UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

SEUPD@CLEF Team FADERIC

A query expansion and reranking approach for the LongEval task

Enrico Bolzonello, Christian Marchiori,
Daniele Moschetta, Riccardo Trevisiol and Fabio Zanini

PROBLEM DESCRIPTION

Performance degradation over time

«LongEval differs from traditional IR and classification shared task with special considerations on evaluating models that **mitigate performance drop over time**»

Project Overview

1 Parser

2 Analyzer
- French Analyzer
- English Analyzer

3 Indexer

4 Searcher
- Query Expansion
- Reranker

5 Results

1 PARSER

ParsedDocument

- Represents a parsed document to be indexed
- Stores ID and Body

DocumentParser

- Basic functionalities to iterate over the elements of a ParsedDocument

LongEvalDocumentParser

- Specific functionalities for documents in the TREC format
- **Input**: TREC format document
- **Output**: ParsedDocument

2 | ANALYZER

Used to process texts from documents and queries

French and English analyzers have been implemented

Common features:

01 Tokenizing

02 Character folding

03 Position Filtering

04 Stopword removal

Specific analyzers features

French analyzer

- Elision removal
- **Stoplist**:
  - Lucene's standard for French
  - Custom (most freq. terms)
- **Stemming**:
  - French Snowball
  - Light

English analyzer

- Possessive removal
- **Stoplist**:
  - Lucene's standard for English
  - Custom (most freq. terms)
- **Stemming**:
  - English Snowball (Porter2)
  - Light

3 | INDEXER

Used to create a searchable database (index) for parsed documents

DirectoryIndexer

- Index all documents located in a certain directory

BodyField

- Represents the body of a document in the index
- Term frequencies and positions are stored
- It is **tokenized**
- It is **stored** (needed for the rerank)

4 | SEARCHER

Approaches used to improve the searcher:

- Tuning BM25 parameters
- Query expansion
  - Fuzzy search
  - N-grams (with proximity search)
  - Synonyms

The searcher implements the **boolean query** procedure, the boolean clauses added are the original query and the expanded query.

Tuning BM25 parameters

Run		FADERIC_French-Stop50-Stem-Shingle-Fuzzy							
Measure		nDCG							
		k1							
		0.6	0.8	1.0	1.2	1.4	1.6	1.8	2.0
b	0.30	0.3941	0.3952	0.3954	0.3957	0.3963	0.3936	0.3948	0.3934
	0.40	0.3967	0.398	0.3984	0.3992	0.3992	0.3992	0.3981	0.3975
	0.50	0.3999	0.4004	0.4008	0.4013	0.4014	0.4014	0.4014	0.4011
	0.60	0.3999	0.4013	0.4017	0.4025	0.4026	0.4024	0.4019	0.4008
	0.70	0.4021	0.4029	0.4034	0.4038	0.4043	**0.4047**	0.4046	0.4039
	0.75	0.4018	0.4028	0.4035	0.4039	0.4038	0.4043	0.4037	0.4035
	0.80	0.401	0.4025	0.4033	0.404	0.4041	0.4043	**0.4047**	0.4039
	0.90	0.3985	0.3998	0.4009	0.4014	0.4018	0.4021	0.4015	0.4011

The default BM25 values in Lucene are

- k1 = 1.2
- b = 0.75

The best performing combination improves performances by 0,223%

Query expansion: Fuzzy

What it is: Allows you to find results even when the words you search for do not exactly match those in the documents.

- We applied fuzzy only if the query contains a **single term**.
- The **fuzzy parameter** can specify the max number of edits allowed in the word. The value is between 0 and 2.
- In our system if the word length is ≥ 10 then the fuzzy parameter is set to 2, otherwise 1 is used.

Query expansion: Word N-grams

What it is: Is a sentence analysis technique of dividing the words of a sentence into overlapping sequences of n consecutive words.

- We avoided generating **unigrams**.
- We decided to generate n-grams with a maximum of **3 words**.
- We then decided to set up a **proximity search** within each n-gram, with a proximity parameter set to 5.
- We applied a **boost** to all n-grams based on the size of the n-gram itself.

Università degli Studi di Padova | Dipartimento di Ingegneria dell'Informazione

Query expansion: Synonyms

Two approaches used:

- **SynonymAnalyzer** for both English and French synonyms.
  - Standard/Custom synonyms dictionary
  - *SynonymGraphFilter* and *FlattenGraphFilter*
  - Boost (based on the size of the processed query)
- **SynonymPOSAnalyzer** only for English synonyms.
  - WordNet dictionary
  - OpenNLP POS Tagging
  - Boost (based on the size of the processed query)

OpenNLP Tag	WordNet Section
JJ	Adjectives
VB	Verbs
RB	Adverbs
NN	Nouns
Others	No synonyms retrieved

Università degli Studi di Padova | Dipartimento di Ingegneria dell'Informazione

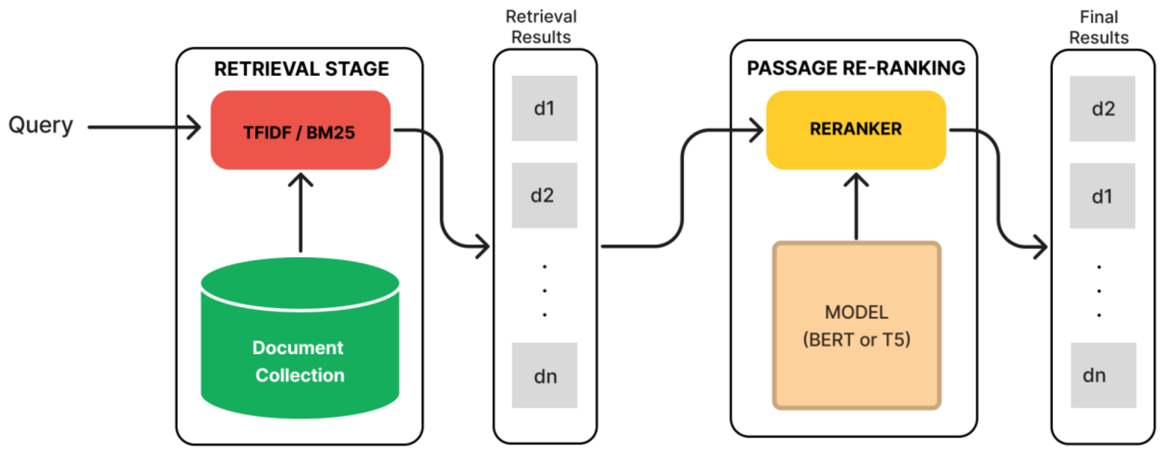
5 | RERANKER

01
Retrieval Stage

TDIDF/BM25 retrieval to get candidates for the ranking. Really **fast**, but not the best approach to rank

02
Passage Re-Ranking

Using Machine Learning, it ranks the documents retrieved by the first stage. Computationally heavy, so really **slow**



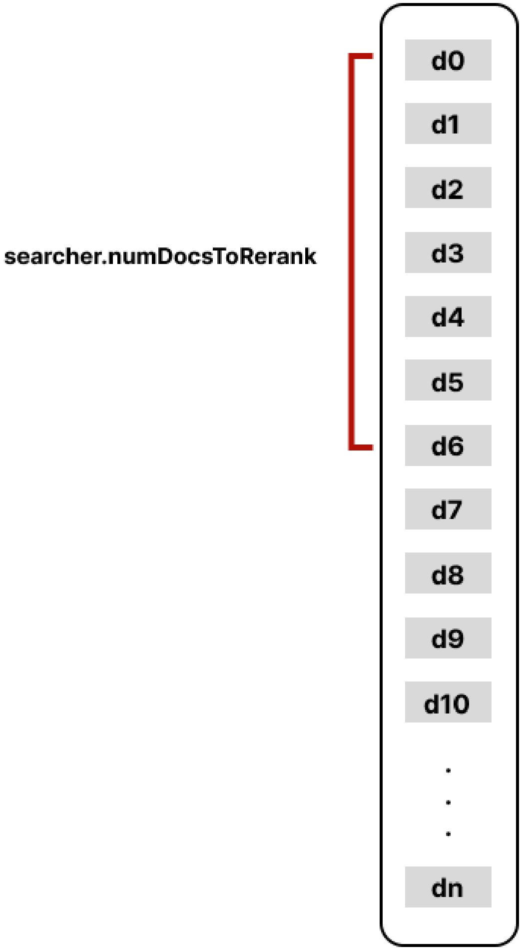
Retrieve-than-Rerank framework

Reranker Features

PyGaggle
Python library which provides deep neural architectures for text ranking and question answering

Number of docs to Rerank
Within the config.xml file, there exists an option that allows for the selection of the desired number of documents to be reranked.

Weight
To incorporate BM25 scores into consideration, a weight to the reranker score can be assigned



Assigning **Weights** to Scores

Normalized Score:

Since the score given by the reranker is between -10 and 10, a normalization is needed

$$nScore_{rr}(i) = \left(Score_{rr}(i) + \min_{j \in [1,n]} Score(j) \right) \cdot \frac{Score_{BM25}(1)}{Score_{rr}(1)}$$

$$finalScore(i) = mntr + (1 - \alpha) \cdot Score_{BM25}(i) + \alpha \cdot nScore_{rr}(i)$$

Maximum score from docs **not reranked**

Weight

In config.xml, *searcher.rerankScoreWeight*

Università degli Studi di Padova | Dipartimento di Ingegneria dell'Informazione

Models

	monot5	bert	own trained
0	0,4075	0,4075	0,4075
10	0,414	0,4207	0,3910
20	0,4119	0,4222	0,3741
50	0,4083	0,4212	0,3405
100	0,405	0,4184	-
250	0,3987	0,4104	-

Three models:

- monot5-base-msmarco-10k
- bert-base-mdoc-bm25
- own trained checkpoint
  based on bert-base-uncased

Università degli Studi di Padova
Dipartimento di Ingegneria dell'Informazione

6 | RESULTS

Document Collections
Results have been analyzed both on training and test data

Languages
Performances have been evaluated for both english and french document collection

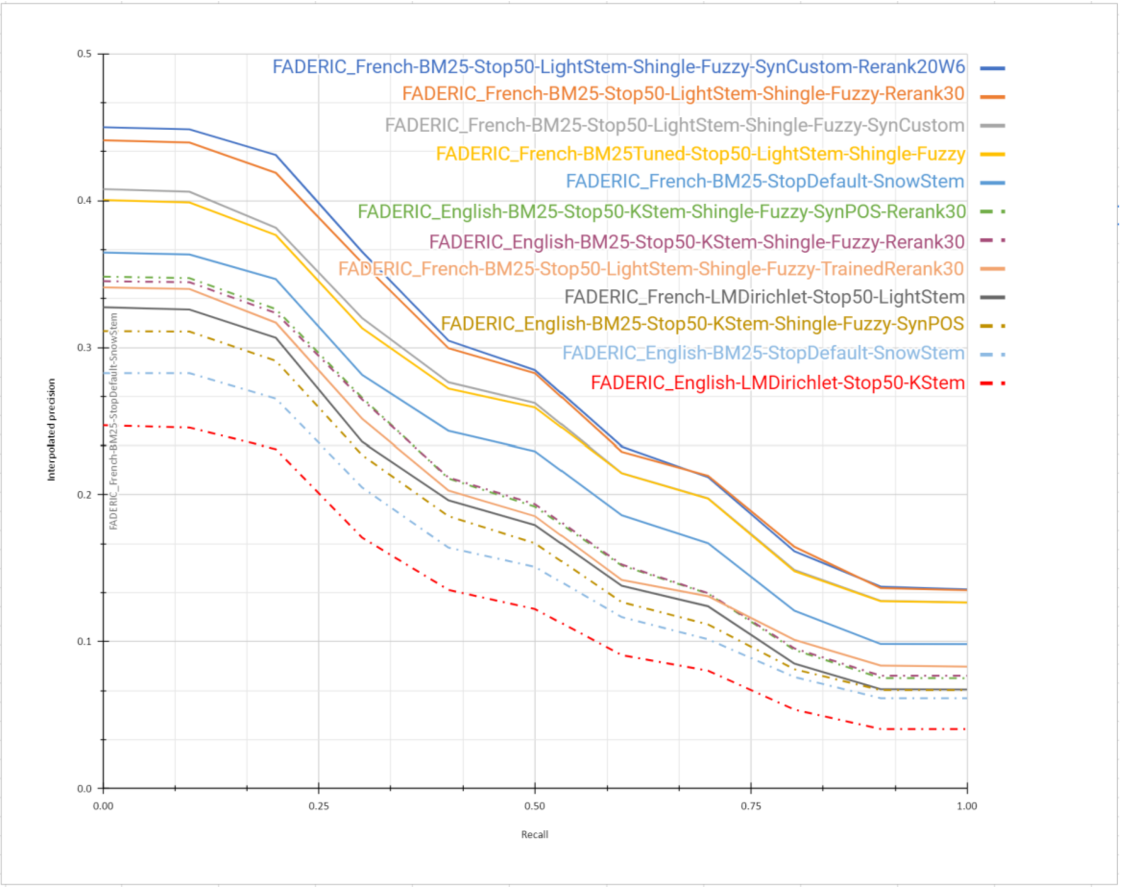
Measures
nDCG, MAP and interpolated precision-recall curve

Run names
Run names are indicative of the components used in the run

Results: Training data

Run name	nDCG	MAP
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom-Rerank20W6	0.4274	0.2671
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-Rerank30	0.4230	0.2632
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom	0.4079	0.2416
FADERIC_French-BM25Tuned-Stop50-LightStem-Shingle-Fuzzy	0.4047	0.2383
FADERIC_French-BM25-StopDefault-SnowStem	0.3786	0.2110
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-SynPOS-Rerank30	0.3271	0.1877
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-Rerank30	0.3527	0.1873
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-TrainedRerank30	0.3599	0.1799
FADERIC_French-LMDirichlet-Stop50-LightStem	0.3398	0.1731
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-SynPOS	0.3081	0.1634
FADERIC_English-BM25-StopDefault-SnowStem	0.2927	0.1490
FADERIC_English-LMDirichlet-Stop50-KStem	0.2612	0.1228

Results: Training data

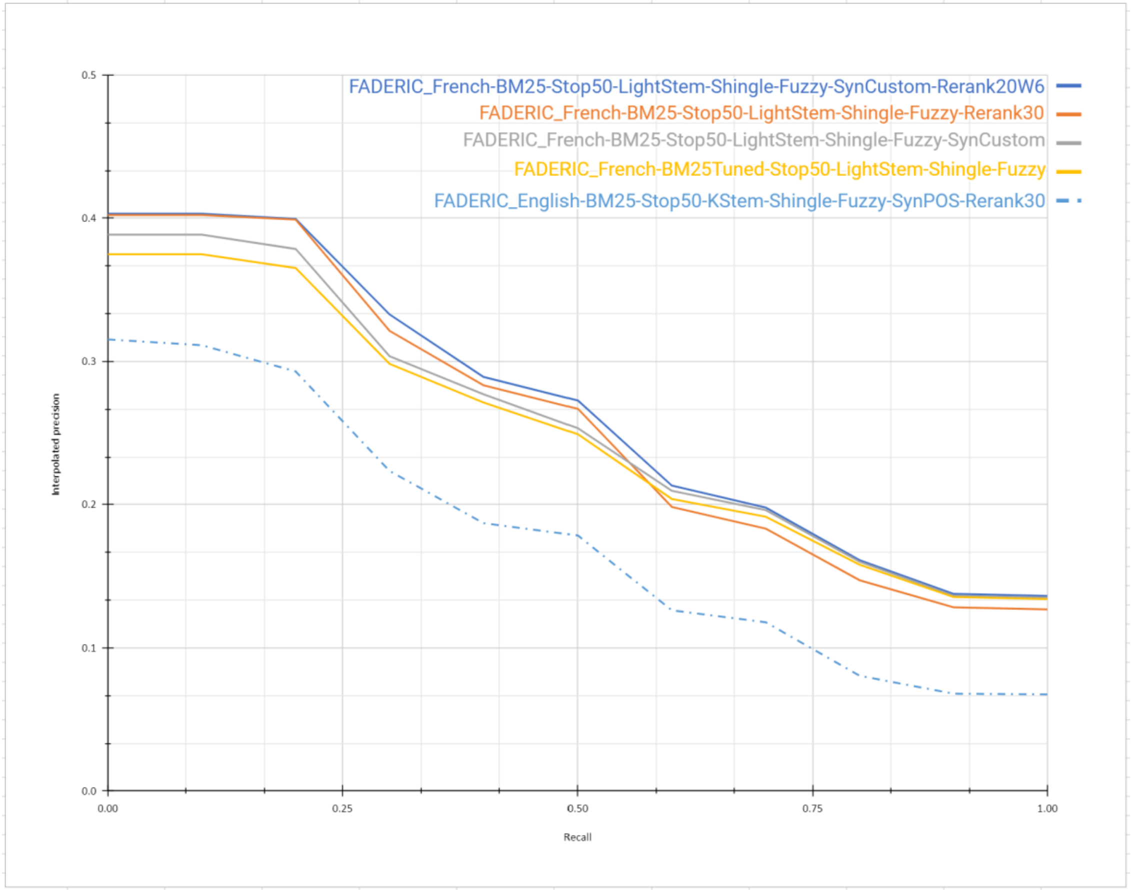


Università degli Studi di Padova | Dipartimento di Ingegneria dell'Informazione

Results: Heldout

Run name	nDCG	MAP
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom-Rerank20W6	0.4169	0.2474
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-Rerank30	0.4147	0.2416
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom	0.4080	0.2376
FADERIC_French-BM25Tuned-Stop50-LightStem-Shingle-Fuzzy	0.4044	0.2324
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-SynPOS-Rerank30	0.3030	0.1626

Results: Heldout

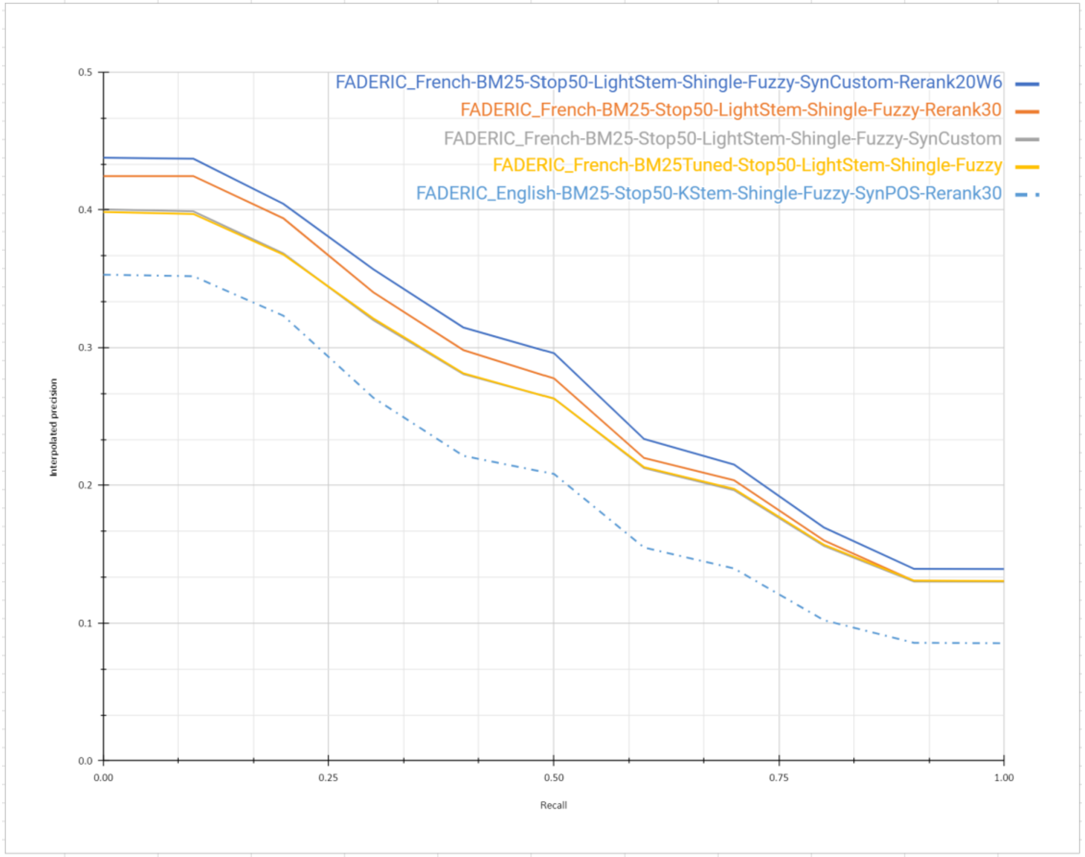


Università degli Studi di Padova | Dipartimento di Ingegneria dell'Informazione

Results: Short term

Run name	nDCG	MAP
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom-Rerank20W6	0.4239	0.2665
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-Rerank30	0.4145	0.2546
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom	0.4034	0.2412
FADERIC_French-BM25Tuned-Stop50-LightStem-Shingle-Fuzzy	0.4034	0.2414
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-SynPOS-Rerank30	0.3296	0.1931

Results: Short term

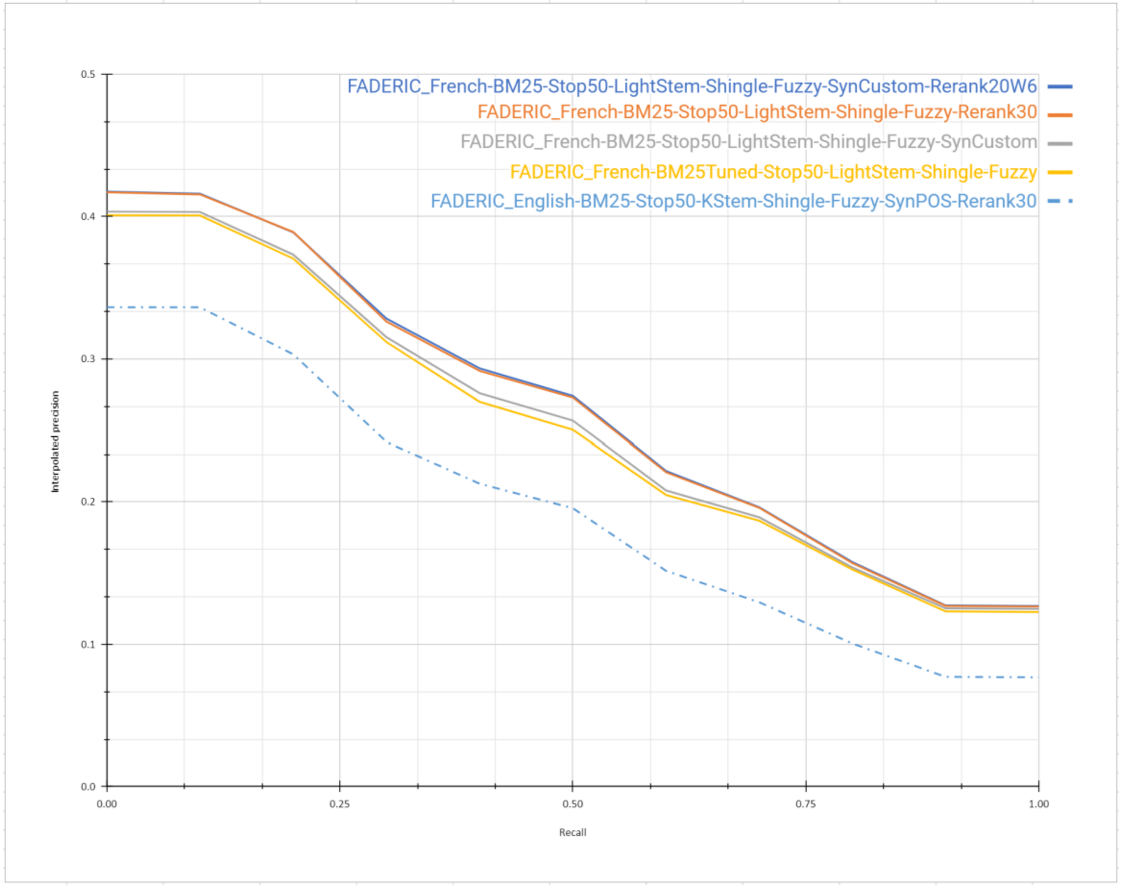


UNIVERSITÀ DEGLI STUDI DI PADOVA | DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

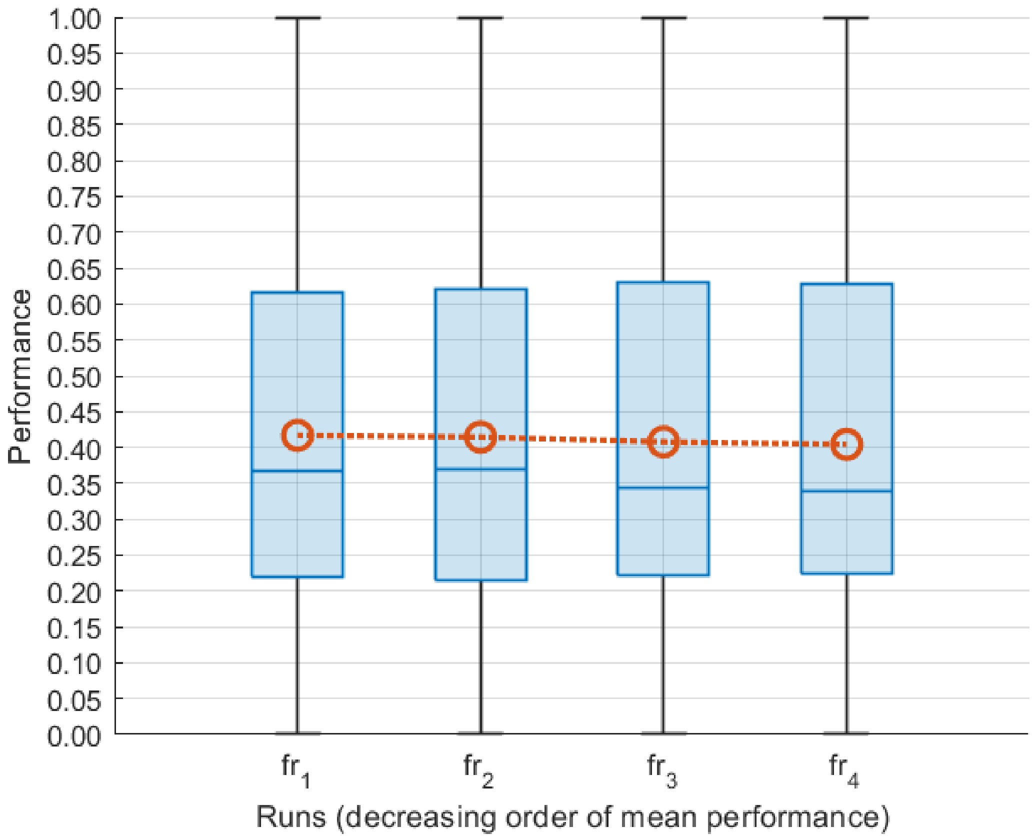
Results: Long term

Run name	nDCG	MAP
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom-Rerank20W6	0.4153	0.2473
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-Rerank30	0.4146	0.2465
FADERIC_French-BM25-Stop50-LightStem-Shingle-Fuzzy-SynCustom	0.4091	0.2384
FADERIC_French-BM25Tuned-Stop50-LightStem-Shingle-Fuzzy	0.4071	0.2350
FADERIC_English-BM25-Stop50-KStem-Shingle-Fuzzy-SynPOS-Rerank30	0.3296	0.1809

Results: Long term



Statistical analysis: Heldout (nDCG)

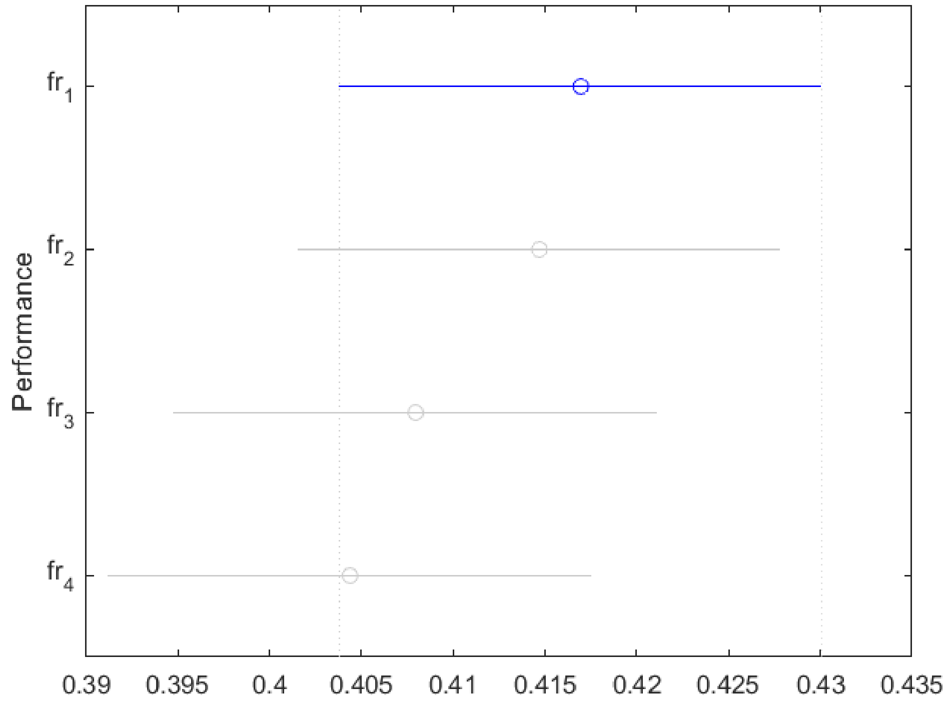


Boxplot on heldout collection (nDCG)

Statistical analysis: Heldout (nDCG)

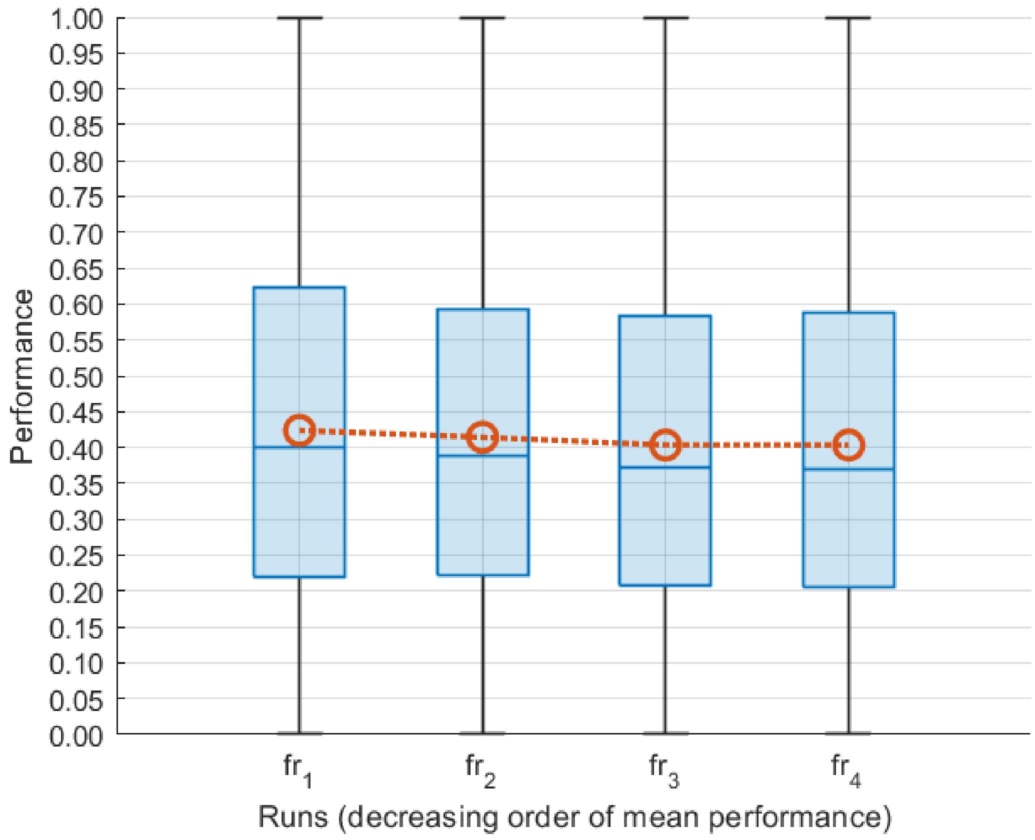
Source	SS	df	MS	F	Prob>F
Columns	0.01	3	0.003	0.64	0.58
Rows	23.54	97	0.242	47.20	1.97E-134
Error	1.49	291	0.005	-	-
Total	25.04	391	-	-	-

ANOVA2 on heldout collection (nDCG) with alpha 0.05



Tukey's HSD on heldout collection (nDCG)

Statistical analysis: Short term (nDCG)

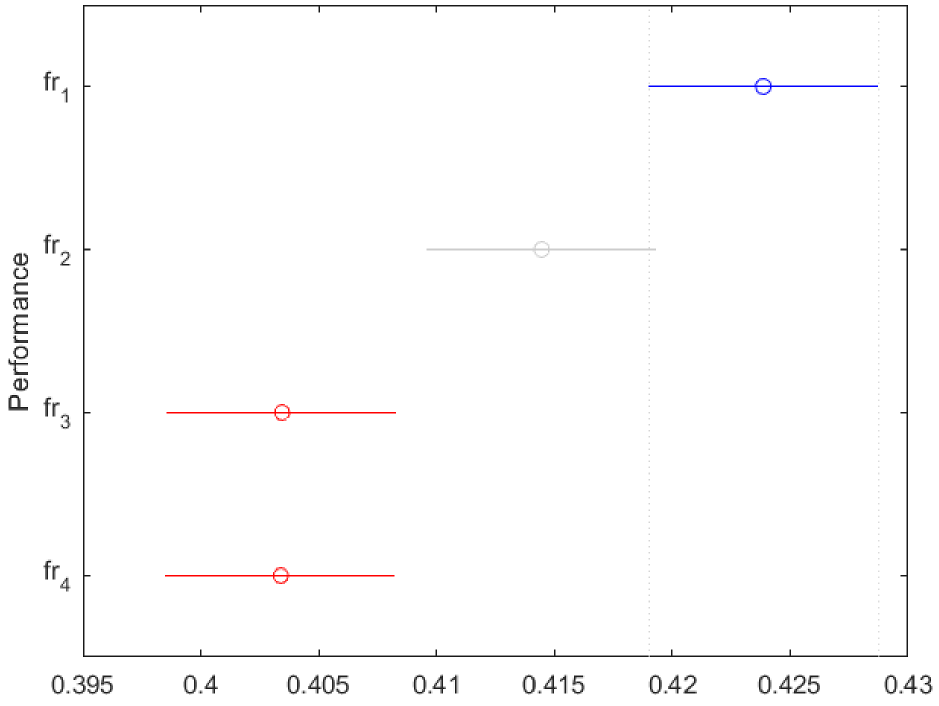


Boxplot on short term collection (nDCG)

Statistical analysis: Short term (nDCG)

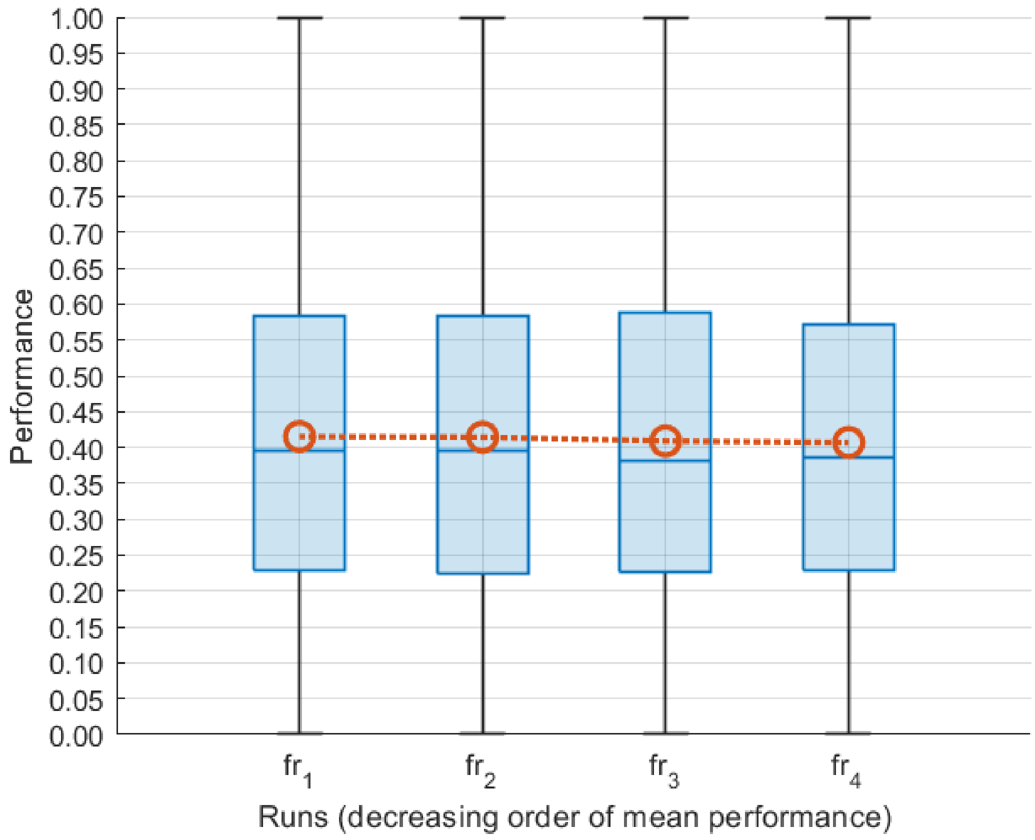
Source	SS	df	MS	F	Prob>F
Columns	0.25	3	0.085	13.58	8.51E-9
Rows	218.58	881	0.248	39.30	0
Error	16.68	2643	0.006	-	-
Total	235.52	3527	-	-	-

ANOVA2 on short term collection (nDCG) with alpha 0.05



Tukey's HSD on short term collection (nDCG)

Statistical analysis: Long term (nDCG)

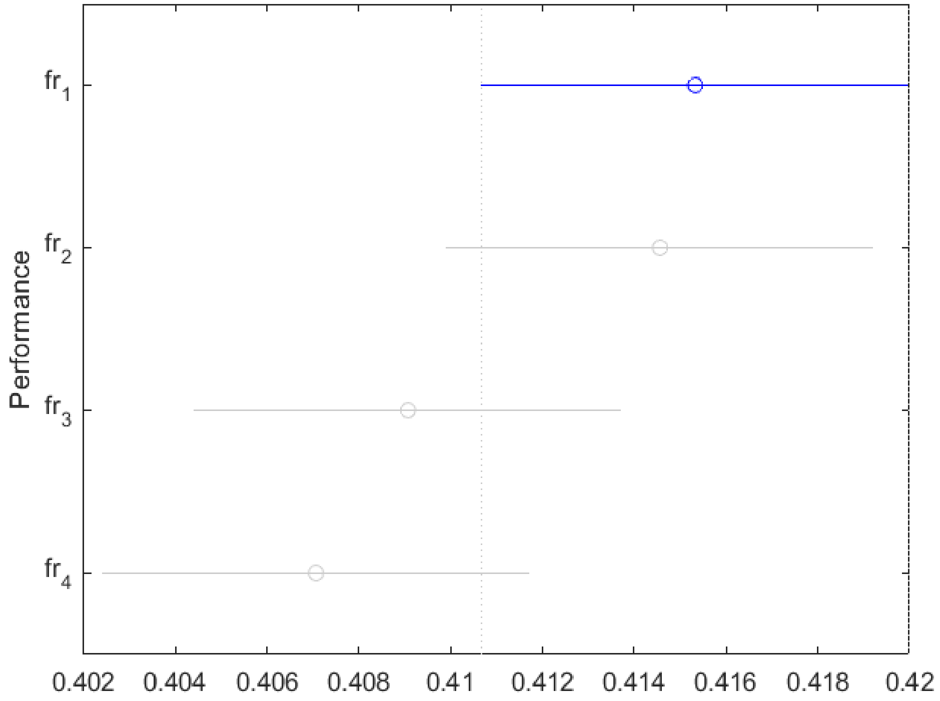


Boxplot on long term collection (nDCG)

Statistical analysis: Long term (nDCG)

Source	SS	df	MS	F	Prob>F
Columns	0.04	3	0.015	1.51	0.056
Rows	202.35	922	0.219	16.17	0
Error	16.78	2766	0.006	-	-
Total	219.18	3691	-	-	-

ANOVA2 on long term collection (nDCG) with alpha 0.05



Tukey's HSD on long term collection (nDCG)

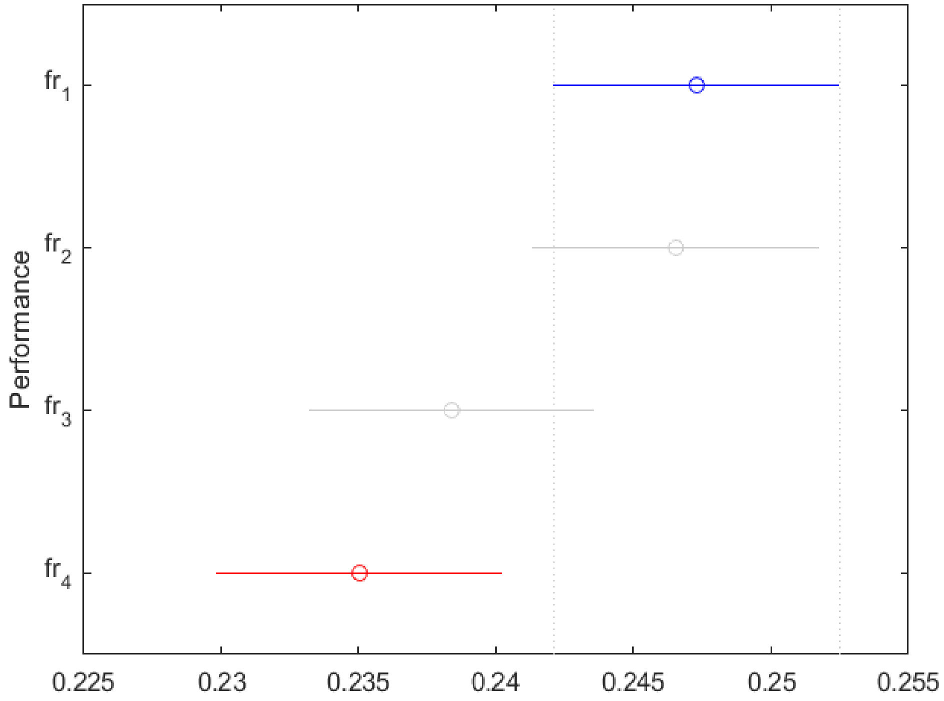
UNIVERSITÀ DEGLI STUDI DI PADOVA

DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

Statistical analysis: Long term (AP)

Source	SS	df	MS	F	Prob>F
Columns	0.10	3	0.011	4.47	0.003
Rows	188.61	922	0.204	27.03	0
Error	20.92	2766	0.007	-	-
Total	209.64	3691	-	-	-

ANOVA2 on long term collection (AP) with alpha 0.05



Tukey's HSD on long term collection (AP)

CONCLUSIONS
AND FUTURE WORKS

01 The system kept **satisfactory performance** on both short and long term

02 **Query expansion** and **reranking** played a major role in overall system performances

03 Possible improvements:

- Better synonym dictionaries
- Query expansion with Neural Networks (NN)
- Better train on our model

UNIVERSITÀ DEGLI STUDI DI PADOVA | DIPARTIMENTO DI INGEGNERIA DELL'INFORMAZIONE

THANK YOU!

Team FADERIC

Università degli Studi di Padova

Dipartimento di Ingegneria dell'Informazione