Enhancing Search Engine Performance on the CLEF 2023 LongEval Corpus with Character N-Grams, Query Expansion, and Named Entity Recognition

TA$K: LongEval CLEF 2023 Lab

Team JIHUMING@UNIPD

Jesús Moncada-Ramírez, Isil Atabek, Huimin Chen

Nicolò Santini, Giovanni Zago

Agenda |

- Introduction
- Methodology
- System Architecture
- Experimental Setup
- Results and Discussion
- Conclusion

Our Team |

Jesús Moncada-Ramírez

Giovanni Zago

Huimin Chen

Nicolò Santini

Isil Atabek

Introduction |

We introduce a search engine for LongEval at CLEF 2023. Our system focuses on temporal performance in English and French documents.

By analyzing text and using NLP techniques, we refine our system. Implemented in Java with Lucene, we developed five top-performing systems based on MAP and NDCG scores.

Introduction |

Our approach involves analyzing English and French versions of the documents using whitespace tokenization, stopword removal and stemming.

We generate character N-grams to identify recurring word structures repeated over documents.

We use query expansion with synonyms and NLP techniques as NER to further refine our system.

Methodology | Parsing

Documents

- JSON version.
- *Iterator/Iterable* architecture.

Topics

- *TRECTopicReader* not working.
- Our **own** parser treating topic documents as **XML** documents.

Methodology | Index

Always (BM25)

FIELD 1	FIELD 2	FIELD 3	FIELD 4
(processed) English version	(processed) French version	Character **N-grams** of both versions	**NER** information
		*3-grams, 4-grams, 5-grams*	*Apache OpenNLP*

Methodology | English

ENGLISH PROCESSING (ANALYZER)

Whitespace tokenization → Breaking based on special characters → Lowercasing

TERRIER stopword list → Query expansion with synonyms → Stemming

Methodology | French

FRENCH PROCESSING (ANALYZER)

Whitespace tokenization → Breaking based on special characters → Lowercasing

French stopword list → Stemming

Methodology | N-grams and NER

N-GRAM GENERATION (ANALYZER)

- Delete all characters **except letters.**
- Generate char N-GRAMs.

NER GENERATION (ANALYZER)

- Use the **FRENCH** documents.
- NER about locations, person names and organizations.

System Architecture |

- LongEval data structure (queries and documents), **Document Parser** and **Topic Reader**
- **Analysis** techniques (tokenization, NER, N-Gram)
- **Index**
- **Search** engine

Experimental Setup | Overview

Goals:

1. Generate multilingual indexes
2. Perform runs over generated indexes and compare results

Requirements:

1. Java JDK version 17, Apache version 2, Lucene version 9.5, and Maven.
2. Source repository on Bitbucket
3. MAP and NDCG scores

Experimental Setup | Indexes

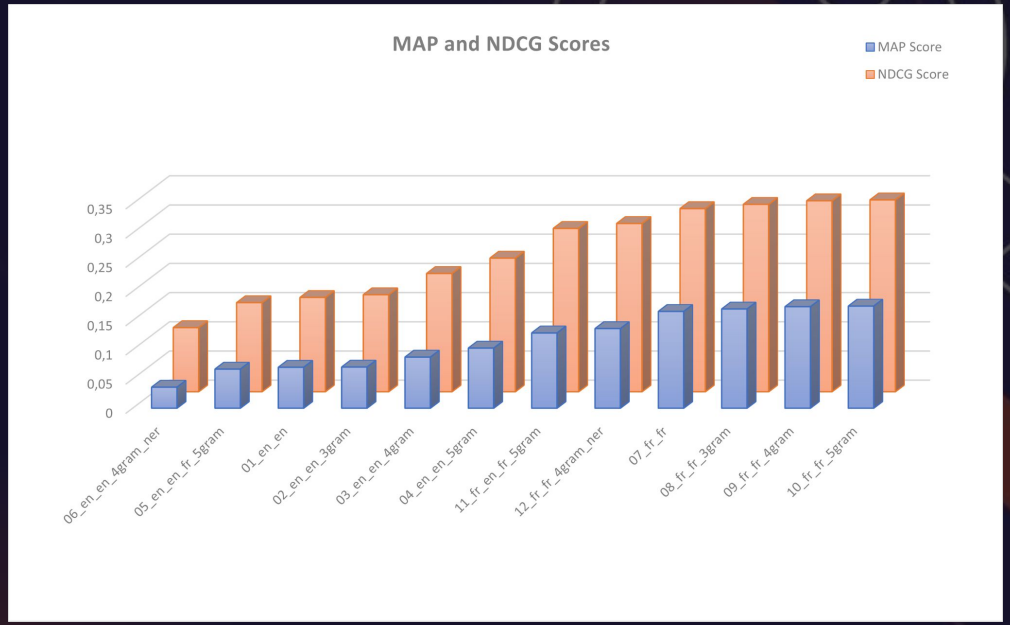
- **2023_04_24_multilingual_3gram**
  - English and French
  - Character 3-gram
- **2023_04_29_multilingual_3gram_synonym**
  - English and French
  - Character 3-gram
  - English query expansion with synonyms
- **2023_05_01_multilingual_4gram_synonym**
  - English and French
  - Character 4-gram
  - English query expansion with synonyms
- **2023_05_01_multilingual_5gram_synonym**
  - English and French
  - Character 5-gram
  - English query expansion with synonyms
- **2023_05_05_multilingual_4gram_synonym_ner**
  - English and French
  - Character 4 gram
  - English query expansion with synonyms
  - NER techniques

Experimental Setup | Runs

Making a lot of runs over different configurations, permit to analyze different aspects of the system and evaluate its effectiveness.

- seupd2223-JIHUMING-01_en_en
- seupd2223-JIHUMING-02_en_en_3gram
- seupd2223-JIHUMING-03_en_en_4gram
- seupd2223-JIHUMING-04_en_en_5gram
- seupd2223-JIHUMING-05_en_en_fr_5gram
- seupd2223-JIHUMING-06_en_en_4gram_ner
- seupd2223-JIHUMING-07_fr_fr
- seupd2223-JIHUMING-08_fr_fr_3gram
- seupd2223-JIHUMING-09_fr_fr_4gram
- seupd2223-JIHUMING-10_fr_fr_5gram
- seupd2223-JIHUMING-11_fr_en_fr_5gram
- seupd2223-JIHUMING-12_fr_fr_4gram_ner

Results and Discussion | Train

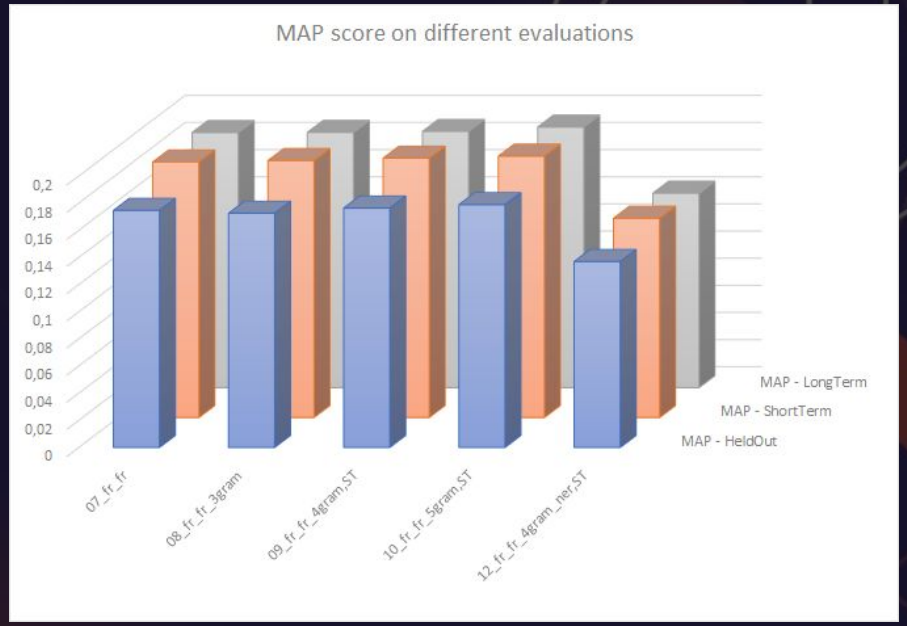


- French queries perform better than their English counterparts
- IR system's effectiveness generally increases with a larger N-gram size.
- The inclusion of NER in the indexing process has a negative impact on the scores

Results and Discussion | Train

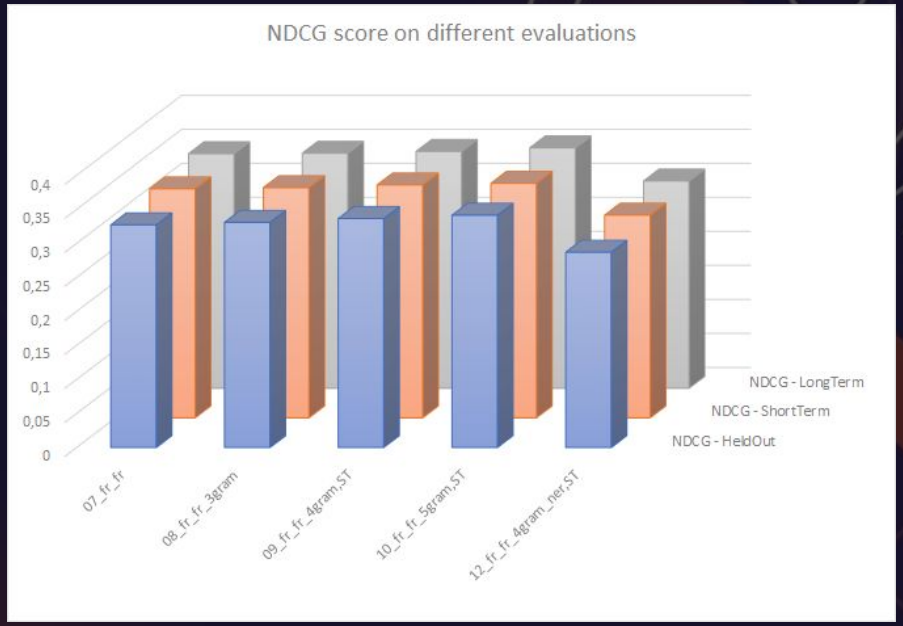
- Five best systems with five best scores
  - Fr_fr_5gram
  - Fr_fr_4gram
  - Fr_fr_3gram
  - Fr_fr
  - Fr_fr_4gram_ner

Results and Discussion | Test



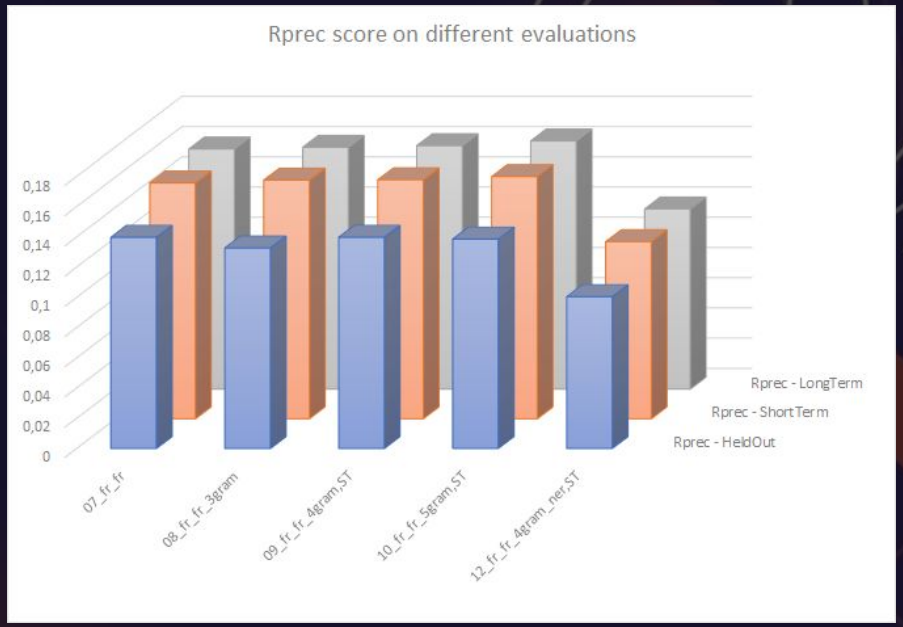
- Mean Average Precision (MAP) evaluating the effectiveness of an IR system in ranking documents/items.
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion | Test



- nDCG (normalized Discounted Cumulative Gain) assesses the quality of the ranking produced by an IR system.
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion | Test

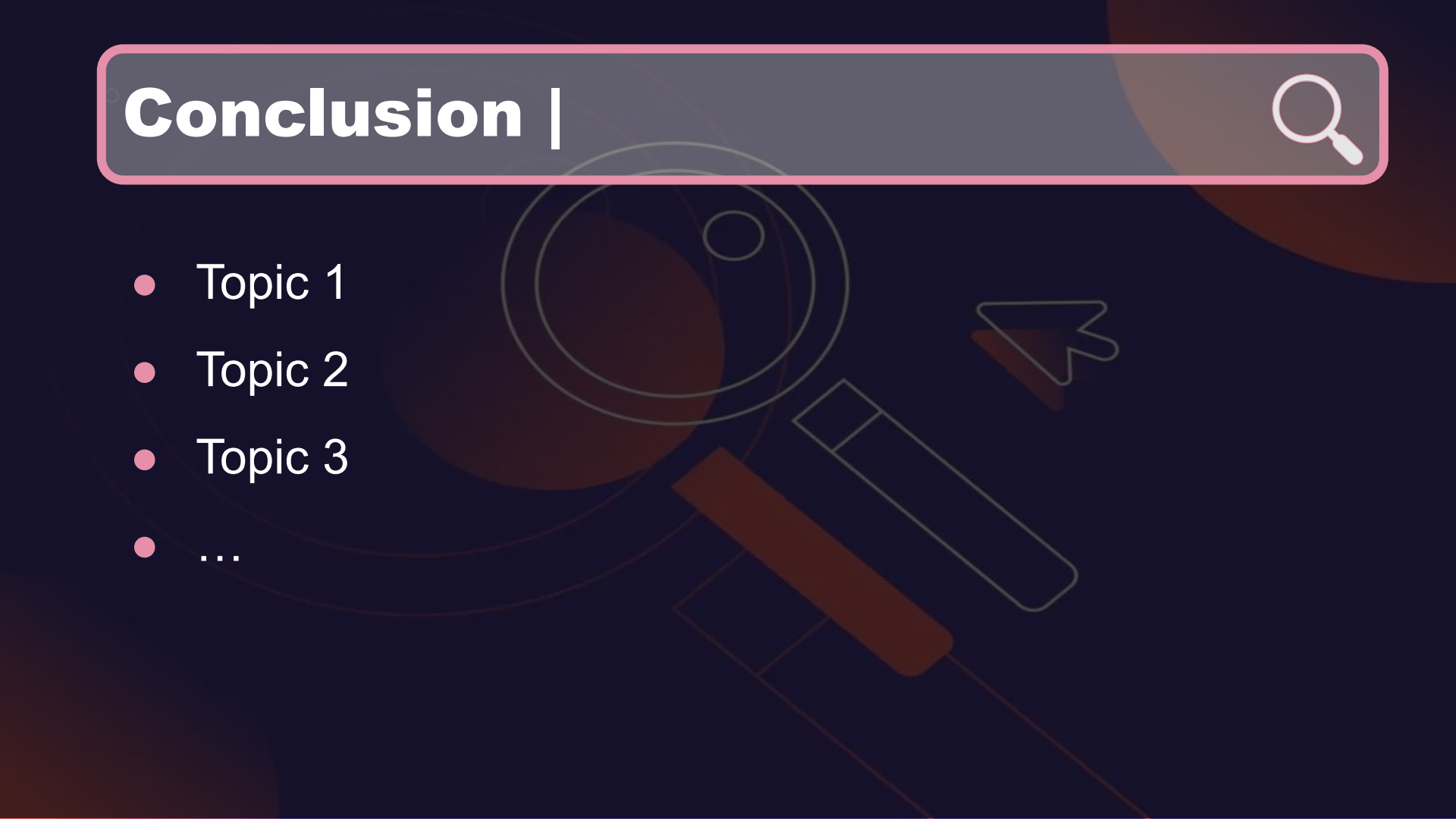


- Rprec (Rank Precision) measures the precision of the retrieved documents/items
- Indicating same score ranks in three data sets.
- Demonstrate the best performance at the long-term.

Results and Discussion |

- From the training data and test data, the three metrics shows the same rankings of five systems' effectiveness
- The long-term data presents the best performance score.

Conclusion |



- Topic 1
- Topic 2
- Topic 3
- …

THANK YOU

Team JIHUMING@UNIPD

Isil Atabek

Huimin Chen

Jesús Moncada-Ramírez

Nicolò Santini

Giovanni Zago