BioASQ Synergy: A strong and simple baseline rooted in relevance feedback

Tiago Almeida[1], Sérgio Matos[1]

[1]University of Aveiro, IEETA

Abstract

This paper presents the participation of the University of Aveiro Biomedical Informatics and Techologies (BIT) group in the Synergy task of the ninth edition of the BioASQ challenge. Given availability of feedback data between rounds, we explored a traditional relevance feedback approach. More precisely, we performed query expansion by selecting the highest tf-idf terms from snippets judged as relevant by experts. Then, the revised query is processed by our BioASQ-8b pipeline consisting of BM25 followed by a lightweight neural reranking model. Our system achieved results above the median, which given its simplicity can be considered satisfactory. Furthermore, in two batches our best results were only second to the runs submitted by the top performing team. Code to reproduce our submissions are available on https://github.com/bioinformatics-ua/BioASQ9-Synergy.

Keywords

Relevance Feedback, BM25, Neural ranking, Covid-19, Document Retrieval, BioASQ Synergy

1. Introduction

In January 2020 the World Health Organization (W.H.O.) declared the 2019 corona virus as a global health emergency. More than one year later, and even with the existence of vaccines, the virus still affects the majority of the world population. Furthermore, studies are still being conducted and new material about the virus is published every day. This causes a wave of knowledge, firstly available through scientific articles, which without effective searching tools ends up deprecating precious research time. So, it becomes imperative to improve the access to this type of unstructured information in order to foster further research about the novel corona virus.

TREC was the first institution to launch a global challenge, TREC-Covid [1], to push the research on search tools for dealing with the exponential growth of the literature about the novel corona virus. The BioASQ organization followed the same path and launched the Synergy task, where the aim is to retrieve the most relevant answers to biomedical questions about this corona virus.

This paper describes the participation of the Biomedical Informatics and Techologies (BIT) group of the Aveiro University in the Synergy challenge, which consisted in retrieving, from

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania

tiagomeloalmeida (T. Almeida); aleixomatos@ua.pt (S. Matos)
https://t-almeida.github.io/online-cv/ (T. Almeida)
0000-0002-4258-3350 (T. Almeida); 0000-0003-1941-3983 (S. Matos)

the CORD-19 [2] collection, documents and snippets that are relevant for a given biomedical question related to the novel corona virus.

Our approach builds on the lessons learned from our participation in the TREC-Covid challenge [3]. In TREC-Covid, due to the nature of the residual evaluation, we observed that relevance feedback approaches drastically benefit from this setup. So, we decided to constructed a strong baseline based on relevance feedback techniques and then tried to rerank this to achieve further improvements.

We achieved satisfying results with our simple approach, losing only to the first place team. In the remaining of the paper we describe in more detail our relevance feedback approach. We then describe the submissions and the results obtained, followed by a general discussion.

2. Relevance Feedback

In this section, to make the paper self-contained, we first made a briefly introduction to the topic of relevance feedback, a well known technique studied in the field of information retrieval in which the main idea is to directly include the user feedback into the retrieval process. In other words, the user will refine the quality of the results by selecting positives example from the initial retrieval order. With more detail, the basic procedure of relevance feedback can be summarize by the following steps:

1. A simple query, encoding the information need, is processed by the system.
2. The results are returned to the user.
3. From that initial list, the user selects some positive and negatives examples.
4. The system creates a new representation of the information need by using the query, the positive examples and the negatives examples.
5. The final retrieved documents are returned to the user.

The main concern when implementing a relevance feedback algorithm regards creating a new representation of the information need from the original query, positives and negatives examples. Following the literature, the most well known method is the Rocchio [4] algorithm. This algorithm operates in the vector space model, where document and queries are represented as vectors. The main idea is to produce a new query vector by combining the original query vector, plus a weighted representation of the positive documents minus a weighted representation of the negative documents. Then the retrieval is done by projecting the new query to the vector space and retrieving by cosine similarity the closest documents. The intuition is to modify the original query in order to move it closer to the positive examples and farther away from the negative examples.

3. Methodology

In this section we describe our main solution, consisting of a combination of BM25 with relevance feedback, and explain the intuition behind this approach. To better understand our rational, we first analyze the format of the Synergy task.

The Synergy task appeared as an effort to help finding answers to biomedical questions about the 2019 novel coronavirus. Unlike the usual BioASQ format, the Synergy task presented a fundamental change concerning its evaluation and flow. More precisely, the Synergy task followed a residual type of evaluation, similar to TREC-Covid, where the test set is reused through all the batches. Additionally, in between batches the golden feedback data, i.e., the relevance information for each question, was made available to the participants. This ends up changing the usual retrieval paradigm, in which one is expected to apply a retrieval system on an unknown question. So, according to the literature, the Synergy tasks becomes suitable to relevance feedback techniques, since some relevant examples were available for a majority of the questions, which satisfies the points 2 and 3 of the relevance feedback procedure. This observation can be confirmed by the TREC-Covid challenge results, where relevance feedback runs were able to achieve top scoring positions, outscoring traditional, neural and transformer based retrieval approaches.

Based on these observations and also inspired by our previous submission to the TREC-Covid challenge [3], we adopted the traditional BM25 ranking function combined with a simple relevance feedback method for constructing a strong baseline for this challenge. Then, we also tried to employ our existing BioASQ 8b neural ranking model to further rerank our baseline.

3.1. Baseline - BM25 with Relevance Feedback

As previously mentioned, we adopted the BM25 ranking function as our retrieval function, since it is known to produce close to state of the art results when well fine-tuned. In order to include relevance feedback in the BM25 algorithm, we follow a similar intuition from the Rocchio algorithm of adding a representation of the positive documents to the query. However, since the BM25 is a probabilistic model and not a vector model, we employed a query expansion technique based on the most representatives terms of each positive document. This new query is then processed by the BM25, hopefully returning a new list of documents that are more similar to the positive documents.

The representative terms of each document were selected as the top-k terms with higher tf-idf score. The intuition behind this assumption is that the terms with higher tf-idf score will largely contribute to the final ranking score. Thus by including them in the new query we are boosting the documents that are most similar, in terms of tf-idf terms, to our positives examples.

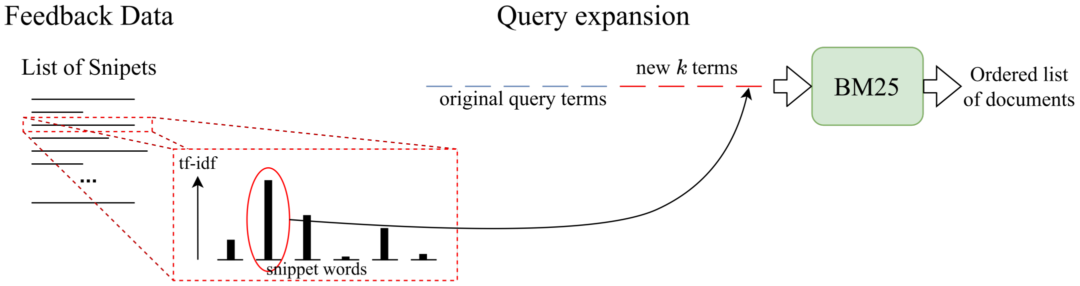

Figure 1: Summary of the procedure to combine relevance feedback with the BM25 ranking function.

After selecting the k most important terms from the collection of positive examples, these are added to the original query in an disjunctive form. Then the normal BM25 ranking function is applied over this new generated query. The overall procedure of relevance feedback is illustrated in Figure 1.

3.1.1. Impact of the Source of Relevance

Another important detail is the source of positive examples that we feed as feedback data. For that there are two alternatives, the text from the list of positive documents (1) and the text from the list of positive snippets (2). In order to chose the best candidate we performed an empirical evaluation using the first round feedback data. More precisely, we performed a 60% random split of the questions, resulting in 60 queries for validation and 41 queries for testing. The validation set was used to finetune the relevance feedback and BM25 parameters, and the final results are reported over the test set. The parameters that were finetuned were the $k1$ and b parameters for BM25, the number of terms to add to the query, k, the maximum number of positive samples per question, S_{max}, and minimum frequency for the query expansion, F_{min}. Additionally, we also finetuned a boost parameter that multiplies the contribution of the original query terms with respect to the added terms. Table 1 shows the range and best value for the parameters. For both experiments we first used random search over a large space of parameters and then proceeded with grid search that best fit each experiment.

Table 1
Set of parameters that were finetuned. In bold we report the best values found for the second round of the BioASQ-Synergy. The notation { X to Y, Z} means that we search between X and Y in Z increments.

Type of search	**BM25**		**RF: Query expansion**			**Boost**
	$k1$	b	k	S_{max}	F_{min}	
Random Search (both)	{0.1 to 1.2, 0.1}	{0.1 to 1, 0.1}	{15 to 80, 5}	{5 to 50, 5}	{5 to 50, 5}	[1,2,4]
Gird Search - Docs	[.4,.6,**.8**,1]	[**.4**,.6,.8]	[5,10,**15**,20,30]	[15,**30**,40,50]	[30,40,**50**]	[**2**,4]
Gird Search - Snippet	[.6,.8,1,**1.2**,1.4]	[.4,.6,**.8**]	[70,**75**,80]	[30,40,**50**]	[**1**]	[2,**4**]

In Table 2 we show the performance of the best and worst set of parameters when using documents and snippets as a source of positive examples. From the experiment, it is clear that the list of snippets are far better candidates than the list of document to extract the most representatives terms to expand the query. We believe that this discrepancy is related to the scope hypothesis [5], that says that a document can address several topics. This will result in the extraction of terms unrelated with the question topic, hence causing query drift.

Furthermore, the snippet hyper-parameter search is also more reliable, with a much smaller difference between the best and worst parameters.

3.2. Neural Rerank

Since it is expected that neural reranking models will bring some improvements over traditional baselines, we also included some runs where we tried to rerank our baseline produced by the previous approach.

For this neural reranking, we relied on our neural architecture that was used in the BioASQ 8b challenge [6]. Following the lessons learned from TREC-Covid [3], we found that reranking

Table 2
Comparison between the source of positives examples.

Positive Examples	Validation Set		Test Set	
	MAP@10	Recall@10	MAP@10	Recall@10
Docs$_{best}$ (1)	19.00	25.66	18.68	24.05
Docs$_{worst}$ (1)	5.37	9.12	4.67	7.32
Snippet$_{best}$ (2)	46.77	46.71	**46.30**	**47.34**
Snippet$_{worst}$ (2)	41.69	44.47	44.73	46.10

over relevance feedback runs is more effective when the number of candidate documents is small.

4. Submission

In this section, we start by describing the data collection and some pre-processing steps. Then we detail each run that was submitted on each batch. Note that all the runs submitted and the results presented are with respect to the document retrieval task.

4.1. Collection and Pre-processing

The Synergy task used the CORD-19 [2] collection, which is a open collection of scientific articles about the 2019 novel coronavirus. Currently, it is updated on a weekly basis and has more than 550 thousand articles gathered from peer-reviewed publications and open archives such as bioRxiv and medRxiv. For the task, only documents that had pmid, abstract and title were considering, meaning that roughly 60% of the articles were discarded.

At each round we indexed the valid set of articles with Elasticsearch using the english text analyzer, which automatically performs tokenization, stemming and stopword filtering. Additionally, we also included an analyzer to perform expansion of Covid-19 related terms by using a synonym expansion list.

Regarding the neural ranking model, we kept the same model architecture described in the 2020 BioASQ 8b challenge [6]. Additionally, we trained 200-dimensinal word embeddings using the GenSim [7] implementation of word2vec [8], with the combination of PUBMED plus CORD-19.

4.2. Runs

The first version of the Synergy task had four rounds, with no feedback data available for the first round. Therefore, we could not apply our relevance feedback baseline for the first round, and used instead a BM25 baseline with neural reranking.

Table 3 presents the summary of all the submissions, where RF stands for relevance feedback and NN for reranking with a neural network that was trained on the feedback data, with NN (TREC-Covid) meaning that the neural network was trained with the trec-covid data and NN (BioASQ) meaning it was trained with the bioASQ data. Furthermore, BM25 was fine-tuned for

Table 3
Summary of the submitted runs for each round of the 2020 BioASQ Task Synergy. RF: relevance feedback; NN: neural network reranking.

Run name	Description		
	Round 1	Round 2	Round 3 and 4
bioinfo-0	BM25	BM25 + RF	BM25 + RF
bioinfo-1	BM25 + Synonyms	BM25 + RF	BM25 + RF + NN
bioinfo-2	BM25 + NN (TREC-Covid)	BM25 + RF + NN	BM25 + RF + NN
bioinfo-3	BM25 + NN (TREC-Covid)	BM25 + RF + NN	BM25 + RF + NN
bioinfo-4	BM25 + NN (BioASQ)	BM25 + RF + NN	BM25 + RF + NN

each round and we set the parameter k to 75, which means that a maximum of 75 new terms were added to the revised query.

5. Results

The overall results are shown in Table 4, together with the median of all submissions and the result of the top performing system in each batch. The results are organized according the Mean Average Precision at ten (MAP@10), which was the measure adopted by organizers to rank all the received submissions. There were a total of 20, 21, 23 and 24 submissions respectively for each round.

Table 4
Summary of the results obtained

Run name	Round 1		Round 2		Round 3		Round 4	
	Rank	MAP	Rank	MAP	Rank	MAP	Rank	MAP
bioinfo-0	13	22.28	7	31.93	15	18.08	6	23.13
bioinfo-1	14	22.08	6	32.59	9	21.26	10	21.44
bioinfo-2	16	18.60	13	27.58	16	18.05	15	20.09
bioinfo-3	18	15.37	15	26.48	13	19.84	11	21.44
bioinfo-4	12	22.52	14	26.58	10	20.80	12	20.91
Median		27.35		28.45		21.26		23.13
Top result		33.75		40.69		32.57		29.83

When looking at the results presented in Table 4, it is important to notice that the main method presented in this paper was only used in rounds 2, 3 and 4. Nonetheless, from the first round results it is possible to observe that the runs that used the TREC-Covid data resulted in the worst performance, below the normal baseline and the run trained with BioASQ Task b data. This is an interesting behavior, since the model that was trained with domain data (TREC-Covid) had worst performance against the model that was trained in a more generic domain (BioASQ). We theorize that this may be related to the differences in the query structure from TREC-Covid, also known as topics, and the more human like questions used in the Synergy task. Another aspect is related to the differences in terms of feedback data. More precisely, TREC-Covid has a very low number of questions but higher number of feedback documents per question, while

BioASQ has a compatible larger number of queries and lower number of feedback documents per question.

Regarding rounds 2, 3 and 4 we achieved competitive performance taking into consideration the simple approach. Notably our best scores correspond to submissions that just used BM25 with relevance feedback, in round 2 and 4, which means that the neural reranking in those rounds lowered the overall performance. However, in round 3 our best performance was achieved with a reranking strategy, making it inconclusive if our reranking technique over the relevance feedback baseline is beneficial or detrimental. In terms of team ranking positions, our technique achieved two second places in rounds 2 and 4, scoring below the strong submissions of the first place team, as well as a third place in round 3.

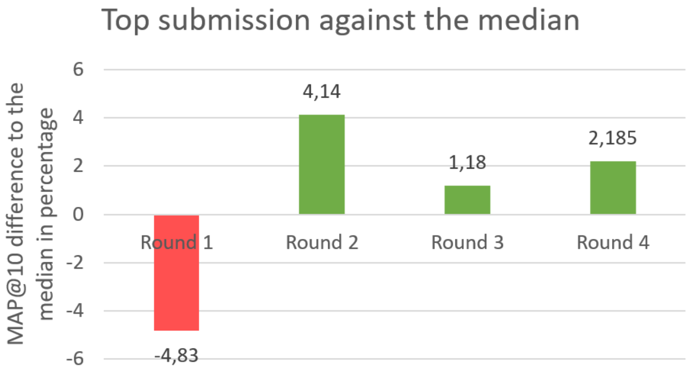

Figure 2: MAP@10 difference between our best run at each round against the median score at that round.

To get a better context of the overall performance in relation to all the submissions we show in Figure 2 the difference in terms of MAP@10 of our best submissions against the median score presented in the leaderboards. Notably, the relevance feedback solution performed as expected and gave us a simple solution that managed to consistently achieve above average results.

6. Conclusion

In this paper we present a simple but strong baseline rooted in a relevance feedback technique. More precisely, we combined the traditional BM25 ranking function with a tf-idf based query expansion, that will add the relevance feedback to the ranking function.

From the results obtained our relevance feedback manage to perform well above average, supporting our initial idea that relevance feedback runs prevail in residual type of evaluations.

Acknowledgments

This work has received support from the EU/EFPIA Innovative Medicines Initiative 2 Joint Undertaking under grant agreement No 806968 and from National Funds through the FCT - Foundation for Science and Technology, in the context of the grant 2020.05784.BD.

References

[1] E. M. Voorhees, T. Alam, S. Bedrick, D. Demner-Fushman, W. R. Hersh, K. Lo, K. Roberts, I. Soboroff, L. L. Wang, TREC-COVID: Constructing a pandemic information retrieval test collection, ArXiv abs/2005.04474 (2020).

[2] L. L. Wang, K. Lo, Y. Chandrasekhar, R. Reas, J. Yang, D. Burdick, D. Eide, K. Funk, Y. Katsis, R. M. Kinney, Y. Li, Z. Liu, W. Merrill, P. Mooney, D. A. Murdick, D. Rishi, J. Sheehan, Z. Shen, B. Stilson, A. D. Wade, K. Wang, N. X. R. Wang, C. Wilhelm, B. Xie, D. M. Raymond, D. S. Weld, O. Etzioni, S. Kohlmeier, CORD-19: The COVID-19 open research dataset, in: Proceedings of the 1st Workshop on NLP for COVID-19 at ACL 2020, Association for Computational Linguistics, Online, 2020. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-acl.1.

[3] T. Almeida, S. Matos, Frugal neural reranking: evaluation on the covid-19 literature, in: Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020, Association for Computational Linguistics, Online, 2020. URL: https://www.aclweb.org/anthology/2020.nlpcovid19-2.3. doi:10.18653/v1/2020.nlpcovid19-2.3.

[4] J. J. Rocchio, Relevance Feedback in Information Retrieval, Prentice Hall, Englewood, Cliffs, New Jersey, 1971. URL: http://www.is.informatik.uni-duisburg.de/bib/docs/Rocchio_71.html.

[5] S. E. Robertson, S. Walker, Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval, in: B. W. Croft, C. J. van Rijsbergen (Eds.), SIGIR '94, Springer London, London, 1994, pp. 232–241.

[6] T. Almeida, S. Matos, BIT.UA at BioASQ 8: Lightweight neural document ranking with zero-shot snippet retrieval, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_161.pdf.

[7] R. Řehůřek, P. Sojka, Software Framework for Topic Modelling with Large Corpora, in: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks, ELRA, Valletta, Malta, 2010, pp. 45–50. http://is.muni.cz/publication/884893/en.

[8] T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, Distributed representations of words and phrases and their compositionality, in: Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, NIPS'13, Curran Associates Inc., Red Hook, NY, USA, 2013, p. 3111–3119.