Search less, research more!
Improving information retrieval in the face of covid-19

Mario Šaško, Ivan Lovrenčić, Nikola Buhiniček

University of Zagreb, Faculty of Electrical Engineering and Computing
Unska 3, 10000 Zagreb, Croatia
{mario.sasko, ivan.lovrencic, nikola.buhinicek}@fer.hr

Abstract

As the persistent COVID-19 outbreak transforms the world, scientists are vigorously researching the virus to find a cure. For researchers to stay updated with the latest developments, they have to find the time to evaluate current advancements and potentially apply them in their research. However, because of the global initiative to solve this crisis, new research is getting published continually. That causes an abundance of research papers, where most of them do not deliver any significance to the on-going cure development. To help researchers find the most prominent improvements, we propose the solution in the form of an information retrieval model, that we adapted to the scientific domain of this problem. Our model will allow researchers to find the most relevant papers for each of their COVID-19 related queries, and in the process, save some valuable time.

1. Introduction

In the last days of 2019, numerous reports about the previously-unknown virus have appeared in Wuhan, China. Not even a few weeks after that, the World Health Organization (WHO) started alerting about a seeming global health crisis. Six months later, the highly-infectious novel Sars-Cov-2 virus has infected more than five million people and more than 340,000 people have died. Furthermore, researchers believe that this emergency will not cease until there is a viable cure, and, naturally, as a part of the global response to find it, researchers are publishing more papers than ever. However, from that arose another challenge. There is now an abundance of research papers being published, and researchers are keen to find a way to retrieve only the most prominent ones to save some valuable time.

To help researchers, the White House and the coalition of leading research groups came up with an initiative to prepare the COVID-19 Open Research Dataset (*CORD-19*[1]). This freely available dataset is given to the global research community in order to apply natural language processing (NLP) and other AI techniques, that could give us a better understanding of the Sars-CoV-2 virus and, sequentially, help scientists find the most relevant research about the ongoing pandemic.

To contribute, we have decided to develop our IR model. The main intention of our approach was to allow researchers to obtain the most relevant papers based on their query input. To achieve that, we have used a stratified model, that through each level, filters and ranks papers based on their query relevance. The model starts with a biomedical named entity recognition (BioNER) to filter out papers that are not related to the current query. Furthermore, the model proceeds by utilizing the word (word2vec) and document (doc2vec) level embeddings to determine the remaining papers significance rank. In the end, we are left with a list of query-related papers, which are sorted based on their relevance score.

In the following section, we will discuss the approaches of other highly-ranked proposed solutions and compare them to our own. Furthermore, in section 3., we will explain how we used the given CORD-19 dataset for evaluation of our model, which will be in detail explained in section 4. Moreover, in section 5., we will talk more about evaluation and results. We had challenges in the evaluation part, as it was hard to evaluate the significance of the research paper from the domain we are not experts. Lastly, we finish with a conclusion and ideas for future work.

2. Related work

The mentioned CORD-19 dataset came along with a series of coronavirus related *tasks*[2] and so far there has been over 1,400 submissions. There could be an article written about each of those solutions, but we will only shortly cover the approaches of few most accepted ones.

The officially accepted submission for this problem is actually using only paper abstracts for relevance determination. The query is processed into keywords which are stemmed and concatenated with corona related terms (covid, cov, -cov, hcov). Then the papers are filtered by checking if the abstract contains all of the queried words and at least one corona related term. Later on, for the remaining papers, relevance scores are calculated based on the keyword counts and the length of the abstract. Based on our work and the issues we encountered, the decision to use only abstracts for retrieval isn't justifiable. Our model is utilising more of the available data, weighting parts of papers differently and not relying just on the short summaries.

There is a different approach, also highly rated, that turned to a probabilistic IR model, BM25, using it along with document embeddings. They are using the BM25 search index, created from paper abstracts, as their base and boosting the performance with an Annoy index made out of 786 dimensional Specter Vectors, which represent document embeddings. This is just one of many submissions

[1] https://bit.ly/2ZMyMTZ

[2] https://bit.ly/3gyyaHv

that is using some level of embeddings while we decided to use both word and document embeddings to get a better insight. Alongside with embeddings, another often used NLP tool is named entity recognition (NER). To adapt more to the task, we used BioNER, a specialised NER for entities commonly searched in this scientific domain.

Another acclaimed solution is more machine learning oriented. They are relying on the article body which they parse using NLP. Each document is then turned into a vector, based on TF-IDF, and the rest of work is done by applying k-means clustering. Our model though, isn't going into ML but there are plenty of ML approaches for this task which is worth mentioning.

3. Dataset

As mentioned in Section 1., we use the CORD-19 dataset to evaluate the model. The dataset contains over 134,000 scholarly articles with approximately half of them being related to COVID-19. Out of the total number, 29315 articles are presented as a full-text JSON. In the task, we focus on fields title, abstract and body and discard the rest, which aligns with our assumption that these fields are the most important ones.

In order to measure the model performance, we sample the dataset and obtain 95 articles. Next, these articles are annotated as relevant or irrelevant with 3x coverage, according to the query. Finally, the majority agreement is used to define gold-standard labels.

During preprocessing, the aforementioned JSON fields are concatenated and tokenized using *spaCy*[3], an advanced natural language processing library. Then, the input is lowercased and passed to the model. The same steps, with additional removal of stop words, are applied when working with the baselines.

4. Baselines

We use two baselines to validate the performance of our model. Both models are part of *Apache Lucene*[4], a high-performance search engine library.

The first one of them, an unigram Language Model, builds one probabilistic model per article in collection. Then for each article, the model outputs the probability of a query belonging to it. We sort these probabilities to obtain the final relevance ranking. To make predictions more robust, the model relies on Bayesian smoothing with Dirichlet priors (Zhai and Lafferty, 2004).

As our second baseline model, we use a BM25 Okapi ranking function (Robertson et al., 1995). The ranking function weighs query terms according to an advanced TF-IDF weighting scheme. Since this is the only function from BM family that properly scores longer articles, we don't consider other variants. Additionaly, the same sorting procedure that is used with the LM model is applied to the function output.

Table 1: Comparison of performance for the query *What is known about about Covid-19 transmission*.

Model	P@5	P@10	P@15	R-Prec	AveP
LM	0.80	0.60	0.47	0.67	0.69
BM25	0.80	0.70	0.47	**0.78**	0.70
Our Model	0.80	0.70	**0.60**	0.67	**0.84**

Table 2: Comparison of performance for the query *What is known about about Covid-19 incubation*.

Model	P@5	P@10	P@15	R-Prec	AveP
LM	0.60	0.50	0.33	0.60	0.64
BM25	0.40	0.50	0.33	0.40	0.55
Our Model	**0.80**	0.50	0.33	**0.80**	**0.97**

5. Results

Two COVID-19-related queries are defined to evaluate the model. For each query, the model assigns relevance scores according to which the ranking of the articles is formed. Since the resulting list contains each article from the sample dataset without the cutoff, only IR-related metrics are used. Due to the volume of the annotated data, the results are not statistically tested.

The results in Table 1 show system performance for the query *What is known about Covid-19 transmission*. Reasonably high R-precision denotes that our model, as well as the baselines, is confident in the relevance of the highest-scoring articles. Furthermore, we assume that by utilizing the word- and document-level embeddings, the model has a chance to infer more subtle similarities between a query and an article which leads to notably higher precision at 15 and average precision. Traditional term-frequency-based approaches fail in such situations.

Similarly, the results for the query *What is known about Covid-19 incubation* are in Table 2. This time our model outperforms the baselines in the ranking of 5 highest-scoring articles. Even higher average precision than in previous case signals that the model, compared to the baselines, ranks relevant articles more consistently and closer to each other.

5.1. First subsection

This is a subsection of the second section.

5.2. Second subsection

This is the second subsection of the second section. Referencing the (sub)sections in text is performed as follows: "in Section 5.1. we have shown . . . ".

5.2.1. Sub-subsection example

This is a sub-subsection. If possible, it is better to avoid sub-subsections.

[3] https://bit.ly/3gEuDYp

[4] https://bit.ly/3ezPxWF

Table 3: This is the caption of the table. Table captions should be placed *above* the table.

Heading1	Heading2
One	First row text
Two	Second row text
Three	Third row text
	Fourth row text

6. Extent of the paper

The paper should have a minimum of 3 and a maximum of 4 pages, plus an additional page for references.

7. Figures and tables

7.1. Figures

Here is an example on how to include figures in the paper. Figures are included in LaTeX code immediately *after* the text in which these figures are referenced. Allow LaTeX to place the figure where it believes is best (usually on top of the page of at the position where you would not place the figure). Figures are referenced as follows: "Figure **??** shows . . . ". Use tilde (˜) to prevent separation between the word "Figure" and its enumeration.

7.2. Tables

There are two types of tables: narrow tables that fit into one column and a wide table that spreads over both columns.

7.2.1. Narrow tables

Table 3 is an example of a narrow table. Do not use vertical lines in tables – vertical tables have no effect and they make tables visually less attractive. We recommend using *booktabs* package for nicer tables.

7.3. Wide tables

Table 4 is an example of a wide table that spreads across both columns. The same can be done for wide figures that should spread across the whole width of the page.

8. Math expressions and formulas

Math expressions and formulas that appear within the sentence should be written inside the so-called *inline* math environment: $2 + 3$, $\sqrt{16}$, $h(x) = \mathbf{1}(\theta_1 x_1 + \theta_0 > 0)$. Larger expressions and formulas (e.g., equations) should be written in the so-called *displayed* math environment:

$$b_k^{(i)} = \begin{cases} 1 & \text{if } k = \text{argmin}_j \|\mathbf{x}^{(i)} - \mu_j\|, \\ 0 & \text{otherwise} \end{cases}$$

Math expressions which you reference in the text should be written inside the *equation* environment:

$$J = \sum_{i=1}^{N} \sum_{k=1}^{K} b_k^{(i)} \|\mathbf{x}^{(i)} - \mu_k\|^2 \tag{1}$$

Now you can reference equation (1). If the paragraph continues right after the formula

$$f(x) = x^2 + \varepsilon \tag{2}$$

like this one does, use the command *noindent* after the equation to remove the indentation of the row.

Multi-letter words in the math environment should be written inside the command *mathit*, otherwise LaTeX will insert spacing between the letters to denote the multiplication of values denoted by symbols. For example, compare $Consistent(h, \mathcal{D})$ and
$\mathit{Consistent}(h, \mathcal{D})$.

If you need a math symbol, but you don't know the corresponding LaTeX command that generates it, try *Detexify*.[5]

9. Referencing literature

References to other publications should be written in brackets with the last name of the first author and the year of publication, e.g., (Chomsky, 1973). Multiple references are written in sequence, one after another, separated by semicolon and without whitespaces in between, e.g., (Chomsky, 1973; Chave, 1964; Feigl, 1958). References are typically written at the end of the sentence and necessarily before the sentence punctuation.

If the publication is authored by more than one author, only the name of the first author is written, after which abbreviation *et al.*, meaning *et alia*, i.e., and others is written as in (Johnson et al., 1976). If the publication is authored by only two authors, then the last names of both authors are written (Johnson and Howells, 1974).

If the name of the author is incorporated into the text of the sentence, it should not be in the brackets (only the year should be there). E.g., "Chomsky (1973) suggested that . . . ". The difference is whether you reference the publication or the author who wrote it.

The list of all literature references is given alphabetically at the end of the paper. The form of the reference depends on the type of the bibliographic unit: conference papers, (Chave, 1964), books (Butcher, 1981), journal articles (Howells, 1951), doctoral dissertations (Croft, 1978), and book chapters (Feigl, 1958).

All of this is automatically produced when using BibTeX. Insert all the BibTeX entries into the file `tar2020.bib`, and then reference them via their symbolic names.

10. Conclusion

In this paper, we present a novel task-specific model for the query-relevance ranking of the CORD-19 articles. The model combines a biomedical named entity recognition with word- and document-level embeddings to determine the relevance score of each article. We use a LM model and a BM25 ranking function as baselines, which our model outperforms on the sample dataset.

Note that our ultimate goal is to prove that a domain-specific model should be considered when tackling IR tasks.

As part of future work, we would like to increase the volume of the sample dataset. This would allow us to validate

[5] `http://detexify.kirelabs.org/`

Table 4: Wide-table caption

Heading1	Heading2	Heading3
A	A very long text, longer that the width of a single column	128
B	A very long text, longer that the width of a single column	3123
C	A very long text, longer that the width of a single column	−32

our results statistically. Additionally, when annotating articles we would like to use expert knowledge and greater coverage. Lastly, the comparison with deep learning approaches would be beneficial.

OVO OSTAVLJAM KAO PODSJETNIK.Conclusion is the last enumerated section of the paper. It should not exceed half of a column and is typically split into 2–3 paragraphs. No new information should be presented in the conclusion; this section only summarizes and concludes the paper.

Acknowledgements

If suitable, you can include the *Acknowledgements* section before inserting the literature references in order to thank those who helped you in any way to deliver the paper, but are not co-authors of the paper.

References

Judith Butcher. 1981. *Copy-editing*. Cambridge University Press, 2nd edition.

K. E. Chave. 1964. Skeletal Durability and Preservation. In J. Imbrie and N. Newel, editors, *Approaches to paleoecology*, pages 377–87, New York. Wiley.

N. Chomsky. 1973. Conditions on Transformations. In S. R. Anderson and P. Kiparsky, editors, *A festschrift for Morris Halle*, New York. Holt, Rinehart & Winston.

W. B. Croft. 1978. *Organizing and searching large files of document descriptions*. Ph.D. thesis, Cambridge University.

F. Feigl, 1958. *Spot Tests in Organic Analysis*, chapter 6. Publisher publisher, 5th edition.

W. W. Howells. 1951. Factors of human physique. *American Journal of Physical Anthropology*, 9:159–192.

G. B. Johnson and W. W. Howells. 1974. Title title title title title title title title title title. *Journal journal journal*.

G. B. Johnson, W. W. Howells, and A. N. Other. 1976. Title title title title title title title title title title. *Journal journal journal*.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Chengxiang Zhai and John Lafferty. 2004. A study of smoothing methods for language models applied to information retrieval. *ACM Transactions on Information Systems (TOIS)*, 22(2):179–214.