The University of Padua IMS Research Group at CENTRE@TREC 2018

Giorgio Maria Di Nunzio, Stefano Marchesin

Department of Information Engineering
University of Padua, Italy
{giorgiomaria.dinunzio, stefano.marchesin}@unipd.it

Abstract. In this paper, we present our participation in one of the tasks of the CENTRE@TREC 2018 Track: the Clinical Decision Support task. We describe the steps of the original paper we wanted to reproduce, identifying the elements of ambiguity that may affect the reproducibility of the results.
The experimental results we obtained follow a similar trend to those presented in the original paper: using clinical trials' "note" field decreases the retrieval performances significantly, while the pseudo-relevance feedback approach together with query expansion achieves the best results across different measures.
In the experimental results we find out that the choice of the stoplist is fundamental to achieve a reasonable level of reproducibility. However, stoplist creation is not described sufficiently well in the original paper.

Keywords: Precision medicine, query expansion, reproducibility

1 Introduction

The CENTRE@TREC challenges TREC participants to reproduce the results of the most interesting systems submitted in previous editions of TREC. Track organizers select these top performing experiments and each participating group is challenged to replicate and/or reproduce one or more of the selected systems by only using standard open source IR systems [4].

In 2018, CENTRE@TREC had three tasks: replicating runs of the TREC 2016 Clinical Decision Support track; replicating runs of the TREC 2013 and TREC 2014 Web track. In this paper, we focus on the first task: the retrieval of biomedical articles relevant for answering clinical questions based on patient records. The target document collection contains 1,251,954 full-text articles, a snapshot of the Open Access Subset of PubMed Central (PMC) from March 28, 2016. The list of topics comprises 30 clinical notes with the following structure: summary, description and note. The paper to reproduce combines pseudo-relevance feedback and semantic query expansion [3]. In our experiments, we were able to partially reproduce three of the four runs described in the original paper.

This paper is organized as follows. Section 2 lists elements of the original paper that may be ambiguous or not clear from an experimental point of view. Section 3 describes our experimental setting and the choices that we made in order to reproduce the results. Section 4 analyzes and compares the results of the original paper with the results of our runs. Finally, Section 5 concludes the paper with some general comments.

2 Analysis of the Methods

Below, we analyze the description of the model components from the original paper. First, a quote from the paper is reported to highlight the component. Then, a description of how we reproduced — or we tried to reproduce — the component follows.

> The document collection for 2016 consists of 1,251,954 PMC articles in NXML format that were indexed with Solr (version 5.5.2).

In our experiments, we used the latest version of Solr (version 7.5.0). One thing that is unclear is whether the authors have used the Data Import Handler (DIH) of Solr to import and index the content. In our case, we decided to parse the collection into a Solr XML data structure that can be used with a curl command to update the index.

> Fields used for indexing were PMCID, title, abstract, body, conclusion, journal title and journal type. Indexing was performed with standard Solr settings which include tokenization, stemming and stopword removal. A master stopword list was constructed which is a combination of standard English stopwords and a list of stopwords constructed manually based on most frequent terms occurring in the document collection.

This is a crucial point in the description of the model. There is no field named "PMCID". We believe that the "pmc" field of the article metadata is the correct mapping. Besides, the most problematic field is "conclusion" as there is no explicit field named conclusion. Therefore, in our experiments, we used pmc, title, abstract, body, journal title and journal type as fields.

Furthermore, the lack of a detailed description for the stoplist used makes the reproduction of the results extremely difficult. According to [2], "the most prominent effects are those of stop lists and IR models, as well as their interactions, while stemmers and n-grams play a smaller role". Following this consideration, we used three 'standard' stoplists (described in the experimental section) and we tried to find the optimal number of 'most frequent terms' that were manually added to the list.

> Okapi BM25 was the similarity measure used for querying and retrieval from the index.

We used the BM25 model implemented in Solr 7.5.0. According to the Solr documentation, the implementation of the BM25 model for version 5.5.2 (original paper) and version 7.5.0 (our implementation) is the same.

For query expansion, the authors of the original paper used MetaMap [1]:

> The MetaMap program was applied for the identification of UMLS concepts in topics. Since UMLS has over 100 semantic types, mappings were restricted to only the following semantic types [...] For all the mapped concepts, synonyms were extracted from UMLS knowledge sources which were thereafter used for query expansion.

We used MetaMap in the same way described in the original paper,

Finally, the description of the Pseudo-Relevance Feedback approach is as follows:

> For a given query, Pseudo-Relevance Feedback (PRF) was implemented to collect words from titles of top k retrieved documents. Stopwords were eliminated and the remaining words were added to the initial query in order to generate a new query that can be reused for searching and retrieval. [...] Although, the value of k had only minor impact on retrieval performance, k = 30 was chosen which was observed to deliver the best results. For the runs that leveraged on UMLS query expansion, terms in the topic after stopword removal in combination with UMLS terms (i.e. expanded queries) were used as initial query.

The paper is clear on the number of k documents to use for PRF. However, the lack of description for the stoplist used — especially regarding the manual selection of frequent medical terms — makes it hard to replicate the performances of PRF.

3 Experimental Setting

We tried to reproduce the following steps from the original paper: pre-processing, indexing and query expansion. A description of each step is presented below.

3.1 Pre-processing

In order to parse the collection of documents, we created an XML file containing the schema used by the update handler of Solr for adding documents to the index. The structure of the file is shown in Listing 1.1, and the code used to produce all the XML files is openly available.[1]

```
<?xml version="1.0"?>
 <add>
  <doc>
```

[1] https://github.com/gmdn/CENTRE-TREC2018

```
   <field name="pmc">1064073</field>
   <field name="title">Role of Notch signaling [...]</field>
   <field name="abstract">Introduction Notch signaling has [...]</field>
   <field name="body">Introduction Stem cells in adult tissues [...]</field>
   <field name="journal-title">Breast Cancer Research</field>
  </doc>
  <doc>
    [...]
  </doc>
</add>
```

Listing 1.1: Example of XML file indexed by Solr.

3.2 Indexing

For indexing, we tried three different stoplists: the default Solr stoplist (33 terms), the SMART stoplist (571 terms)[2], the SMART stoplist together with a list of the 300 most frequent words[3] in the collection . The union of the SMART stoplist with the most frequent words produced a stoplist of 731 words. We name the three variants in the experimental results as: Lucene, SMART, and SMART+.

The stoplists can be found at the GitHub repository.[4]

3.3 Query Expansion

For query expansion, we followed the steps described in the paper. We used MetaMap [5] to identify UMLS concepts for each query and for each title of the first 30 documents retrieved. First, we restricted the mappings of UMLS concepts to only those semantic types listed in [3]. Then, we extracted all the synonyms of the matched concepts.

4 Comparison of Results

In this section, we analyze the results obtained with our experiments by comparing the impact of the three stoplists with the original runs. We replicated three out of four runs of the original paper: MRKPrfNote, MRKSumCln, MRKUmlsSolr. In Tables 1a and 1b, we report the values of the infNDCG and infAP measures obtained in [3], along with the performances of each of our three runs based on the different stoplists (i.e. Lucene, SMART and SMART+). We also add the performance of each run in terms of Rprec in Table 1c.

The results we obtained are close to those reported in the original paper. Regardless of the stoplist used, we found that:

[2] http://members.unine.ch/jacques.savoy/clef/
[3] We used the Solr Terms Component to extract the most frequent terms.
[4] https://github.com/gmdn/CENTRE-TREC2018
[5] https://metamap.nlm.nih.gov

	infNDCG			
run	original	Lucene	SMART	SMART+
MRKPrfNote	0.150	0.125	0.133	0.138
MRKSumCln	0.222	0.187	0.197	0.202
MRKUmlsSolr	0.226	0.190	0.205	0.199

(a) infNDCG values

	infAP			
run	original	Lucene	SMART	SMART+
MRKPrfNote	0.018	0.012	0.014	0.014
MRKSumCln	0.027	0.021	0.023	0.024
MRKUmlsSolr	0.029	0.022	0.024	0.024

(b) infAP values

	RPrec			
run	original	Lucene	SMART	SMART+
MRKPrfNote	0.102	0.084	0.093	0.093
MRKSumCln	0.156	0.144	0.147	0.147
MRKUmlsSolr	0.168	0.151	0.153	0.156

(c) RPrec values

Table 1: A comparison of different performance measures among the original runs and the reproduced runs with different stoplists.

- Using the "note" field of clinical trials decreases the performance significantly (MRKPrfNote run);
- The query expansion-based approach (MRKSumCln run) increases performance significantly over the plain BM25 approach (the values of the performance for plain BM25 are not presented in the original paper nor in this paper);
- The PRF approach (MRKUmlsSolr run) increases the performance over the simple query expansion approach (MRKSumCln run).

A deeper analysis shows that we were not able to match the scores of the original paper. However, by increasing the size of the stoplist we were able to get closer and closer to the values reported in the original paper.

In Figure 1, we show a topic by topic analysis of the difference between the reproduced run MRKUmlsSolr and the original one for the three performance measures. A negative value (red bars below zero) means that the original run performed better than the reproduced run. For most of the topics, the difference in the inferred measures, infNDCG in Figure 1a and infAP in 1b, is in favour of the original run. On the other hand, for RPrec we observe a more evenly distributed difference across topics. A paired t-test confirms that there is no statistical significant difference between the two runs (the reproduced run and the original one).

5 Conclusions

In this paper, we presented the results of our participation in the CENTRE@TREC 2018 Track. Our objective was the reproduciblity of a paper aiming to retrieve biomedical articles relevant for answering clinical questions based on patient records. In particular, we wanted to reproduce the query expansion and pseudo-relevance feedback components.

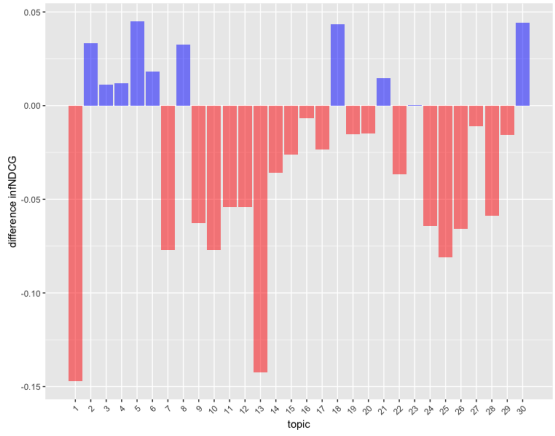
We analyzed the steps of the original paper we wanted to reproduce, and we described the elements of ambiguity that may affect the reproducibility of the results. We provided a detailed list of the choices we made for each step and, in particular, the selection of terms for the stoplist.

The experimental results we obtained follow a similar trend to those of the original paper: using the "note" field of clinical trials decreases the retrieval performances in a significant way, while pseudo-relevance feedback and query expansion obtain the best results across different measures.

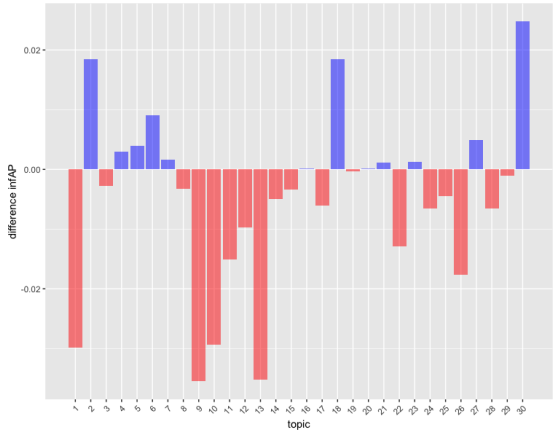
We were able to get close to the performances of the best run reported in [3] by increasing the number of terms in the stoplist. The results showed that the best results for some performance measures (such as the inferred NDCG and inferred Average Precision) are more difficult to achieve compared to other measures (such as RPrec).

References

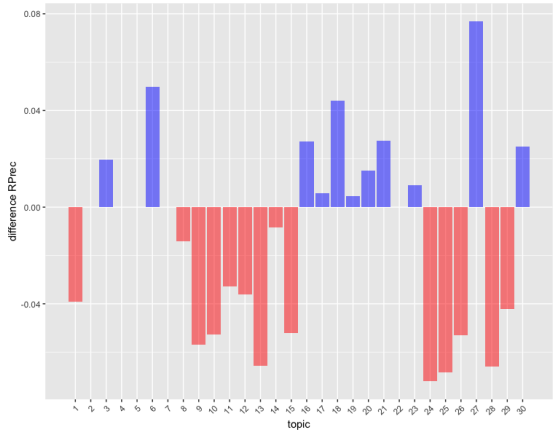
1. Aronson, A.R., Lang, F.M.: An overview of metamap: historical perspective and recent advances. Journal of the American Medical Informatics Association : JAMIA **17**(3), 229–236 (May-Jun 2010). https://doi.org/10.1136/jamia.2009.002733, `https://www.ncbi.nlm.nih.gov/pubmed/20442139`
2. Ferro, N., Silvello, G.: Toward an anatomy of IR system component performances. JASIST **69**(2), 187–200 (2018). https://doi.org/10.1002/asi.23910, `https://doi.org/10.1002/asi.23910`
3. Gurulingappa, H., Toldo, L., Schepers, C., Bauer, A., Megaro, G.: Semi-supervised information retrieval system for clinical decision support. In: Proceedings of The Twenty-Fifth Text REtrieval Conference, TREC 2016, Gaithersburg, Maryland, USA, November 15-18, 2016 (2016), `http://trec.nist.gov/pubs/trec25/papers/MERCKKGAA-CL.pdf`
4. Soboroff, I., Ferro, N., Sakai, T.: Overview of the TREC 2018 centre trask. In: Proceedings of The Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018 (2018), `https://trec.nist.gov/`



(a) infNDCG values



(b) infAP values



(c) RPrec values

Fig. 1: Topic by topic comparison of the difference of values for the MRKUmlsSolr reproduced run and the original one.