INTERNATIONAL HELLENIC UNIVERSITY

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

A Combination of BERT and BM25 for Patent Search

Vasileios Stamatis[1], Michail Salampasis[1], Konstantinos Diamantaras[1], Allan Hanbury[2]

[1]International Hellenic University, Thessaloniki, Greece

[2]Vienna University of Technology, Vienna, Austria

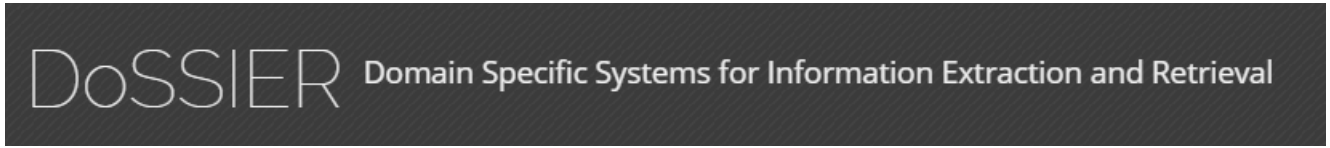
INTERNATIONAL HELLENIC UNIVERSITY

Project information

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

- This work is part of the DoSSIER project (https://dossier-project.eu/)
- DoSSIER is an EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval.



- This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 860721

INTERNATIONAL HELLENIC UNIVERSITY

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

Patent Search

Patent documents:

- Long documents.
- Specific structure.
- Many domain specific words.

Materials:

- Patent examiners use very old technologies.
- There are limited resources.
- There is a need to advance the search technologies used.

Models:

- Boolean Model.
- Probabilistic model. (BM25)
- Neural models. (BERT)

We want to combine lexical and semantics signals of relevance.

Research Question

- How can the BERT model be adapted to improve retrieval effectiveness in patent prior art search?

Method

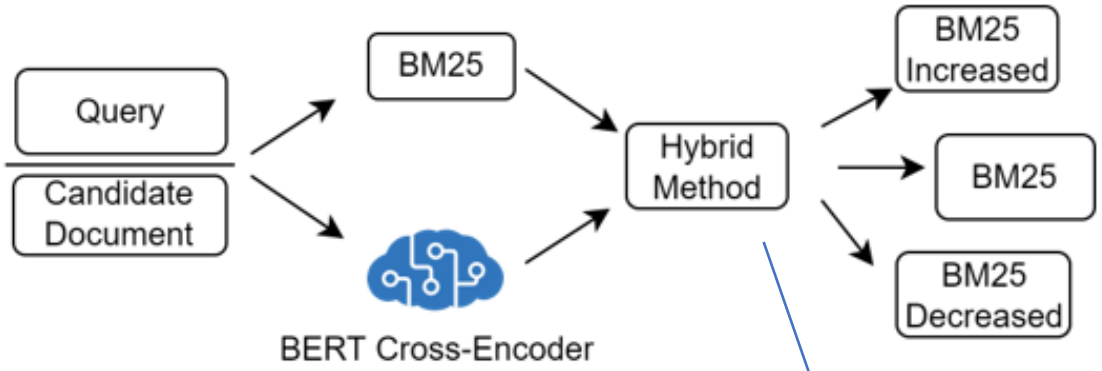


Figure 1: Hybrid Retrieval method

$$score = bm25 + c * bm25 * bert \ (1)$$

Method

- BM25 score was on average between (200, and 1000).
- BERT score was between (-3, and 3). We used only the patent abstracts as input for BERT.
- We optimized the parameter c in formula (1) and conducted a grid search and we found that formula (1) optimized for c = 0.25 using our dataset.

Method

For instance, if BERT estimates a candidate document as non-relevant with a low value i.e. -3, then the score would be:

$$score = bm25 - 3 * 0.25 * bm25$$

$$score = bm25 - \frac{3}{4} bm25$$

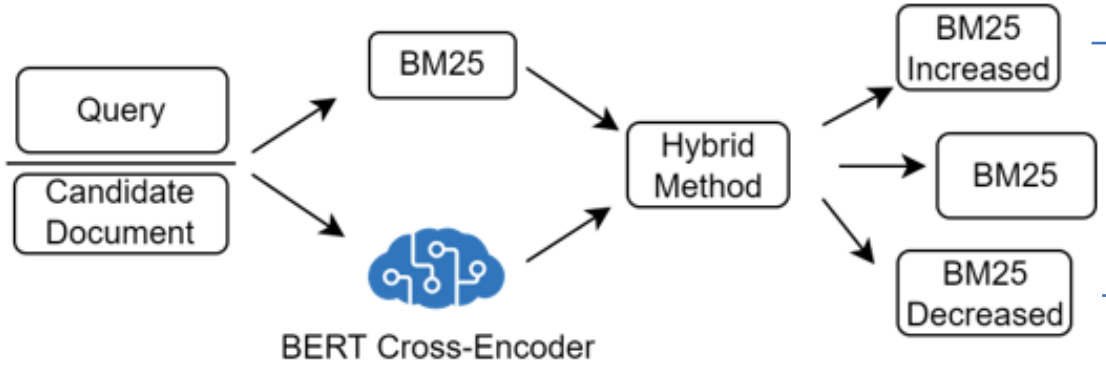
Method

Respectively, if BERT estimates a candidate document as relevant with a high value i.e. 3, then the score would be:

$$score = bm25 + 3 * 0.25 * bm25$$

$$score = bm25 + \frac{3}{4} bm25$$

Method



Max score: BM25 + ¾ BM25

(1.75* BM25)

MIN score: BM25 - ¾ BM25

(0.25* BM25)

Figure 1: Hybrid Retrieval method

INTERNATIONAL HELLENIC UNIVERSITY

Datasets

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

The experiments presented in this paper are based on:

- CLEF-IP 2011 collection.
- Our new IPA dataset.

Datasets

MSCA ITN/ETN No. 860721
DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

To create the IPA dataset:

1. We iterate all MAREC documents, and for each document with an English abstract, we process its citations.

2. For every citation, we extracted it's English abstract and wrote one relevant instance in the CSV file (abstract_doc | abstract_citation | 1)

3. And one non-relevant instance using the abstract from a random document from the MAREC collection (abstract_doc | abstract_random | 0).

4. Finally, we removed all the lines containing abstracts used in CLEF-IP topics.

5. The whole dataset contains approximately 78 million pairs of abstracts.

Results

MSCA ITN/ETN No. 860721
DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

Baselines:

- BM25.
- Cross-Encoder BERT (CE BERT).
- Bi-Encoder BERT (BE BERT).

- BERT used in a zero-shot setting and fine-tuned using IPA dataset.

INTERNATIONAL HELLENIC UNIVERSITY

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

Results

MODEL	MAP @100	PRES @100	RECALL @100
BM25	0.0881	0.2115	0.2761
CE BERT (zero-shot)	0.0005	0.0050	0.0090
CE BERT (fine-tuned)	0.0088	0.0877	0.1544
BE BERT (fine-tuned)	0.0114	0.0521	0.0916
BE BERT (zero-shot)	0.0226	0.0868	0.1242
Hybrid (zero-shot)	0.0006	0.0045	0.0069
Hybrid (fine-tuned)	0.0930	0.2191	0.2859
Random	0.0017	0.0188	0.0421

Table 1: Results of the different models

Our fine-tuned hybrid method achieved the best scores and outperformed all the baselines, especially the BM25 by 5.56% at MAP, 3.6% at PRES, and 3.5% at RECALL @100

INTERNATIONAL HELLENIC UNIVERSITY

MSCA ITN/ETN No. 860721

DoSSIER
Domain Specific Systems for Information Extraction and Retrieval

Next Steps

- Explore more complex ways to combine BERT and BM25.
  - Especially ways to include the whole patent document.
- Use other parts of patent documents as input for BERT.
- Include more datasets for the experiments.
- Add more baselines for comparison.