LIG-Health at Adhoc and Spoken IR Consumer Health Search: expanding queries using UMLS and FastText

Philippe Mulhem, Gabriela Gonzalez Saez, Aidan Mannion, Didier Schwab, Jibril Frej

LIG-Health at Adhoc and Spoken IR Consumer Health Search: expanding queries using UMLS and FastText.

Philippe Mulhem, Gabriela Gonzalez Saez, Aidan Mannion, Didier Schwab, and Jibril Frej

Univ. Grenoble Alpes, CNRS, Grenoble INP**, LIG, 38000 Grenoble, France

Abstract. This paper describes the work done by the LIG of Grenoble for the Adhoc and the Spoken Consumer Health search. Our focus for this participation is to study the effectiveness of simple query expansions for health related retrieval. We focused on several query expansions, using knowledge-based or embedding-based techniques, with and without weighting of expansions, with and without Pseudo Relevance Feedback. The results obtained for Adhoc queries show that our baseline run outperforms the query expansions proposed. The results obtained for spoken queries show that several speakers lead to very different results, and that merging the results from several users improve the quality of the system.

Keywords: Query Expansion, UMLS, FastText, Query fusion

1 Introduction

This paper describes the experiments achieved by the LIG-Health team for the CLEF 2020 evaluation campaign [7]. We did participate to the task Consumer Health Search of CLEF eHealth 2020 [12], and more spcifically to the adHoc subtask and to the spoken queries subtask [6]. The people involved are for these experiments are members of the Information Retrieval group (MRIM) and the Natural Language Processing group (GETALP) of the Laboratoire d'Informatique de Grenoble[1].

Our work targeted the two subtasks proposed: adhoc and spoken queries. For both subtasks, we explored the use of two query expansions methods: one knowledge-supported using the UMLS meta-thesaurus [2], and one using embeddings using FastText [3]. Binary and weighted expansions were processed in both case. For the retrieval stage, we considered both "Straight" (SR) and Relevance Feedback (RF) cases. We study how some simple processes may be adapted for both text and spoken queries. In the case of spoken queries, query expansions

** Institute of Engineering Univ. Grenoble Alpes

[1] http://liglab.imag.fr

may be questionable because of the possible errors of speech to text steps. In all the cases, we made use of the assessments of CLEF eHealth 2018 to select the submissions.

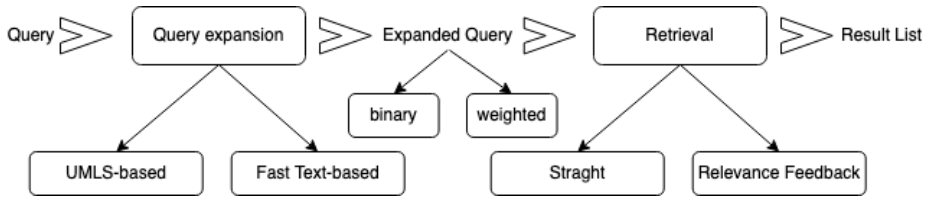


Fig. 1. Overview of Adhoc LIG-Health runs.

We tackled the spoken queries by considering all the transcriptions provided, and applying the two expansions and the two retrieval described above, in a way to evaluate the best configuration to submit. For the fusion of runs, we did consider a simple fusion of result lists.

The remaining of the paper is organized as follows. In Section 2 we describe in detail the two expansions approaches used, before describing our proposal in Section 3. Section 4 focuses on the features and parameters of the Information Retrieval used. The official results are presented in part 6. We discuss the results in Section 7 before concluding in Section 8.

2 Expansion Approaches

2.1 FastText-based

This first expansion proposed relies on FastText [3, 10]. FastText proposes a framework to learn and manage words embeddings. It is able to consider subwords (using ngrams) as opposed to more classical embeddings models like Word2Vec [11], which create embeddings only for whole-word tokens. The FastText embedding vector of a word is the sum of the vectors of its component ngrams.

We used the pre-trained word vectors for English language, trained on Common Crawl and Wikipedia using FastText. The features of the model used are as follows;

- Continuous bag-of-words (CBOW) with positional weighting
- Vector embeddings of dimension $d = 300$
- Character n-grams of length 5
- Context window of size 5
- Sampling of 10 negative examples per positive example

Using such embeddings, in our experiments, we expand each query using terms with a cosine similarity greater than an experimentally determined threshold t with the original query terms - i.e. denoting the cosine similarity function as

FT_{cos}, for each term w in the preprocessed query, we calculate its FastText embedding vector $f(w)$ and then add all terms w' for which $FT_{cos}(f(w), f(w')) \geq t$ to the query.

2.2 UMLS-based

The second expansion strategy used in this work relies on the Unified Medical Language System (UMLS) Metathesaurus [2], a comprehensive biomedical thesaurus incorporating a network of semantically related concepts linking a large number of medical language resources. From the many information sources in UMLS Metathesaurus, we restricted our expansion search to one that is specifically designed to deal consumer-level medical vocabulary: the Open Access and Collaborative Consumer Health Vocabulary[2], known as the CHV, which contains more than 88 000 synonyms for more that 57 000 concepts.

The CHV is used to get the synonyms of query terms, and we denote the function mapping a term to its CHV synonym as CHV_{syn} in the following. As the synonyms were often too general or too numerous in initial experiments, we introduced in addition a filtering step based on FastText similarity FT_{cos}. Given that the goal of the UMLS-based expansion is to find expansion terms that are semantically rather than syntactically related to the query terms, the CHV synonyms were only included in the expanded query if their FastText embedding had a cosine similarity less than 0.6 with the original query term they were associated with.

3 Query expansions proposed

The two query expansions proposed are described now. Each of them has two versions: the *binary* one and the *weighted* one. As their names suggest, the binary expansions do not consider any weight to query terms, and the weighted ones are able to indicate a level of importance of a term in the query. We detail them in the following.

3.1 Embedding-based only expansion.

This approach is quite similar to [1], one major difference is the embeddings consider subwords, as described above in part 2.1. We do not use any manually defined knowledge for these expansions. Formulas 1 and 2 describe the binary expansion based on FastText. In formula 1, the set VOC_FT denotes the vocabulary that FastText manages. The manually-defined threshold considered here, 0.75, is lower than the one in the UMLS expansion: it is consistent with [13] and a trade off between quality of suggested terms and the quantity of terms found.

$$T\_exp\_FT_{binary}(q_i) = \{e | e \in VOC\_FT\ q \wedge FT_{cos}(q_i, e) >= 0.75\} \quad (1)$$

[2] https://www.nlm.nih.gov/research/umls/sourcereleasedocs/current/CHV/index.html

$$Exp\_Query\_FT_{binary}(q) = q \cup_{q_i \in q} T\_exp\_FT_{binary}(q_i) \quad (2)$$

For the weighted FT-based expansion, the principle is the same as before, but:

- the initial query terms have a weight of 1;
- the expanded terms are weighted by the cosine similarity between their FastText embedding and that of their synonym in the original query;
- if one expansion term occurs several times in the expansion, each (weighted) occurrence is considered in the expansion.

More formally, the formulas 3 and 4 describe such expansion :

$$T\_exp\_FT_{weighted}(q_i) = \{(e, FT_{cos}(q_i, e))|e \in VOC\_FT\ q \wedge FT_{cos}(q_i, e) >= 0.75\} \quad (3)$$

$$Exp\_Query\_FT_{weighted}(q) = q \cup_{q_i \in q} T\_exp\_FT_{weighted}(q_i) \quad (4)$$

3.2 UMLS-based expansion.

The use of knowledge-based expansions of query is well studied, as in [8]. In the specific case of medical search, the use of UMLS meta thesaurus is classical, as in [16]. The binary expansion is processed in follows query term by query term q_i, as described in formulas 5 and 6. In our experiments for each query term q_i from a query q, we look for the synonyms of q_i in the consumer health vocabulary. Then, we apply a filtering that keeps the term if his similarity, using FastText [3], is larger that 0.8. Again, this threshold has been manually defined and is consistent with [13] (even if [5] showed that such threshold can not be considered as a rule of thumb). This filtering allows to consolidate the trust we have in the synonyms provided by CHV.

$$T\_exp\_UMLS_{binary}(q_i) = \{e|e \in CHV_{syn}(q_i) \wedge FT_{cos}(q_i, e) >= 0.8\} \quad (5)$$

$$Exp\_Query\_UMLS_{binary}(q) = q \cup_{q_i \in q} T\_exp\_UMLS_{binary}(q_i) \quad (6)$$

For the weighted UMLS-based expansion, the principle is the same as before, but:

- the initial query terms have a weight of 1;
- as CHV does not weight synonymy relationships, we propose that the expanded terms get the weight provided by FastText;
- if one expansion term occurs several times on the expansion, each (weighted) occurrence is considered in the expansion.

More formally, the formulas 7 and 8 describe such expansion :

$$T\_exp\_UMLS_{weighted}(q_i, q) = \{(e, FT_{cos}(q_i, e))|e \in CHV_{syn}(q_i) \backslash q \wedge FT_{cos}(q_i, e)) >= 0.8\} \quad (7)$$

$$Exp\_Query\_UMLS_{weighted}(q) = \{(q_i, 1)|q_i \in q\} \cup_{q_i \in q} T\_exp\_UMLS_{weighted}(q_i) \quad (8)$$

4 Information Retrieval System

The information retrieval system used for the experiments is Terrier v5.2[3] [9]. We did not index the corpus, but we used the index provided by the organizers. This had a impact on the retrieval, as simple tests made us find out that the index seem corrupted, leading to duplicate document identifiers in result lists. We did then post-process the result list in a way to remove these duplicated documents. Because this removal was applied on the top-1000 documents, our results lists are less than 1000 long.

The IR model used is BM25 [14], with b=0.75 after preliminary experiments, other parameters by default. Many experiments show that BM25 is a very good model to be used [15]. The Relevance Feedback model is Bose Einstein (*bo1* model of Terrier), with default parameters (3 top documents considered, and 10 terms for expansion). The Bo1 relevance feedback model provides very good results.

5 Runs description

The different runs submitted were the best four runs of several configurations. As described above in section 3, adding expanded configurations, we get a total 10 runs:

1. Noexp: no expansion, straight query processing (i.e., without Relevance Feedback) ‡;
2. Noexp_RF: no expansion, RF query processing †;
3. FT_Straight_binary: FastText-based query expansion, binary expansion mode, no RF ‡;
4. FT_Straight_weighted: FastText-based weighted query expansion, no RF;
5. FT_RF_binary: FastText-based binary expansion, RF query processing †;
6. FT_RF_weighted: FastText-based weighted query excpansion, RF query processing;
7. UMLS_Straight_binary: UMLS-based query expansion, binary expansion mode, straight query processing ‡;
8. UMLS_Straight_weighted: UMLS-based weighted query expansion, straight query processing;

[3] http://terrier.org/

9. UMLS_RF_binary: UMLS-based binary query excpansion, RF query mode †;
10. UMLS_RF_weighted: UMLS-weighted with RF query processing ‡ †.

A described below, we select among these configurations our submissions for the two subtasks *Adhoc* (marked ‡) and *Spoken queries* (marked †).

5.1 Adhoc subtask

For the selection of our submitted run, we did evaluate the quality on the qrels of CLEF eHealth Adhoc 2018, using the MAP, of the 10 configurations above. The results obtained are presented in Table 1. The best reference run between Noexp and Noexp_RF plus the top three runs were submitted as our official runs. For the Adhoc subtask, dedicated to retrieve documents when asking one query, a set of 50 queries are provided. The runs with the best results over these 50 queries are chosen (marked with a ‡ in section 5).

Table 1. LIG-Health configurations results for the Clef2018 eHealth Adhoc subtask (‡ selected for submission).

Configuration			MAP	selected
expansion	query processing	expansion mode		
Noexp	Straight	/	0.2575	‡
Noexp	RF	/	0.2471	
FT	Straight	binary	0.2239	‡
FT	Straight	weighted	0.2239	
FT	RF	binary	0.2137	
FT	RF	weighted	0.2137	
UMLS	Straight	binary	0.2287	‡
UMLS	Straight	weighted	0.2287	
UMLS	RF	binary	0.2155	
UMLS	RF	weighted	0.2225	‡

From Table 1, we see that, using the CLEF eHealth 2018 reference, the best run is the non-expanded and non-RF one. When binary configurations achieve the same quality that their binary counterpart, we choose the binary configuration. This explain why the UMLS and TF-based binary expansions with straight query processing are selected. Overall, we notice that the Relevance Feedback query processing underperforms straight query processing for the FastText-based expansions, and that the weighted FastText expansions behave the same than their binary counterparts.

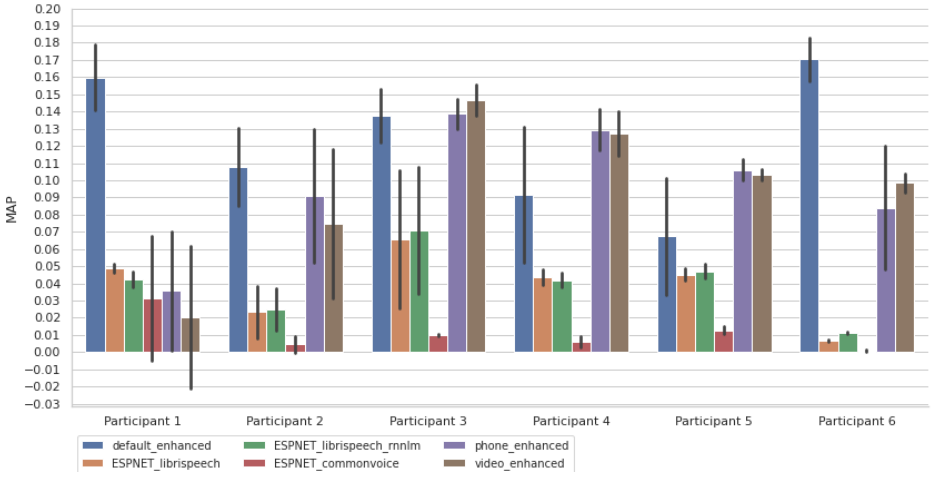
5.2 Spoken subtask

On the Spoken subtask, the 50 topics from the Adhoc task had been recorded by six users (Participant_1 to Participant_6). Per participant, six transcrip-

tions are provided: default_enhanced transcription, ESPNET_commonvoice, ESPNET_librispeech, ESPNET_librispeech_rnnlm, phone_enhanced, video_enhanced. Weexplore all if these transcriptions for each participant. This leads to a total of 36 (= 6 participants × 6 transcriptions) versions of the set of queries.

The full selection of the four submitted runs per user considers the 10 configurations described previously over these versions of queries. It follows two steps: the first one select one transcription per user, and in a second step we choose the configurations used for the submission. More precisely:

Fig. 2. MAP evaluations (with standard deviation error bars) of non-expanded spoken transcriptions per user, wrt. CLEF eHealth Adhoc 2018 assessments.



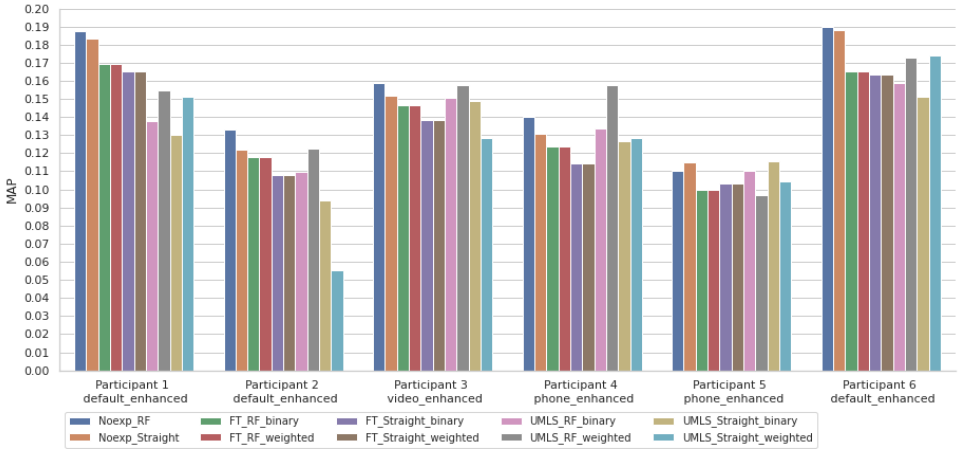
1. **Selection of the one transcription per participant**
   We first choose the transcriptions that achieves the higher MAP values (according to CLEF eHEalth 2018 qrels) over the non-expanded runs. These results are presented in Figure 2. We see that the transcription quality varies a lot depending on the speaker: fort instance the default_enhanced transcription is very good for the participants 1, 2, 3 and 6, but fails for participants 4 and 5. By analyzing this figure, we select the following transcriptions:
   - default_enhanced transcription for Participant 1
   - default_enhanced transcription for Participant 2
   - video_enhanced for Participant 3
   - phone_enhanced for Participant 4
   - phone_enhanced for Participant 5
   - default_enhanced transcription for Participant 6
2. **Selection of four configurations per participant over the chosen transcription of step 1**
   Then, we computed the averaged MAP (over the 6 participants, using CLEF

eHealth 2018 Adhoc assessments provided) for each IR configuration system. These results are presented in Figure 3. Considering these averaged maps, we choose the top-four configurations with only one non-expanded run (marked as † in section 5):

(a) Noexp_RF: no expansion, RF query processing;
(b) FT_RF_binary: FastText-based binary query expansion, RF query processing;
(c) UMLS_RF_binary: UMLS-based binary query expansion, RF query processing
(d) UMLS_RF_weighted: UMLS-based weighted query expansion, RF query processing.

We notice that all the selected configurations use Relevance Feedback.

Fig. 3. MAP evaluations of expanded spoken transcriptions per user for the selected transcriptions, wrt. CLEF eHealth Adhoc 2018 assessments.



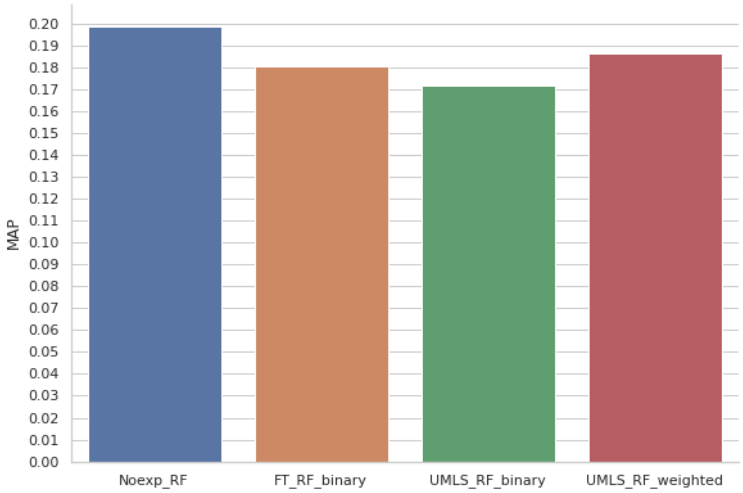
Fused runs

We also submitted 4 runs that fuse the results of the same configuration for each participant. To integrate these results, we used a simple sum of scores. This allows to study if the integration of several transcriptions from several participants outperforms single participant transcriptions. The MAP evaluation of these four configurations on the CLEF eHealth 2018 assessments, as presented in Figure 4, show a slight increase compared to the top results per user.

6 Results

We present here the official results obtained by our runs, for the adhoc and spoken queries subtasks. We consider the following evaluation measures: MAP

Fig. 4. MAP evaluations fused spoken configurations, wrt. CLEF eHealth Adhoc 2018 assessments.



to assess globally the quality of the configurations, Bpref [4] that takes into account the fact that the evaluation relies on incomplete assessments, and the classical ndcg@10 that focuses on the relevance of the top-10 results. For these runs, the other measures provided by the organizers, like RBP-based ones, lead to similar rankings of the configurations tested.

6.1 Adhoc

Our official results for the Adhoc query runs are presented in Table 2. In this table, we see that the best MAP and ndcg@10 results are obtained without any query expansion, and without any relevance feedback. This means that none of the query expansions are able to increase the quality regarding these evaluation measures. However, binary UMLS and TF-based expansions, with straight query processing, slightly outperform the un-expanded runs with straight query processing.

We studied also the results obtained for RBP measures in table3. We see that the UMLS expanded run outperforms the non-expanded one for RBP and RBP readability.

6.2 Spoken queries

The official results of our results per participant are presented in Table 4. This table confirms our evaluations on 2018 assessments: there are large variations of

Table 2. LIG-Health official results for the Adhoc subtask (bests in bold).

Run			MAP	Bpref	ndcg@10
expansion	query processing	expansion mode			
Noexp	straight	/	**0.2627**	0.3640	**0.5919**
UMLS	straight	binary	0.2340	0.3665	0.5769
UMLS	RF	weighted	0.2258	0.3616	0.5918
FT	straight	binary	0.2318	**0.3669**	0.5617

Table 3. LIG-Health RBP official results for the Adhoc subtask (bests in bold).

Run			RBP 0.80	rRBP 0.80	cRBP 0.80
expansion	query processing	expansion mode			
Noexp	straight	/	0.7094	0.2993	**0.4615**
UMLS	straight	binary	0.7058	0.3062	0.4614
UMLS	RF	weighted	**0.7172**	**0.3123**	0.4593
FT	straight	binary	0.6912	0.2909	0.4555

quality (for all the measures) depending on the participant considered. From our four configurations submitted, the best MAP for participant 6 is 0.1744 (Noexp with RF), where for participant 5 the best MAP (UMLS-based binary expansion with RF) is only 0.1036. The best measures (among the 3 presented) per user are obtained without any query expansion in 12 cases on 24. The UMLS-based binary expansion only provides twice the best measures (MAP and Bpref) for participant 5. The weighted expansion using UMLS outperforms other configurations in 4 cases. The weighted UMLS expansion outperforms the binary UMLS expansion on 12 cases over 18. FT-based expansion never produces the best results, but the second best Bpref and ndcg@10 for participant 1.

For the submitted merged results, we see that the merging always increase (over the best single participant) the evaluations measures for each configuration. Here, the binary UMLS-based expansion outperforms its weighted counterpart for Bpref and ndcg@10. FT-based expansions underperform UMLS-based expansions.

For the RBP evaluations of the merged spoken runs, the no-expanded run still outperforms our other submissions.

7 Discussion

We focus first on the Adhoc subtask. According to the MAP values obtained on the 2018 assessments, our official results are consistent: the best run is the non-expanded one without relevance feedback. We see that the UMLS expanded query with relevance feedback performs as well as the non-expanded run for first results, as the P@10 and ndcg@10 are almost equal. Binary UMLS and FT-based expansions slightly outperform our non-expanded runs for the Bpref measure, this shows that such expansions can be beneficial. We see in Figures 5 and 6

Table 4. LIG-Health official per participant results for the spoken queries subtask (best per user in bold).

Participant	Run			MAP	Bpref	ndcg@10
	expansion	query processing	expansion mode			
1	Noexp	RF	/	**0.1726**	**0.3192**	**0.5178**
	UMLS	RF	binary	0.1271	0.2783	0.4540
	UMLS	RF	weighted	0.1416	0.2928	0.4628
	FT	RF	binary	0.1565	0.3096	0.5179
2	Noexp	RF	/	**0,1206**	**0.2634**	0,4247
	UMLS	RF	binary	0.1017	0.2384	0.3963
	UMLS	RF	weighted	0.1133	0.2575	**0.4419**
	FT	RF	binary	0.0995	0.2314	0.3885
3	Noexp	RF	/	0.1447	0.3023	0.4583
	UMLS	RF	binary	0.1385	0.2984	0.4217
	UMLS	RF	weighted	**0.1485**	**0.3114**	**0.4605**
	FT	RF	binary	0.1274	0.2664	0.4290
4	Noexp	RF	/	**0.1301**	**0.2880**	**0.4310**
	UMLS	RF	binary	0,1246	0.2852	0,4042
	UMLS	RF	weighted	0,1282	0.2877	0.4273
	FT	RF	binary	0.1090	0.2582	0.3805
5	Noexp	RF	/	0.1035	0.2412	**0.3539**
	UMLS	RF	binary	**0.1036**	**0.2470**	0.3462
	UMLS	RF	weighted	0.0917	0.2227	0.3275
	FT	RF	binary	0.0952	0.2287	0.3097
6	Noexp	RF	/	**0.1744**	**0.3238**	**0.4807**
	UMLS	RF	binary	0.1478	0.3019	0.4355
	UMLS	RF	weighted	0.1594	0.3072	0.4439
	FT	RF	binary	0.1509	0.2921	0.4468

Table 5. LIG-Health official results for the spoken queries subtask - Merged runs. Best results in bold.

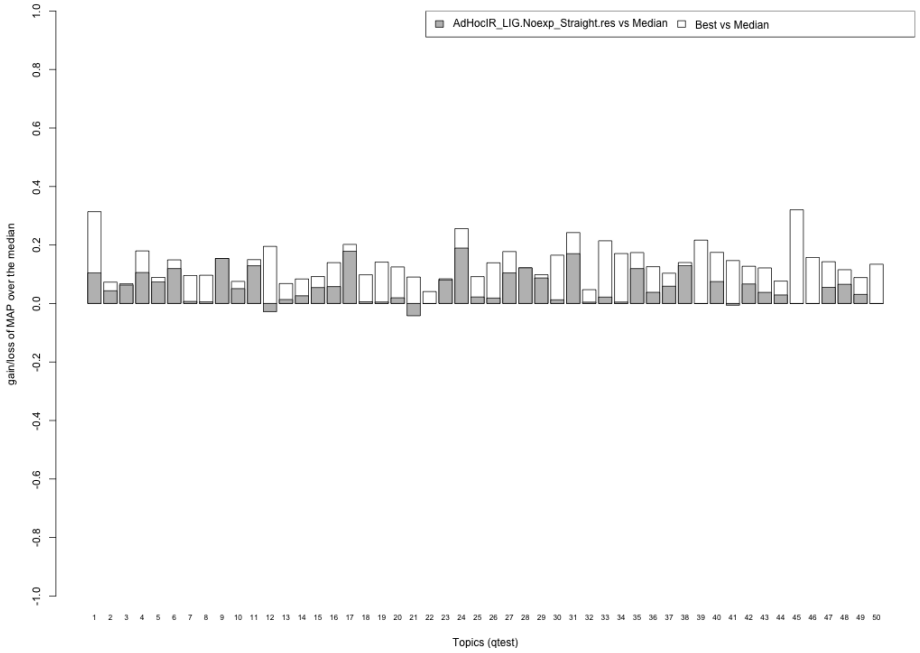
Run			MAP	Bpref	ndcg@10
expansion	query mode	expansion mode			
Noexp	RF	/	**0.1810**	**0.3279**	**0.5411**
UMLS	RF	binary	0.1582	0.3085	0.5203
UMLS	RF	weighted	0.1671	0.2964	0.5203
FT	RF	binary	0.1626	0.3054	0.4873

Table 6. LIG-Health RBP official results for the spoken queries subtask - Merged runs. Best results in bold.

Run			RBP 0.80	rRBP 0.80	cRBP 0.80
expansion	query mode	expansion mode			
Noexp	RF	/	**0.6186**	**0.2601**	**0.4067**
UMLS	RF	binary	0.6017	0.2557	0.3864
UMLS	RF	weighted	0.5847	0.2407	0.3722
FT	RF	binary	0.5629	0.2264	0.3592

(i.e, our two best results according to MAP), that the expansion is much more unstable compared to the media results.

Fig. 5. MAP evaluations per query for Adhoc Straight without expansion. (Image courtesy of the CLEF eHealth 2020 Task 2 organizers.)

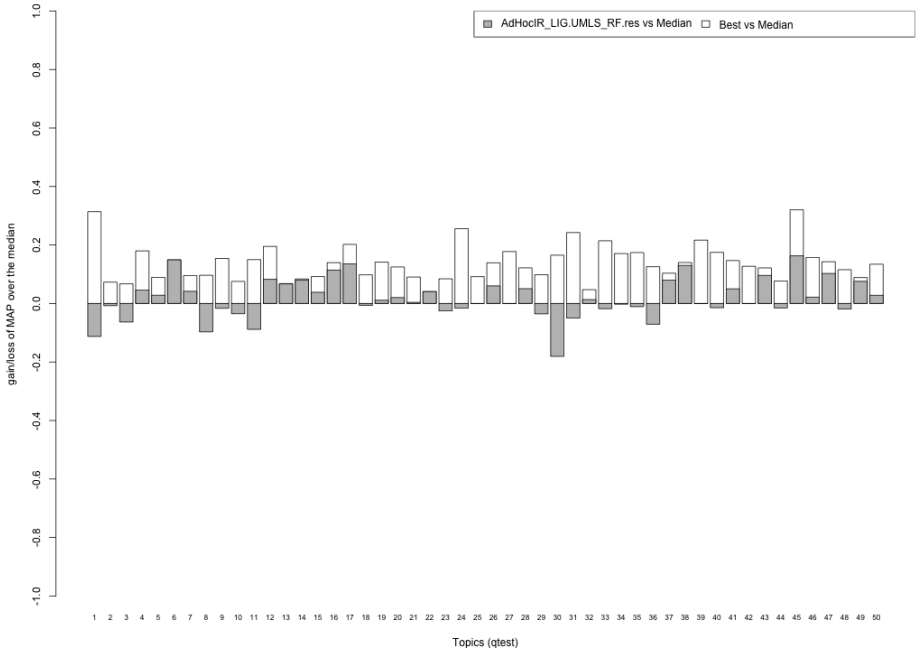


When considering spoken participant runs, we show again that some expansions proposed are able to outperform the non-expansions configurations. More precisely, UMLS-based expansions obtain larger MAP and Bpref values than non-expansions for 33% (= 2/6) of the participants.

The fused spoken evaluation results that we get from the spoken queries are consistent with the adhoc results: the expansions underperform the non-expanded runs. With merged results, the expansions are never close to the quality of the non-expanded runs: the reason is that the expansions over each user tend to disperse the initial query expression already subject to transcription errors. We present in 7 and 8 (i.e, our two best results according to ndcg@10) the results per query. We see again here that the expansion underperforms the median results more often than the non-expanded run.

A more detailed fusion process may improve the overall quality, but in any case merging results for spoken queries needs to be able to retrieve similar queries asked by several users, which is not an easy task.

Fig. 6. MAP evaluations per query for UMLS binary expanded with Relevance Feedback. (Image courtesy of the CLEF eHealth 2020 Task 2 organizers.)



This work is only focusing on simple expansions, and these expansions do not succeed in increasing the quality of the results. The expansion terms are not strongly enough related to teh initial query. Future experiments will be conducted to check exactly why these expansions fail.

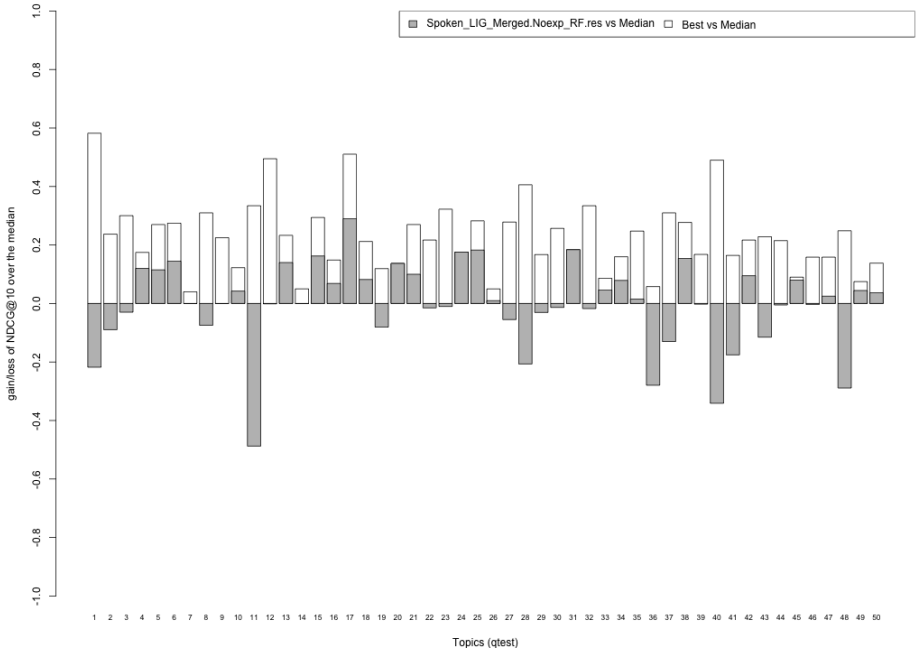
For the official evaluation measures related to credibility, the UMLS expanded runs outperform by 4.3% (cRBP 0.80) the non-expanded one for the Adhoc search, but still the non-expanded runs achieve a higher quality than the expanded ones for the spoken runs.

8 Conclusion

We presented in this paper the configurations of the retrieval for Adhoc and Spoken queries subtasks of the consumer Health Search task from CLEF eHealth 2020. We focused our proposal on several query expansions. The query expansions rely on UMLS meta-thasaurus, and on words embeddings using FastText.

The main findings is that the expansion proposed for classical Adhoc underperforms simple retrieval (with or without relevance Feedback strategy). For spoken runs, we were able to detect that some query expansions (based on UMLS) do compete well with simple retrieval without expansion.

Fig. 7. ndcg@10 evaluation per query for Merged run, with Relevance Feedback, without expansion. (Image courtesy of the CLEF eHealth 2020 Task 2 organizers.)



Other approaches based on reranking should be studied in the future, in a way to avoid the noise generated by the expansions of queries.

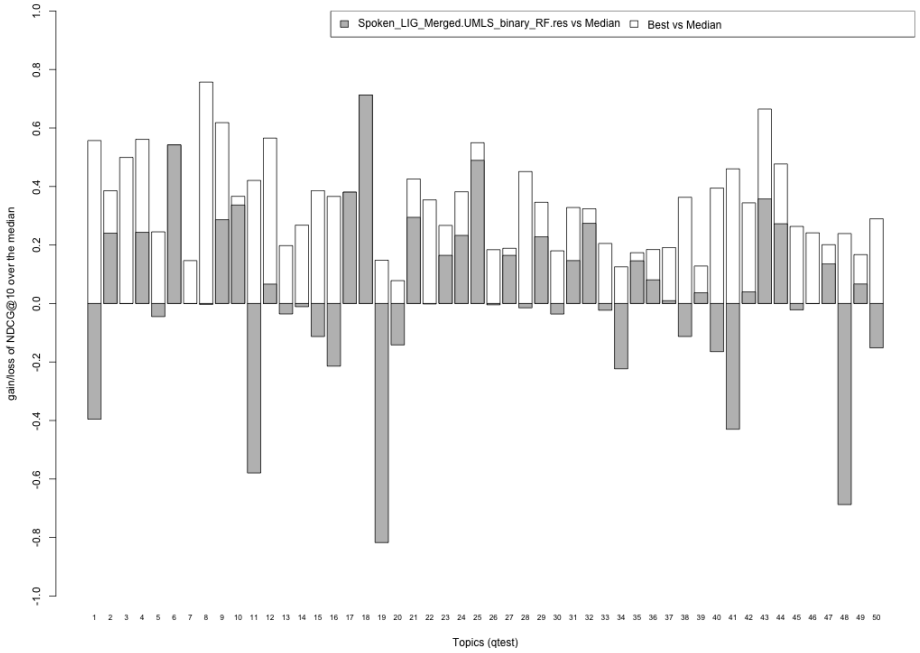
Acknowledgement

This work was partially supported by the ANR Kodicare project, grant ANR-19-CE23-0029 of the French Agence Nationale de la Recherche.

References

1. Mohannad Almasri, Catherine Berrut, and Jean-Pierre Chevallet. A Comparison of Deep Learning Based Query Expansion with Pseudo-Relevance Feedback and Mutual Information. In *Conférence ECIR*, volume 42, pages 369 – 715, Padoue, Italy, March 2016.
2. Olivier Bodenreider. The Unified Medical Language System (UMLS): Integrating Biomedical Terminology. *Nucleic Acids Res.*, 32(Database-Issue):267–270, 2004.
3. Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching Word Vectors with Subword Information. *CoRR*, abs/1607.04606, 2016.
4. Chris Buckley and Ellen M. Voorhees. Retrieval Evaluation with Incomplete Information. In *Proceedings of the 27th Annual International ACM SIGIR Conference*

Fig. 8. ndcg@10 evaluation per query for Merged run, with UMLS binary query expansion, with Relevance Feedback. (Image courtesy of the CLEF eHealth 2020 Task 2 organizers.)



on Research and Development in Information Retrieval, SIGIR '04, page 25–32, New York, NY, USA, 2004. Association for Computing Machinery.

5. A. Elekes, M. Schaeler, and K. Boehm. On the Various Semantics of Similarity in Word Embedding Models. In *2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL)*, pages 1–10, 2017.
6. 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.
7. 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, Linda Cappellato, and Nicola 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.
8. Alexander Kotov and ChengXiang Zhai. Tapping into Knowledge Base for Concept Feedback: Leveraging Conceptnet to Improve Search Results for Difficult

Queries. In Eytan Adar, Jaime Teevan, Eugene Agichtein, and Yoelle Maarek, editors, *WSDM*, pages 403–412. ACM, 2012.
9. Craig Macdonald, Richard McCreadie, Rodrygo LT Santos, and Iadh Ounis. From Puppy to Maturity: Experiences in Developing Terrier. *Proc. of OSIR at SIGIR*, pages 60–63, 2012.
10. Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. Advances in Pre-Training Distributed Word Representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*, 2018.
11. Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed Representations of Words and Phrases and their Compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.
12. Antonio Miranda-Escalada, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, and Martin Krallinger. Overview of Automatic Clinical Coding: Annotations, Guidelines, and Solutions for non-English Clinical Cases at CodiEsp Track of CLEF eHealth 2020. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum*, CEUR Workshop Proceedings, 2020.
13. Navid Rekabsaz, Mihai Lupu, and Allan Hanbury. Exploration of a Threshold for Similarity Based on Uncertainty in Word Embedding. In Joemon M Jose, Claudia Hauff, Ismail Sengor Altıngovde, Dawei Song, Dyaa Albakour, Stuart Watt, and John Tait, editors, *Advances in Information Retrieval*, pages 396–409, Cham, 2017. Springer International Publishing.
14. Stephen Robertson, S. Walker, S. Jones, M. M. Hancock-Beaulieu, and M. Gatford. Okapi at trec-3. In *Overview of the Third Text REtrieval Conference (TREC-3)*, pages 109–126. Gaithersburg, MD: NIST, January 1995.
15. Peilin Yang and Hui Fang. A Reproducibility Study of Information Retrieval Models. ICTIR '16, page 77–86, New York, NY, USA, 2016. Association for Computing Machinery.
16. Liu Zhenyu and Chu Wesley W. Knowledge-based Query Expansion to Support Scenario-specific Retrieval of Medical Free Text. *Information Retrieval*, 10(2):173–202, 2007.