A Study on Reciprocal Ranking Fusion in Consumer Health Search. IMS UniPD at CLEF eHealth 2020 Task 2

Giorgio Maria Di Nunzio[1,2], Stefano Marchesin[1], and Federica Vezzani[3]

[1] Dept. of Information Engineering – University of Padua
[giorgiomaria.dinunzio,stefano.marchesin]@unipd.it
[2] Dept. of Mathematics – University of Padua
[3] Dept. of Linguistic and Literary Studies – University of Padua
federica.vezzani@unipd.it

Abstract. In this paper, we describe the results of the participation of the Information Management Systems (IMS) group at CLEF eHealth 2020 Task 2, Consumer Health Search Task. In particular, we participated in both subtasks: Ad-hoc IR and Spoken queries retrieval. The goal of our work was to evaluate the reciprocal ranking fusion approach over 1) different query variants; 2) different retrieval functions; 3) w/out pseudo-relevance feedback. The results show that, on average, the best performances are obtained by a ranking fusion approach together with pseudo-relevance feedback.

1 Introduction

CLEF eHealth is an evaluation challenge where the goal is to provide researchers with datasets, evaluation frameworks, and events to evaluate the performance of IR systems in the medical IR domain. In the CLEF eHealth 2020 edition [5], the organizers set up two tasks to evaluate retrieval systems on different domains. In this paper, we report the results of our participation to the Task 2 "Consumer Health Search" [4]. This task investigates the problem of retrieving documents to support the needs of health consumers that are confronted with a health issue. In particular, we participated in both the subtasks available: the Ad-hoc IR task and the Spoken queries retrieval task.

The contribution of our experiments to both subtasks can be summarized as follows:

- A study of a manual query variation approach similar to [7, 8];
- An evaluation of a ranking fusion approach [3] on different document retrieval strategies, with or without pseudo-relevance feedback [10].

Table 1. Examples of query variants given the original query for subtask 1.

id	type	text
151001	original	anemia diet therapy
151001	variant 1	anaemia diet cure
151001	variant 2	diet treatment for the decrease in the total amount of red blood cells (RBCs)or hemoglobin in the blood
152001	original	emotional and mental disorders
152001	variant 1	psychiatric disorder
152001	variant 2	psychological disorder
152001	variant 3	mental illness
152001	variant 4	mental disease
152001	variant 5	mental disorder
152001	variant 6	nervous breakdown
152001	variant 7	emotional disturbance such as: anxiety, bipolar, conduct, eating, obsessive-compulsive (OCD) and psychotic disorders

The remainder of the paper will introduce the methodology and a brief summary of the experimental settings that we used in order to create the official runs that we submitted for this task.

2 Methodology

In this section, we describe the methodology for merging the ranking list provided by different retrieval methods for different query variants.

2.1 Subtask 1: Ad-hoc IR

Query variants : In this subtask, we asked to an expert in the field of medical Terminology to rewrite the original English query into as many variants as she preferred. The aim of the query rewriting was to describe in the best possible way (given the knowledge of the user) the information need expressed by the query. In table 1, we show the variants for the first two queries (151001, 152001). These examples show how the number of variants as well as the complexity of the request (from a few keywords to complex sentences) may change across queries.

Retrieval models : For each query, we run three different retrieval models: the Okapi BM25 model [9], the divergence from randomness model [1], the language model using Dirichlet priors [11]. We used the RM3 Positional Relevance model to implement a pseudo-relevance feedback strategy including query expansion [6].

Ranking fusion : Given different ranking lists, we used the reciprocal ranking fusion (RRF) approach to merge them [2].

Table 2. Examples of query variants for subtask 2. Only the first three variants are shown.

id	type	text
151001	participant 1	anemia diet changes
151001	participant 2	Diet for anemia
151001	participant 3	What food can i eat on this diet
152001	participant 1	causes of withdrawal
152001	participant 2	What diseases may cause mental health?
152001	participant 3	what mental health conditions can cause mood alterations cause somebody to become more withdrawn

2.2 Subtask 2: Spoken queries retrieval

Query variants : In this subtask, there are already available a number of query variants that were (audio) recorded by six users. For this task, we used the different transcriptions of these audio files: clean transcript, default variant, phone enhanced variant, video enhanced variant. In Table 2, we show three examples of variants (out of six) for the first two queries.

Retrieval models : for this subtask, we used only the Okapi BM25 retrieval model and the RM3 pseudo-relevance feedback model.

Ranking fusion : given different ranking participants and different transcripts, we used the RRF approach to merge them.

3 Experiments

In this section, we describe the experimental settings and the results for each subtask.

3.1 Search Engine

For all the experiments, we used the Elasticsearch search engine[4] and the indexes provided by the organizers of the task. We used the following parameter settings for each retrieval model:

- BM25, $k2 = 1.2$, $b = 0.75$
- LMDirichlet, $\mu = 2000$
- DFR, *basic_model* = if, *after_effect* = b, *normalization* = h2

The RM3 pseudo-relevance feedback model was implemented with the following strategy: pick the 10 most relevant terms from the top 10 ranked documents, add these terms to the original query with a weight equal 0.5 (while the original terms are weighted 1.0), run the expanded query and produce the final ranking list.

[4] `https://www.elastic.co/products/elasticsearch`

3.2 Runs

For each subtask, we submitted four runs.

Subtask 1 . For the Ad-hoc retrieval subtask, the runs are:

- clef_bm25_orig: Only BM25 (no rank fusion) using the original query only;
- clef_original_rrf: Reciprocal rank fusion with BM25, QLM, DFR models and the original query;
- clef_original_rm3_rrf: Reciprocal Rank fusion with BM25, QLM, DFR approaches using RM3 pseudo relevance feedback and the original query;
- clef_variant_rrf: BM25 and reciprocal rank fusion on the rankings produced by the original and manual variants of the query.

Subtask 2 . For the spoken queries retrieval subtask, the runs are:

- bm25_rrf: Reciprocal rank fusion with BM25 on the six variants of the query;
- bm25_rrf_rm3: Reciprocal rank fusion with BM25 on the six variants of the query using pseudo relevance feedback with 10 documents and 10 terms (query weight 0.5);
- bm25_all_rrf: Reciprocal rank fusion with BM25 on all transcripts of the six variants of the query (a total of 18 variants per query)
- bm25_all_rrf_rm3: Reciprocal rank fusion of BM25 with all transcripts using RM3 pseudo relevance feedback.

3.3 Results

The organizers of this task provided the results (averaged across topics) achieved by many baselines compared to the runs of each participant. In Table 3, we show a summary of these results.

A preliminary analysis of the results shows that, in terms of standard evaluation measures such as MAP, Rprec, and bref, the use of the RM3 relevance feedback model improves the effectiveness of the search engine (see Table 3).

For subtask 1, the use of reciprocal ranking together with RM3 produced satisfactory results, in most cases better than any baseline for many performance measures. The run with manual query variants without relevance feedback did not show any significant improvements.

For subtask 2, the use of pseudo relevance feedback achieved better results. It is interesting to see that, despite the noise of the formulation of the query by different participants, Precision@5 (P_5) was better, in general, than most of the baselines.

In terms of understandability (rRBP) and credibility (cRBP) of the retrieved results [12], we report in Table 4 the values of these two measures by cut-off (0.50, 0.50, 0.95) and ordered by map (same ordering of Table 3). From this set of results, one interesting thing emerges: the readability of the Ad-hoc manual query variant seems to improve compared to the runs that use the original query. This will be part of our future work.

Table 3. Summary of the results for subtask 1 and 2. The upper part of the table shows the performances of many baselines (Base). The second and the third part of the table (bottom part) show the performance of our experiments for subtask 1 (Adhoc) and subtask 2 (Spoken).

run	map	Rprec	bpref	recip_rank	P_5
AdHocIR_Base.elastic_BM25f_noqe.out	0.271	0.344	0.421	0.911	**0.808**
AdHocIR_Base.terrier_DirichletLM_noqe.out	0.271	0.357	0.416	0.869	0.736
AdHocIR_Base.terrier_BM25_cli.out	0.264	0.357	0.392	0.760	0.620
AdHocIR_Base.terrier_BM25_gfi.out	0.263	0.357	0.392	0.713	0.628
AdHocIR_Base.terrier_BM25_noqe.out	0.263	0.345	0.396	0.852	0.716
AdHocIR_Base.terrier_TF_IDF_noqe.out	0.261	0.347	0.396	0.854	0.764
AdHocIR_Base.terrier_TF_IDF_qe.out	0.250	0.328	0.380	0.875	0.740
AdHocIR_Base.terrier_BM25_qe.out	0.245	0.323	0.378	0.854	0.704
AdHocIR_Base.elastic_BM25_QE_Rein.txt	0.176	0.252	0.307	0.793	0.684
AdHocIR_Base.terrier_DirichletLM_qe.out	0.145	0.217	0.272	0.878	0.688
AdHocIR_Base.indri_tfidf_noqe.out	0.121	0.209	0.240	0.758	0.600
AdHocIR_Base.indri_okapi_qe.out	0.119	0.204	0.239	0.740	0.604
AdHocIR_Base.indri_tfidf_qe.out	0.119	0.199	0.234	0.685	0.608
AdHocIR_Base.elastic_BM25f_qe.out	0.111	0.163	0.211	0.892	0.720
AdHocIR_Base.indri_okapi_noqe.out	0.110	0.195	0.223	0.786	0.600
AdHocIR_Base.indri_dirichlet_noqe.out	0.079	0.160	0.181	0.748	0.540
AdHocIR_Base.indri_dirichlet_qe.out	0.048	0.110	0.123	0.637	0.436
AdHocIR_Base.Bing_all.txt	0.014	0.017	0.016	0.832	0.632
AdHocIR_IMS.original_rm3_rrf.txt	**0.283**	**0.364**	**0.432**	0.864	0.780
AdHocIR_IMS.original_rrf.txt	0.281	0.362	0.423	**0.916**	0.800
AdHocIR_IMS.bm25_orig.txt	0.248	0.328	0.391	0.888	0.796
AdHocIR_IMS.variant_rrf.txt	0.202	0.288	0.371	0.855	0.744
Spoken_IMS.bm25_rrf_rm3.txt	0.219	0.306	0.404	0.856	0.744
Spoken_IMS.bm25_all_rrf_rm3.txt	0.214	0.304	0.398	0.827	0.700
Spoken_IMS.bm25_rrf.txt	0.196	0.280	0.374	0.854	0.760
Spoken_IMS.bm25_all_rrf.txt	0.195	0.286	0.372	0.841	0.772

4 Final remarks and Future Work

The aim of our participation to the CLEF eHealth Task 2 was to test the effectiveness of the reciprocal ranking fusion approach together with a pseudo-relevance feedback strategy. The initial results show a promising path, but a failure analysis and a topic-by-topic comparison is needed to understand when and how the different combination in the retrieval pipeline are significantly better than simple models.

5 Acknowledgements

This work was partially supported by the ExaMode Project, as a part of the European Union Horizon 2020 Program under Grant 825292.

Table 4. Understandability (rRBP) and Credibility (cRBP) results at different levels of cut-off for each run.

run	rRBP 0.50	rRBP 0.80	rRBP 0.95	cRBP 0.50	cRBP 0.80	cRBP 0.95
AdHocIR_IMS.original_rm3_rrf.txt	0.322	0.314	0.304	0.523	0.504	0.453
AdHocIR_IMS.original_rrf.txt	0.339	0.323	0.302	**0.567**	**0.522**	**0.468**
AdHocIR_IMS.bm25_orig.txt	0.347	0.320	0.292	0.551	0.513	0.448
AdHocIR_IMS.variant_rrf.txt	**0.353**	**0.351**	**0.310**	0.513	0.486	0.414
Spoken_IMS.bm25_rrf_rm3.txt	0.296	0.289	0.250	0.485	0.449	0.381
Spoken_IMS.bm25_all_rrf_rm3.txt	0.289	0.285	0.257	0.469	0.435	0.383
Spoken_IMS.bm25_rrf_rm3.txt	0.296	0.289	0.250	0.506	0.464	0.373
Spoken_IMS.bm25_all_rrf.txt	0.308	0.298	0.248	0.504	0.462	0.372

References

1. Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Trans. Inf. Syst.*, 20(4):357–389, October 2002.
2. Gordon V. Cormack, Charles L A Clarke, and Stefan Buettcher. Reciprocal rank fusion outperforms condorcet and individual rank learning methods. In *Proceedings of the 32nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '09, pages 758–759, New York, NY, USA, 2009. Association for Computing Machinery.
3. D. Frank Hsu and Isak Taksa. Comparing rank and score combination methods for data fusion in information retrieval. *Information Retrieval*, 8(3):449–480, 2005.
4. Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Zhengyang Liu, Gabriella Pasi, Gabriela Saez Gonzales, Marco Viviani, and Chenchen Xu. Overview of the CLEF eHealth 2020 task 2: Consumer health search with ad hoc and spoken queries. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings, 2020.
5. Lorraine Goeuriot, Hanna Suominen, Liadh Kelly, Antonio Miranda-Escalada, Martin Krallinger, Zhengyang Liu, Gabriella Pasi, Gabriela Saez Gonzales, Marco Viviani, and Chenchen Xu. Overview of the CLEF eHealth evaluation lab 2020. In Avi Arampatzis, Evangelos Kanoulas, Theodora Tsikrika, Stefanos Vrochidis, Hideo Joho, Christina Lioma, Carsten Eickhoff, Aurélie Névéol, and Linda Cappellato andNicola Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction: Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020)* , LNCS Volume number: 12260, 2020.
6. Yuanhua Lv and ChengXiang Zhai. Positional relevance model for pseudo-relevance feedback. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '10, pages 579–586, New York, NY, USA, 2010. Association for Computing Machinery.
7. Giorgio Maria Di Nunzio, Federica Beghini, Federica Vezzani, and Geneviève Henrot. An interactive two-dimensional approach to query aspects rewriting in systematic reviews. IMS unipd at CLEF ehealth task 2. In *Working Notes of CLEF 2017 - Conference and Labs of the Evaluation Forum, Dublin, Ireland, September 11-14, 2017.*, 2017.
8. Giorgio Maria Di Nunzio, Giacomo Ciuffreda, and Federica Vezzani. Interactive sampling for systematic reviews. IMS unipd at CLEF 2018 ehealth task 2. In *Working Notes of CLEF 2018 - Conference and Labs of the Evaluation Forum, Avignon, France, September 10-14, 2018.*, 2018.

9. Stephen E. Robertson and Hugo Zaragoza. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval*, 3(4):333–389, 2009.
10. Ian Ruthven and Mounia Lalmas. A survey on the use of relevance feedback for information access systems. *Knowl. Eng. Rev.*, 18(2):95–145, June 2003.
11. Chengxiang Zhai and John Lafferty. A study of smoothing methods for language models applied to ad hoc information retrieval. In *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '01, pages 334–342, New York, NY, USA, 2001. Association for Computing Machinery.
12. Guido Zuccon. Understandability biased evaluation for information retrieval. In Nicola Ferro, Fabio Crestani, Marie-Francine Moens, Josiane Mothe, Fabrizio Silvestri, Giorgio Maria Di Nunzio, Claudia Hauff, and Gianmaria Silvello, editors, *Advances in Information Retrieval*, pages 280–292, Cham, 2016. Springer International Publishing.