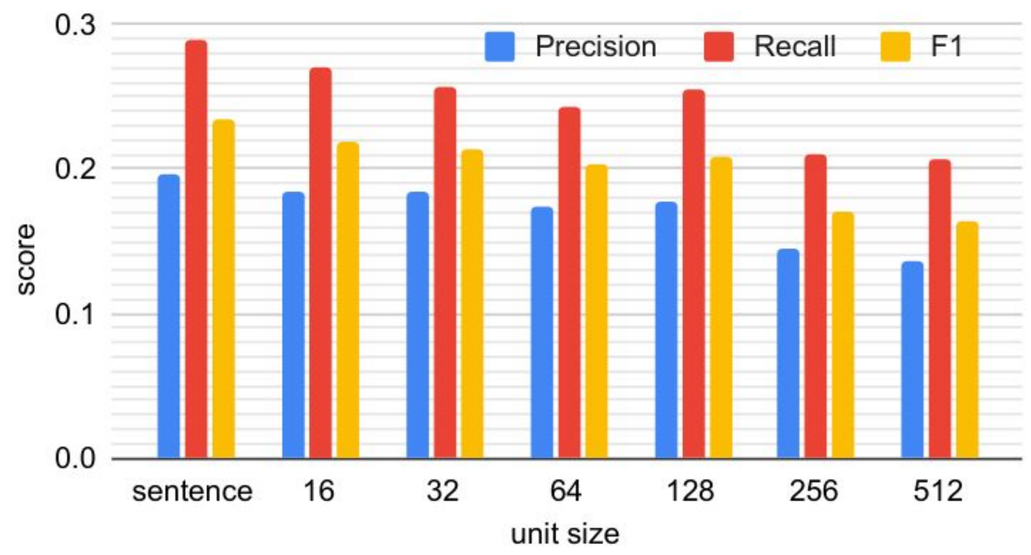
Results on COLIEE'21 dataset

	Model	Precision	Recall	F1
	Probabilistic lexical matching baselines			
a	BM25	0.0770	0.1959	0.1113
b	BM25 + KLI	0.0983	0.1980	0.1313
c	BM25optimized + KLI	0.1700	0.2536	0.2035
d	TLIR	0.1533	0.2556	0.1917
	Cross-encoders			
e	BERT	0.1340	0.2263	0.1683
f	Legal BERT	0.1440	0.2463	0.1817
g	MTFT-BERT [2] (Previous state-of-the-art)	0.1744	**0.2999**	0.2205
	Sentence-based baseline			
h	SDR_{inf}	.1470	0.2063	0.1716
	Proposed methods			
i	RPRS without frequency	0.1890	0.2799	0.2256
j	RPRS w/freq	**0.1960**†*	0.2891*	**0.2336**†*

- Our proposed methods outperforms previous state-of-the art models on the official F1 measure

Effect of using other units than sentences

- We investigate the effect of using embeddings of different textual units than sentences



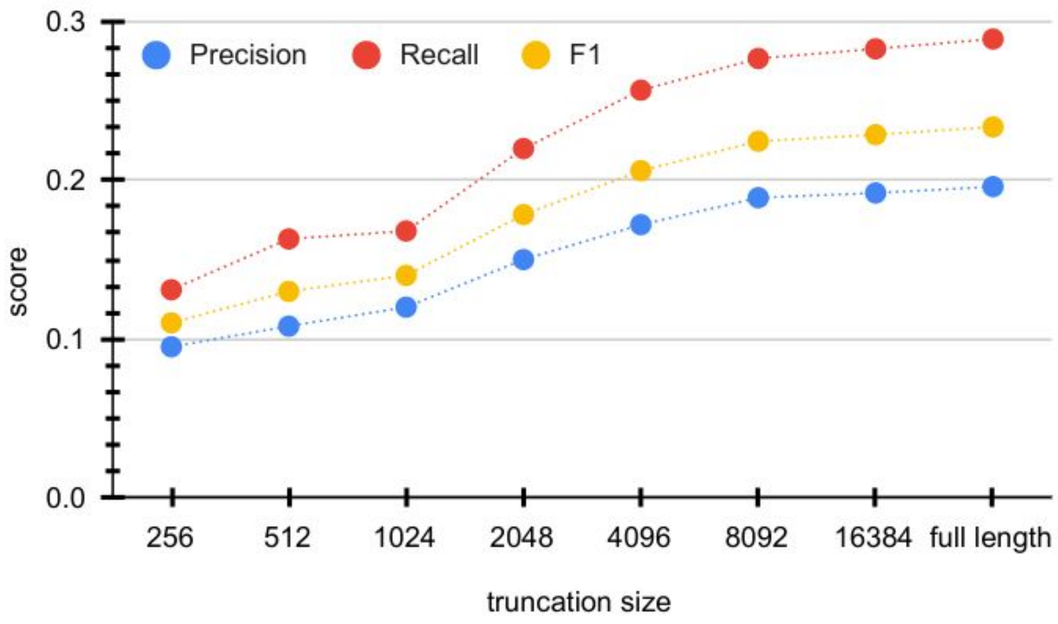
- The above Figure shows that sentences are better units than sequences with pre-defined fixed-length, also compared to sequences with fixed-length 16 and 32 which are similar in length to the average and median sentence length (29.4 and 24.8 tokens). This confirms the relevance of sentences as units for retrieval, one of the premises of RPRS.

Ablation study

		P	R	F1
	Full method (RPRS w/freq)	**0.1960**	**0.2891**	**0.2336**
a	No QP (Eq.4)	0.1670	0.2421	0.1976
b	No DP (Eq.5)	0.1410	0.2180	0.1712
c	No b and k1 params	0.1860	0.2691	0.2199
	RPRS without freq (Eq.1)	0.1890	0.2799	0.2256
d	No min function in PRS	0.1798	0.2644	0.2140

- The above table shows the full method with all components including QP, DP, and b and $k1$ parameters obtains highest effectiveness and supports the necessity of each component.

Effect of covering the full length of queries and documents



- Our re-ranker takes advantage of seeing the whole content of query and documents.

Main Results and Findings

- Our experiments on **seven** datasets show that our proposed method takes advantage of the long queries and documents and sets a new state-of-the-art performance on each dataset.
- The efficiency of our method is due to: (1) using a bi-encoder, SBERT; (2) the pre-processing, embedding, and indexing of document sentences could be done before the query time; (3) the only calculation is based on cosine similarity that is efficient operation.