Precise Document Retrieval by Minimizing Kolmogorov Distance with Document Generation

CS898 Course Project

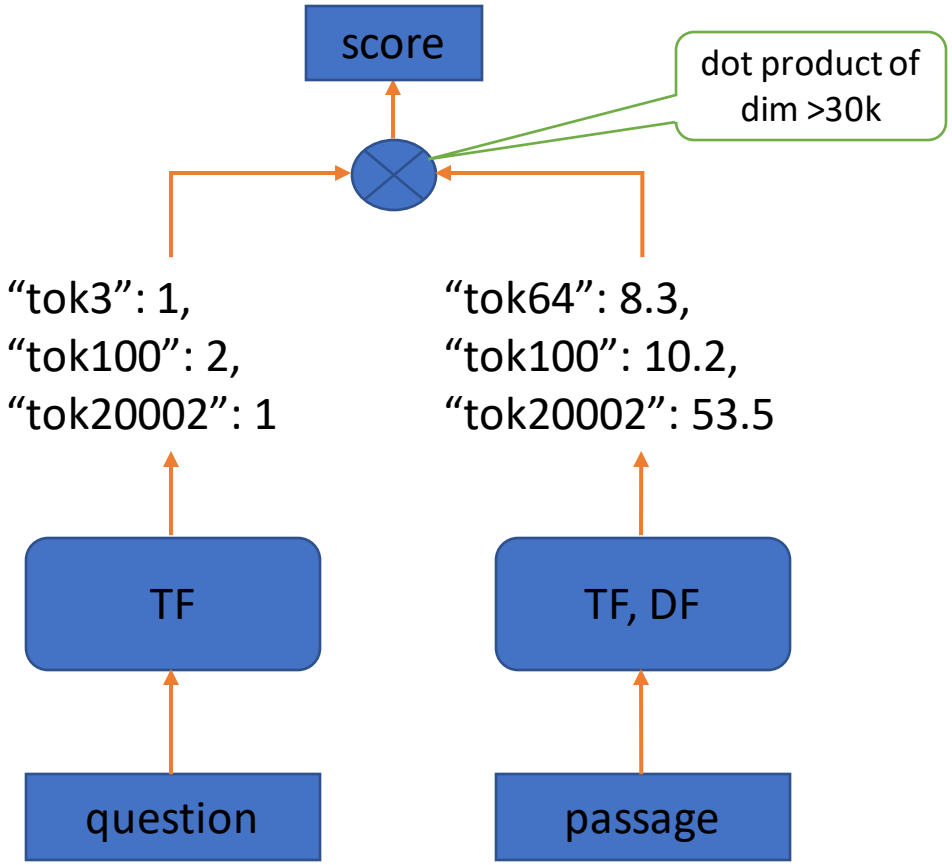
Xueguang Ma

Overview

- Text Retrieval
  - Lexical Retrieval
  - Dense Retrieval
- Kolmogorov Distance
  - Query-Documents Distance
- Minimizing Kolmogorov Distance by Document Generation

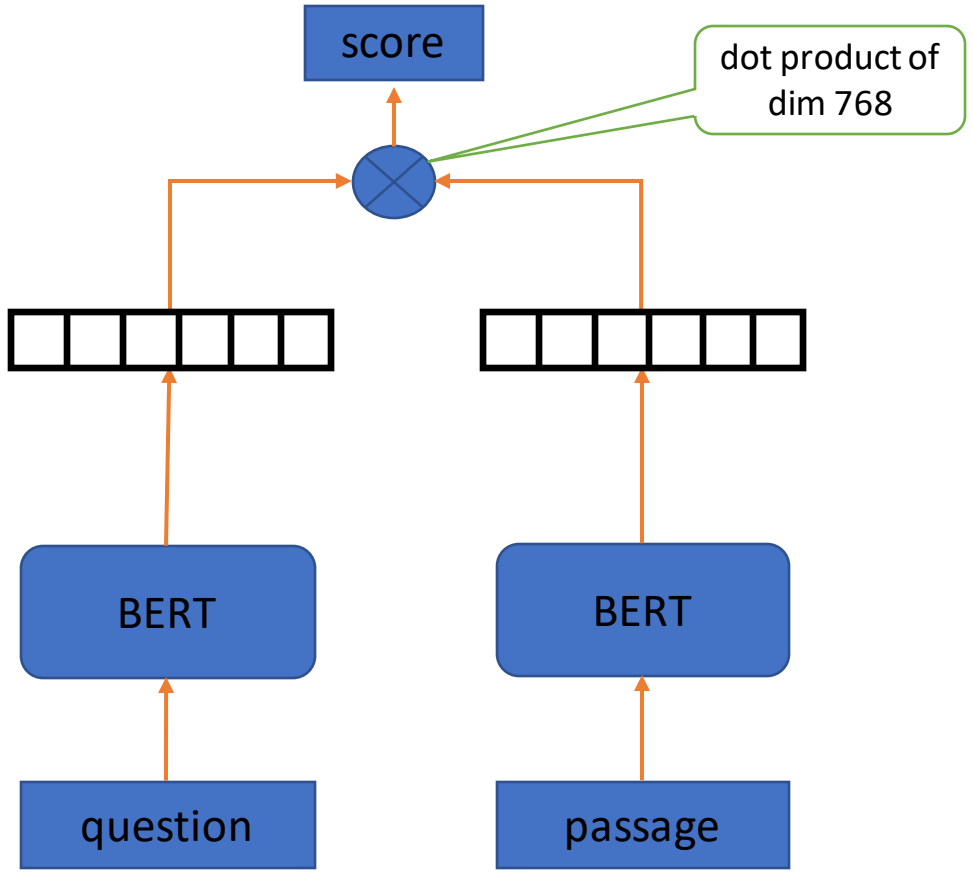
Text Retrieval

- Lexical Retrieval
  - BM25
  - TF-IDF



Text Retrieval

- Dense Retrieval
  - DPR



Information Distance

$$\begin{aligned} E(x,y) &= \max\{K(x|y), K(y|x)\} \\ &= K(xy) - \min\{K(x), K(y)\} \end{aligned}$$

$$NCD(x,y) = \frac{C(xy) - \min\{C(x), C(y)\}}{\max\{C(x), C(y)\}}$$

The similarity metric. Ming Li et.al.

Information Distance

```
import gzip
def compute_kolmogorov_distance(query, text):
    C_query = len(gzip.compress(query.encode()))
    C_text = len(gzip.compress(text.encode()))
    C_query_text = len(gzip.compress((query + '' + text).encode()))
    ncd = (C_query_text - min(C_query, C_text)) / max(C_query , C_text)
    return ncd
```

Less is More: Parameter-Free Text Classification with Gzip
Zhiying Jiang et.al.

Dataset

- TREC DL19, DL20
  - Web Search Dataset
  - 8.8 Million Document Corpus
  - 0-3 Relevancy Score Judged
    - We consider score >=2 as relevant

Experiments

- Observe the accuracy that positive pairs has lower Kolmogorov distance than negative pairs NCD(positive) < NCD(negative)
- i.e. check how information distance align with query—document relevancy judgment.

- Q, P1, P2, …, N1, N2 …

Query – Document Distance

	DL19	DL20
Accuracy (min)	11.62%	12.96%
Accuracy (avg)	23.25%	33.33%

Why?

- Query – Document similarity is not symmetric
- There is gap in query/document distribution, makes its hard to measure information distance.

- Solution:
  - Hypothetical Document Generation
  - Casting query-document distance measure to document—document distance.

Hypothetical Document Generation

InstructGPT, Capture the user intent and generate a Hypothetical document.

1. Capture the query intent
2. Map query into document distribution

Hypothetical Document Generation

Example:

Question: where was Michael klim born? (ncd = 0.85)

Generated Doc: ...Michael Klim was born on August 12, 1976 in Melbourne, Australia. He is the son of Polish immigrants and was raised in the city's western suburbs. ….(ncd = 0.71)

Relevant Doc: Michael Klim was born in 1977 in Poland. He is married to Lindy Rama. They have one child...

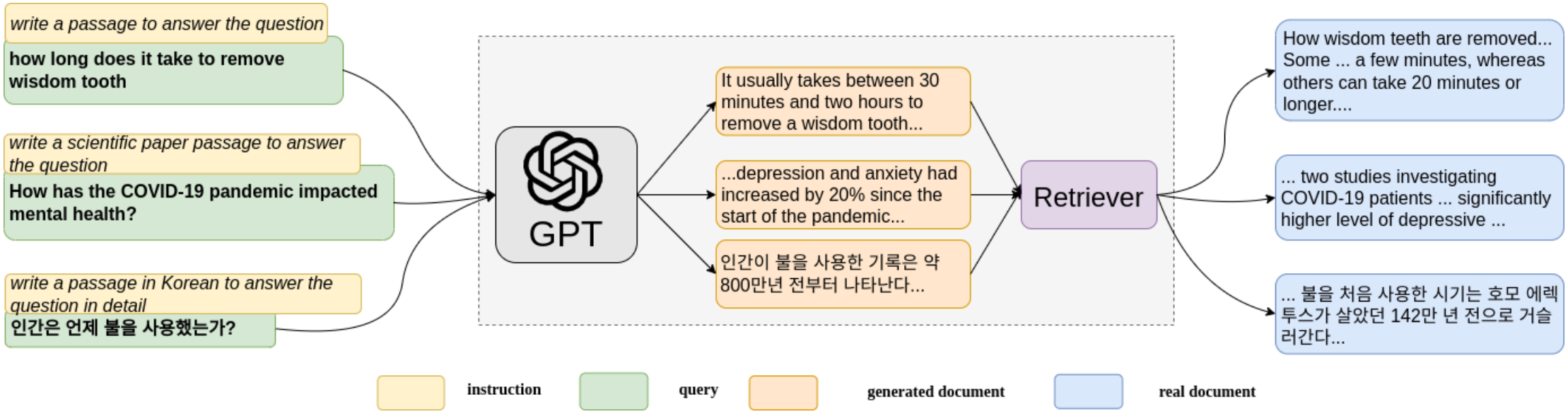
- **Q**, P1, P2, …, N1, N2 …
- Q -> **D**, P1, P2, …, N1, N2, ...

Pseudo Document– Document Distance

Query--Document	DL19	DL20
Accuracy (min)	11.62%	12.96%
Accuracy (avg)	23.25%	33.33%

Pseudo Document -- Document	DL19	DL20
Accuracy (min) (GPT3.5)	**65.12%**	**40.74%**
Accuracy (min) (Curie)	51.16%	37.04 %
Accuracy (avg) (GPT3.5)	**100%**	94.44%
Accuracy (avg) (Curie)	**100%**	**96.29%**

Retrieval with Hypothetical Document Generation



Retrieval with Hypothetical Document Generation

		DL19			DL20	
	map	Ndcg@10	Recall@1k	map	Ndcg@10	Recall@1k
BM25	30.1	50.6	75.0	28.6	48.0	78.6
Contriever	24.0	44.5	88.0	24.0	42.1	75.4
HyDE (BM25)	39.7	59.4	85.6	38.4	55.5	86.3
HyDE (Contriever)	41.8	61.3	88.0	38.2	57.9	84.4

Thank You

[1] Less is More: Parameter-Free Text Classification with Gzip

Zhiying Jiang, Matthew Y.R. Yang, Mikhail Tsirlin, Raphael Tang, and Jimmy Lin

[2] The similarity metric

Ming Li, Xin Chen, Xin Li, Bin Ma, and Paul MB Vitányi.

[3] Precise Zero-Shot Dense Retrieval without Relevance Labels

Luyu Gao*, Xueguang Ma*, Jimmy Lin, and Jamie Callan