SEUPD@CLEF: Team <Acronym> on <Short Description>

Notebook for the LongEval Lab at CLEF 2023

Jesús Moncada-Ramírez[1], Isil Atabek[1], Huimin Chen[1], Michele Canale[1], Nicolò Santini[1] and Giovanni Zago[1]

[1]*University of Padua, Italy*

Abstract

A clear and well-documented LaTeX document is presented as an article formatted for publication by CEUR-WS in a conference proceedings. Based on the "ceurart" document class, this article presents and explains many of the common variations, as well as many of the formatting elements an author may use in the preparation of the documentation of their work.

Keywords

CLEF 2023, Information retrieval, LongEval, English, French, Search Engines

1. Introduction

This report aims at providing a brief explanation of the Information Retrieval system built as a team project during the Search Engine course 22/23 of the master's degree in Computer Engineering and Data Science at the University of Padua, Italy. As a group in this subject, we are participating in the 2023 CLEF LongEval: Longitudinal Evaluation of Model Performance [1]. This annual evaluation campaign focuses on the longitudinal evaluation of model performance in information retrieval and natural language processing.

The LongEval collection [2] relies on a large set of data provided by Qwant (a commercial privacy-focused search engine that was launched in France in 2013). Their idea regarding the dataset (collected in June 2022) was to reflect changes of the Web across time, providing evolving document and query sets. The training collection consists of 672 **queries**, 98 held-out queries, and 9656 evaluation assignments. The **documents** were chosen based on queries using the Qwant click model, in addition to random selection from the Qwant index. The training queries are categorized into twenty **topics**, such as: car-related, antivirus-related, employment-related, energy-related, recipe-related, etc. In addition to the original French version, the collection also includes English translations of the documents and queries using the CUBBITT [3] system.

CLEF 2023: Conference and Labs of the Evaluation Forum, September 18–21, 2023, Thessaloniki, Greece

jesus.moncadaramirez@studenti.unipd.it (J. Moncada-Ramírez); isil.atabek@studenti.unipd.it (I. Atabek); huimin.chen@studenti.unipd.it (H. Chen); michele.canale.1@studenti.unipd.it (M. Canale); nicolo.santini.1@studenti.unipd.it (N. Santini); giovanni.zago.3@studenti.unipd.it (G. Zago)

CEUR Workshop Proceedings (CEUR-WS.org)

The paper is organized as follows: Section 2 introduces related works; Section 3 describes our approach; Section 4 explains our experimental setup; Section 5 discusses our main findings; finally, Section 6 draws some conclusions and outlooks for future work.

2. Related Work

Describe related works, i.e. previous approaches to solve your problem you have started or improved from.

3. Methodology

Our final search engine can be divided into the following parts: parsing of the documents and queries, indexing, text processing (analyzers), and run generation (effective search).

Document parsing was performed using the JSON version of the documents. On the other hand, query parsing was based on an XML parser.

In the index, we decided to include four fields: (1) the (processed) English version of the documents, (2) the (processed)French version, (3) character N-grams of both versions concatenated, and (4) some NER information extracted from the French (original) version. As similarity function we have used BM25 [4] as it takes into account both term frequency and document length.

The text added to the fields must first be processed, for this we have developed four different analyzers. The English analyzer is based on whitespace tokenization, breaking of words and numbers based on special characters, lowercasing, applying the Terrier [5] stopword list, query expansion with synonyms based on the WordNet synonym map [6], and stemming. The French analyzer is based on whitespace tokenization, breaking of words and numbers based on special characters, lowercasing, applying a French stopword list [7] and stemming. To generate the character Ngrams we consider only the letters of the documents (i.e. we discard numbers and punctuation). To perform NER we apply NLP techniques based on Apache OpenNLP [8]) to the original (French) version of the documents. Specifically, we used NER applied to locations, person names and organizations.

We conducted some experiments to generate the runs, i.e., we have tried different combinations of the explained techniques. Thus, our searcher will always use BM25 [4], but the rest of characteristics depend on the run it is generating. See Section 4 for more details.

4. Experimental Setup

Our work was initiated based on the experimental setups outlined below.

- Evaluation measures: MAP (Mean Average Precision) and NDCG (Normalized Discounted Cumulative Gain) scores.

- [9, Repository].
- During the development and the experimentation, personal computers were used.
- Java JDK version 17, Apache version 2, Lucene version 9.5, and Maven.

In order to do different run experiments our team has created several indexes from the provided collection during the development of the final version of the project. In other words, the first created indexes only include several characteristics explained in this report, while the last indexes correspond to the final version of the project.
All the created indexes are **multilingual**, which allows us to take full advantage of the (bilingual) data collection. Additionally, we did some experiments with character N-grams generating different versions of indexes with 3-grams, 4-grams and 5-grams. Our motivation for experimenting with this was to compare how the size of different character N-grams affect to the effectiveness of our system. 3-grams are able to collecting more specific information about our documents, while 4-grams and 5-grams are more open to the context. An additional functionality of some indexes is query expansion, but as commented, this is only applied to the English body. One index uses Named Entity Recognition which provides not only the search for keywords but also identifying and extracting specific named entities.
In order to do different run experiments our team has created several indexes from the provided collection during the development of the final version of the project. In other words, the first created indexes only include several of the characteristics explained in this report, while the last indexes correspond to the final version of the project.
All the created indexes are multilingual, which allows us to take full advantage of the (bilingual) data collection. Additionally, we did some experiments with character 3-grams, 4-grams and 5-grams. Our motivation for experimenting with this was to compare how the size of different character N-grams affect to the effectiveness of our system. 3-grams are able to collect more specific information, while 4-grams and 5-grams allow considering bigger structures with more context and information. An additional functionality of some indexes is query expansion, but as commented, this is only applied to the English body. Finally, we created indexes with NER, which provides not only the search for keywords but also identifying and extracting specific named entities.

The subsequent indexes are:

- `2023_04_24_multilingual_3gram`: both languages of documents, using character 3-grams.
- `2023_04_29_multilingual_3gram_synonym`: both languages, character 3-grams, (English) query expansion with synonyms.
- `2023_05_01_multilingual_4gram_synonym`: both languages, character 4-grams, (English) query expansion with synonyms.
- `2023_05_01_multilingual_5gram_synonym`: both languages, character 5-grams, (English) query expansion with synonyms.
- `2023_05_05_multilingual_4gram_synonym_ner`: both languages, character 4-grams, (English) query expansion with synonyms, NER techniques.

The indexes also can be found in the following Google Drive folder.

After creating indexes, we were able to conduct multiple runs to evaluate the effectiveness of our system. These runs not only experiment with some of the techniques specified here, but also consider different versions (English or French version) of the queries. With them we can compare and analyze different aspects of our system's performance, such as precision and recall. We then computed the MAP and NDCG scores for each run, which allowed us to further evaluate the performance of our system. The results will be commented in the Section 5. The runs are the following:

- `seupd2223-JIHUMING-01_en_en`: English topics; using English body field.
- `seupd2223-JIHUMING-02_en_en_3gram`: English topics; using English body field and 3-gram field.
- `seupd2223-JIHUMING-03_en_en_4gram`: English topics; using English body field and 4-gram field.
- `seupd2223-JIHUMING-04_en_en_5gram`: English topics; using English body field and 5-gram field.
- `seupd2223-JIHUMING-05_en_en_fr_5gram`: English topics; using English and French body fileds and 5-gram field.
- `seupd2223-JIHUMING-06_en_en_4gram_ner`: English topics; using English body field, 4-gram field and NER technique.
- `seupd2223-JIHUMING-07_fr_fr`: French topics; using French body field.
- `seupd2223-JIHUMING-08_fr_fr_3gram`: French topics; using French body field and 3-gram field.
- `seupd2223-JIHUMING-09_fr_fr_4gram`: French topics; using French body field and 4-gram field.
- `seupd2223-JIHUMING-10_fr_fr_5gram`: French topics; using French body field and 5-gram field.
- `seupd2223-JIHUMING-11_fr_en_fr_5gram`: French topics; using English and French body fields and 5-gram field.
- `seupd2223-JIHUMING-12_fr_fr_4gram_ner`: French topics; using French body field, 4-gram field and NER technique.

The process of creating the indexes typically took around 1 hour, with the exception of the indexes that included NER, which took approximately 16 hours. On the other hand, generating the runs was a much quicker process, taking consistently less than a minute and a half to complete.

5. Results and Discussion

The analysis shows that the highest MAP score (0.1748) is achieved by `fr_fr_5gram`, followed by `fr_fr_4gram` (0.1737) and `fr_fr_3gram` (0.1698), while the lowest MAP score (0.0360) is obtained by `en_en_4gram_ner`. Similarly, the highest NDCG score (0.3208) belongs to

Table 1
MAP and NCDG scores for all runs

Index	Run	MAP Score	NCDG Score
01	en_en	0.0700	0.1614
02	en_en_3gram	0.0704	0.1661
03	en_en_4gram	0.0874	0.2025
04	en_en_5gram	0.1028	0.2288
05	en_en_fr_5gram	0.0669	0.1525
06	en_en_4gram_ner	0.0360	0.1098
07	fr_fr	0.1656	0.3135
08	fr_fr_3gram	0.1698	0.3208
09	fr_fr_4gram	0.1737	0.3269
10	fr_fr_5gram	0.1748	0.3285
11	fr_en_fr_5gram	0.1288	0.2797
12	fr_fr_4gram_ner	0.1362	0.2881

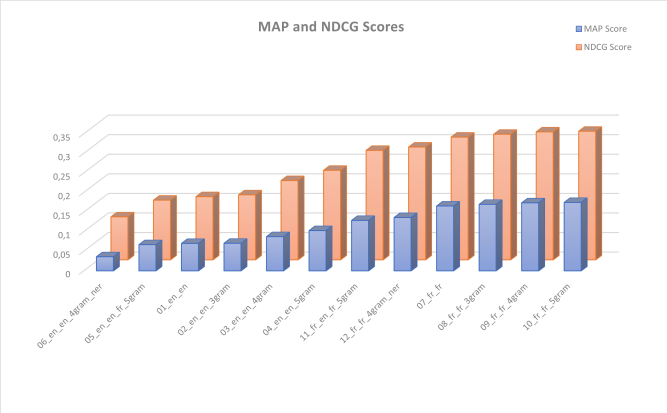

Figure 1: All scores sorted by MAP score

`fr_fr_4gram_ner`, followed by `fr_fr_5gram` (0.3285) and `fr_fr_4gram` (0.3269), whereas the lowest NDCG score (0.1098) corresponds to `en_en_4gram_ner`.
Results suggest that French queries perform better than their English counterparts, possibly due to the training data's French origin and later translation into English. Moreover, the IR system's effectiveness generally increases with a larger N-gram size, as indicated by the higher scores of `en_en_5gram` and `fr_fr_5gram`. Conversely, the inclusion of NER in the indexing process seems to have a negative impact on the scores, as shown by the lower scores of `en_en_4gram_ner` and `fr_fr_4gram_ner`. The use of query expansion with synonyms in English does not seem to improve the search results to any great extent.

Here we can see a chart ranking of the five best scores, they are the runs that have been presented at CLEF: It's interesting to notice that the cross-language approaches (`en_en_fr_5gram` and `fr_en_fr_5gram`) are out of the five bests systems. It turns out that searching for English

Figure 2: Best MAP and NDCG scores

words in French documents and vice versa messes up the search, lowering the score. Another interesting aspect is that the worst-performing index is the one with named entity recognition in English (`en_en_4gram_ner`): it combines translated queries and NER, which appears to be the two worst-performing approaches.
In general, we focus more on trying multiple approaches, this is why our score has such a big space for improvement. As already said, French queries with bigger N-gram sizes perform better. Instead of relying on single-word matches, the queries could take place with more context, resulting in better search results.

Following the competition workflow, we created the indexes based on the test data and re-executed the top five runs (see Figure 2). These runs will be the ones delivered to CLEF.

6. Conclusions and Future Work

Provide a summary of what are the main achievements and findings.

Discuss future work, e.g. what you may try next and/or how your approach could be further developed.

7. Group Members Contribution

Jesús Moncada-Ramírez

Isil Atabek

Huimin Chen

Michele Canale

Nicolò Santini

Giovanni Zago has set up the Trello application we used to divide the work among us, also providing a simple tutorial on how to use it properly. He took care of writing most of the parts in the documentation (including pages mockups) and also some application pages, i.e. editing users information as user or as administrator.

References

[1] CLEF2023, LongEval CLEF 2023 Lab, https://clef-longeval.github.io/, 2023.

[2] CLEF2023, LongEval Train Collection, https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-5010, 2023.

[3] M. Popel, M. Tomkova, J. Tomek, Ł. Kaiser, J. Uszkoreit, O. Bojar, Z. Žabokrtský, Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals, Nature Communications 11 (2020) 1–15. URL: https://www.nature.com/articles/s41467-020-18073-9. doi:10.1038/s41467-020-18073-9.

[4] S. E. Robertson, S. Walker, S. Jones, M. Hancock-Beaulieu, M. Gatford, Okapi at trec-3, in: Text Retrieval Conference, 1994.

[5] I. Ounis, G. Amati, V. Plachouras, B. He, C. Macdonald, C. Lioma, Terrier: A High Performance and Scalable Information Retrieval Platform, in: M. Beigbeder, W. Buntine, W. G. Yee (Eds.), Proc. of the ACM SIGIR 2006 Workshop on Open Source Information Retrieval (OSIR 2006), 2006.

[6] Princeton University, About WordNet, https://wordnet.princeton.edu/download/current-version, 2005.

[7] G. Diaz, A. Suriyawongkul, Stopword list in French, https://github.com/stopwords-iso/stopwords-fr/tree/master, 2023.

[8] Apache, Apache OpenNLP, https://opennlp.apache.org/, 2023.

[9] Atabek, Canale, Chen, Moncada-Ramirez, Santini and Zago, JIHUMING, https://bitbucket.org/upd-dei-stud-prj/seupd2223-jihuming/src/master/, 2023.