BIT.UA@TREC 2020 Precision Medicine Track

Tiago Almeida[0000−0002−4258−3350] and Sérgio Matos[0000−0003−1941−3983]

DETI/IEETA, University of Aveiro, 3810-193 Aveiro, Portugal
{tiagomeloalmeida,aleixomatos}@ua.pt

Abstract. The TREC Precision Medicine and the previous CDS track have produced a variety of approaches on document retrieval to support clinical decisions. To the best of our knowledge, the top-performing approaches always relied on traditional information retrieval solutions, such as BM25 with query expansion, to find the biomedical articles relevant to the given topics. Although deep learning solutions have been tried, these struggle to reach the top scores on this particular task.
To further explore and assess the effectiveness of deep learning methods in the PM retrieval task, we reformulate this relevance problem of evidence finding as a question-answering problem, where a query is formulated with the topic information and a neural information retrieval model generates a ranked list of documents. More precisely, we adopted a two-stage retrieval pipeline, where we first reduce the searching space using BM25 with gene name expansion and then apply a lightweight neural IR model, with only 620 trainable parameters, that was previously trained and tested on the BioASQ challenge.
In terms of overall performance, in phase 1 we achieved the overall best score of 0.5325 nDCG, ten percentage points above the reported median. In phase 2 we achieved a best score of 0.3289 nDCG@30, four percentage points above the reported median. Our source code is available from `https://github.com/T-Almeida/TREC-PM-2020`.

Keywords: Deep Learning · Neural Networks · Information Retrieval · BM25 · Lightweight Neural Model.

1 Introduction

This paper describes the participation, for the first time, of the Biomedical Informatics and Technologies (BIT) group from the University of Aveiro, Portugal, in the TREC Precision Medicine track.

The precision medicine track has an already extensive literature describing a wide variety of solutions, where the great majority of top-performing runs are usually based on traditional information retrieval solutions, like BM25 with query expansions, being even capable of outperforming a majority of deep learning (DL) solutions. Although DL approaches are currently less effective in this track, some groups have presented interesting results [5,6].

In our approach, we tried to adopt a deep learning solution by reformulating the TREC-PM problem as a well studied DL task, i.e, we reformulate the

problem of finding evidence as a question-answering problem. More precisely, we followed a two-stage pipeline, where, following the literature, we adopted BM25 [8] with query expansion for creating strong baselines as our first-stage step. Next, we employed a lightweight shallow interaction-based neural network [1], with only 620 trainable parameters, to act as a question-answering reranker, hoping to improve the rank of our strong baseline.

In the remainder of the paper, we describe our methodology for the construction of the submitted runs, then address the results that were obtained, followed by a conclusion section.

2 Methodology

Figure 1 presents an overview of our employed system architecture. As already mentioned, we first construct a strong baseline using the BM25 [8] ranking function combined with a list of gene synonyms. Then we reformulate the topic as a natural language question to be fed to our neural interaction model further refining the previous ranking order. Moreover, our neural model was pretrained on the BioASQ [9] training data, with the intuition of exploring some domain similarity between the two tasks, since BioASQ is a biomedical challenge over the same Pubmed/MEDLINE corpus. Through this section, we describe the most important steps that comprise our submissions.

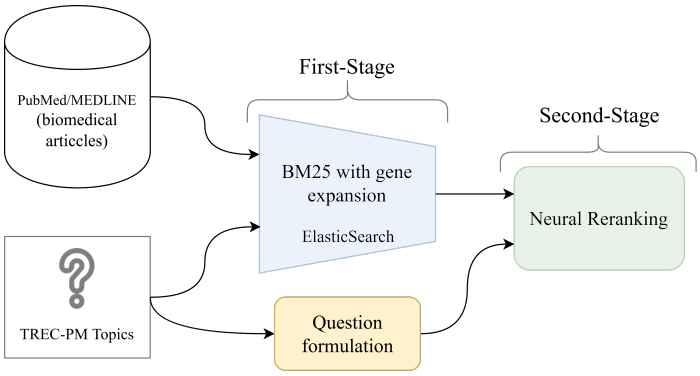

Fig. 1: General system architecture.

2.1 Synonyms

Given the success of employing synonyms described in the literature, we also adopted this solution building a synonyms list for the genes and other for drugs, with the second one being related to the *treatment* field on the topics.

Human gene names and synonyms were obtained from HUGO Gene Nomenclature Committee (HGNC), and expanded with other gene and protein synonyms from Enseml, UniProt and NCBI. Additionally, ortholog gene names were also included to increase recall.

In a final step we ended up discarding the synonyms list for the drugs since it did not match any drug present on the *treatment* topic field.

2.2 Question formulation

This is our key step to change into a question-answering alike problem. The objective is to use the information need described on the topic and transforming it into a natural language question that can be understood by a neural retrieval model.

Due to time constraints, we were unable to explore this procedure exhaustively and followed a heuristic approach to build a template question that all the topics would be converted to. For that, we looked at the types of question present on the BioASQ data and noticed the following frequent pattern: "Is ***treatment*** effective for treatment of ***disease***?", e.g., "Is dasatinib effective for treatment of glioblastoma?". We then decided to use this as our base for the template since it encodes two of the fields present on the topics and it is a template that the model sees during training.

So, this step produces a natural question following the template: "Is ***T*** effective for treatment of ***D G***?", where T is the treatment, D is the disease and G is the gene specification of the disease.

2.3 First-Stage: BM25 with synonyms

We relied on the BM25 implementation in ElasticSearch, which also offers a mechanism to perform query expansion with synonyms. More precisely, we added a Synonyms Token Filter[1] to the ElasticSearch "english" tokenizer.

Moreover, we also finetuned the BM25 hyper-parameters using the TREC-PM 2019 test data, despite addressing slightly different problems.

2.4 Second-Stage: Neural reranking model

The adopted neural model was original proposed in [2], and further improved in [1]. Figure 2 describes the overall architecture, where we employed an interaction network to learn and pool relevant signals and an aggregation network to weigh all the evidence found on different document passages.

In a more detailed way, each document is split into sentences that are further combined with the query to build interaction matrices. Then, taking these matrices in the interaction network, 3-by-3 convolutions are adopted to learn

[1] https://www.elastic.co/guide/en/elasticsearch/reference/current/analysis-synonym-tokenfilter.html

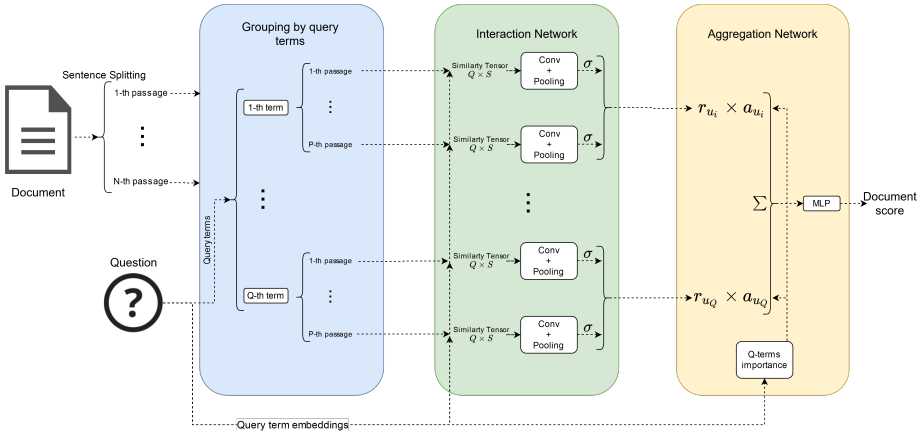

Fig. 2: Lightweight neural model data and operation flow.

n-gram patterns that are then extracted by pooling operations, lastly, the resulting feature vector is linearly combined with a trainable vector to compute a sentence relevant signal, with 0 for irrelevant and 1 for relevant. Next, the job of the aggregation network is to weigh the importance of each sentence in order to produce the final document score. For that, we follow the heuristic to first weigh each sentence by the importance of each query term, as suggested in [4], where this importance is learned by taking into consideration the embedding representation of the query term.

More importantly, this model inner-working follows some of the best-reported ideas from shallow interaction-based models resulting in a completely transform-free architecture. As intuition, it was designed to weigh the importance of the document sentences by taking into consideration the context where the exact match with the query terms occurs. In other words, this model produces a more refined judgment of the previously exact match signal considered in the first stage of the pipeline.

Data preparation. We built a regex-based tokenizer and trained 200 dimension word embeddings for the produced vocabulary. The tokenizer consisted of filtering off non-alphanumeric characters with the exception of the hyphen character since in this biomedical domain this character composes more complex words like chemical compositions. We ran this tokenizer over the PubMed/Medline documents collection and the development and test topics, resulting in a vocabulary with approximately 4 million tokens.

We used the word2vec skip-gram algorithm from the Gensim [7] library to obtain the word embeddings. Specifically, we adopted the default configuration present on the Gensim library for training with word2vec skip-gram.

2.5 Training and hardware

The neural model was pretrained using a pairwise cross-entropy loss on the BioASQ training collection. We also tried to further train on the TREC 2019 test data. The model architecture parameters were the same as used in [1], so we redirect the reader for further information or details.

Additionally, our experiments ran on a machine with Intel(R) Xeon(R) CPU E5-2670 v3 @ 2.30GHz and 192GB of RAM, highlighting that the neural model only ran on the CPU not requiring a GPU.

2.6 Runs identification

The TREC Precision Medicine track allowed five submissions that we utilized in the following way:

- **baseline**: Corresponds to the first-stage ranking order, produced with the BM25 with gene synonyms and the all the topic fields concatenated.
- **nnrun1**: Corresponds to the full pipeline presented.
- **nnrun2**: Similar to nnrun1, but uses a different checkpoint of the trained neural model.
- **nnrun3**: Similar to nnrun1, but training the model also on the TREC 2019 test data.
- **rrf**: Reciprocal rank fusion [3] of the previous four runs.

3 Results and discussion

In this section, we will address the results obtained in phase 1, precision medicine assessment, and phase 2, evidence assessment.

3.1 Phase 1

First, we present the results of phase 1 of our runs and compare it to the median, as summarized in Table 1.

Table 1: Summary of our runs results for the phase 1 comparatively to the TREC average of the median.

Submissions	nDCG	P@10	R-prec
baseline	0.5325	0.5516	0.4207
nnrun1	0.5145	0.5097	0.4027
nnrun2	0.5071	0.5161	0.4029
nnrun3	0.4652	0.4903	0.3604
rrf	0.5245	0.5161	0.4107
Median	0.4316	0.4645	0.3259

In general, all of our submissions were able to achieve scores higher than the presented median. However, when looking with more detail it was our baseline run that achieved the higher scores in all the metrics, which means that our current neural solution injures the original rank instead of improving its results. Nevertheless, the produced ordering stayed above the median which is a good indication since we believe that the neural model could be further improved. Only regarding the neural submissions, the *nnrun1* was the best in terms of nDCG@10 losing to the *nnrun2* on the remaining two metrics.

Our weakest submission was *nnrun3*, which shows that training with the TREC 2019 test data ended up injuring the model. On the other hand, the other neural runs were capable of staying closer to the already strong baseline. This also demonstrates that training with the BioASQ data seemed to be more effective than training with last year's TREC data, which may be due to this year's task being slightly different.

Per topic analysis. We now present two visualizations to look with more detail at the individual query performance of our submitted runs.

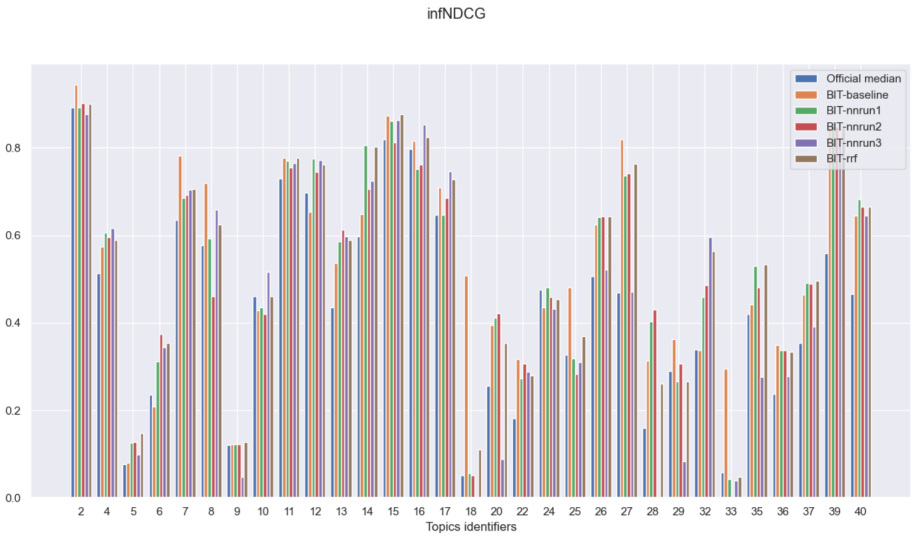

Fig. 3: nDCG of all the submitted runs in comparison to the official median.

In Figure 3 we show the performance in terms of normalized discounted cumulative gain per each topic for all the submitted runs comparatively to the official TREC median. It is observable that for a majority of topics our submissions were able to match the official median, and for topics 18 and 33 our baseline run clearly outperforms the median and also our remaining runs.

We also compare, in Figure 4, our best run with the official median and best values, in terms of nDCG, where we achieved the same scores as the best

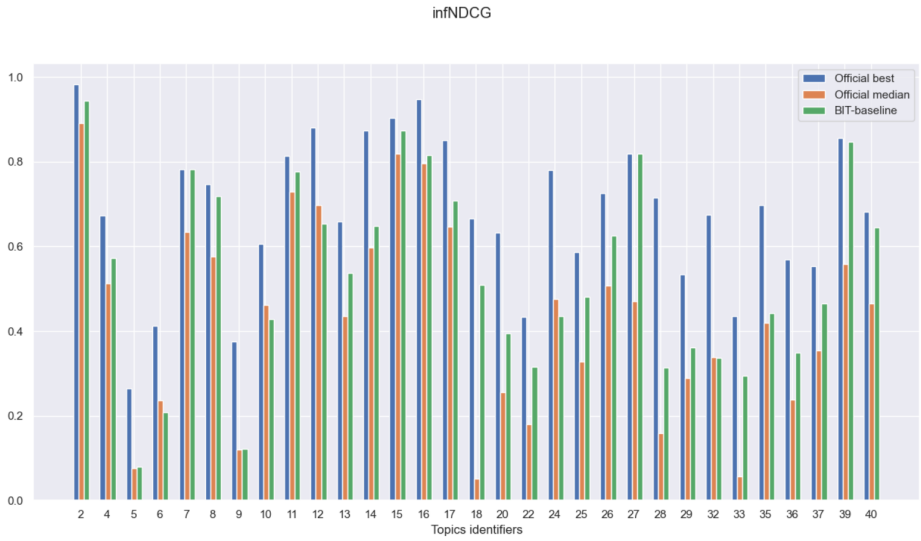

Fig. 4: nDCG of our best run against the official median and best values for each topic.

reported for topics 7 and 27, while being close for a great part of the remaining topics.

In Appendix A we also present the same visualization for the other available evaluation metrics, namely P@10 and R-prec.

3.2 Phase 2

Regarding phase 2, we present the results of our runs on Table 2, in comparison to the median.

Table 2: Summary of our runs results for the phase 2 comparatively to the TREC average of the median.

Submissions	nDCG@30	nDCG@5
baseline	0.3092	0.2720
nnrun1	0.3266	0.2964
nnrun2	0.3214	0.2944
nnrun3	0.3030	0.2683
rrf	0.3289	0.2911
Median	0.2857	0.2529

Similarly to phase 1, in phase 2 all of our submissions achieved scores higher than the presented median, for both metrics. More interestingly, the neural reranking runs were able to outperform the baseline run, except for the *nnrun3*. This observation reinforces our hypothesis of converting the problem of finding evidence to the more well-studied question-answering problem.

Fig. 5: nDCG@30 of all the submitted runs in comparison to the official median.

Per topic analysis. In Figure 5 we show the performance in terms of nDCG@30 per each topic for all the submitted runs comparatively to the official TREC median. In phase 2, the results were closer to the median, comparatively to phase 1, being only the topics 24, 34 and 39 where our system considerably outperforms the median.

Similarly to the previous phase, we also compare, in Figure 6, our best run with the official median and the best values, in terms of nDCG@30, where we achieved close to the best reported score for topic 39, while being superior to the median for a great part of the remaining topics.

In Appendix A we present the same visualization for the other available evaluation metrics, namely nDCG@5.

4 Conclusion

We proposed a two-stage retrieval pipeline to address the Precision Medicine challenge, where we demonstrate a strong baseline comparatively to the TREC best and median scores, which will be extremely useful as a starting point for future work, leaving more time to focus on the reranking portion of our system, which failed to achieve the expected improvements.

We also demonstrate a successful naive neural approach to extract evidence by reformulating this as a question-answering problem.

In future work, we aim to improve the question formulation step and explore transform models, like BERT, for the reranking stage.

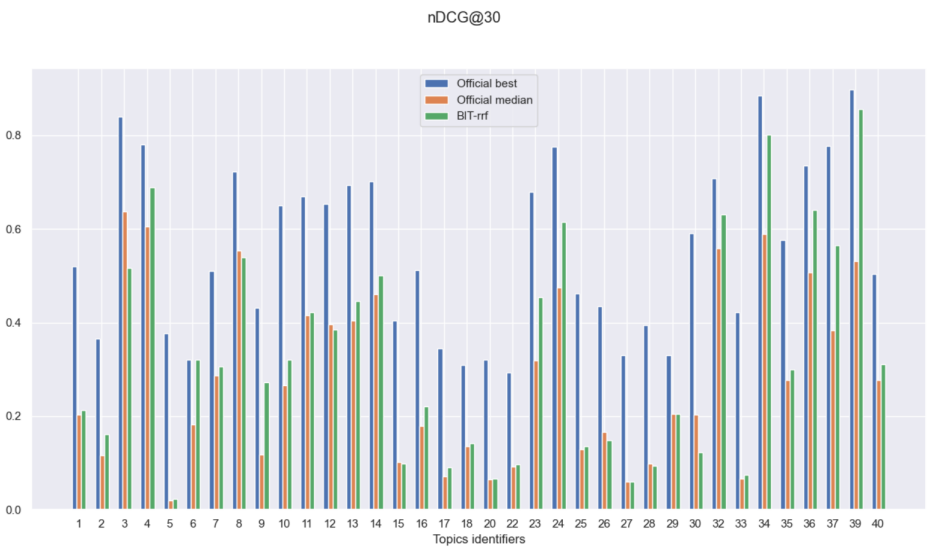

Fig. 6: nDCG of our best run against the official median and best values for each topic.

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 project UIDB/00127/2020.

References

1. Almeida, T., Matos, S.: BIT.UA at bioasq 8: Lightweight neural document ranking with zero-shot snippet retrieval. In: Cappellato, L., Eickhoff, C., Ferro, N., Névéol, A. (eds.) Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020. CEUR Workshop Proceedings, vol. 2696. CEUR-WS.org (2020), `http://ceur-ws.org/Vol-2696/paper_161.pdf`
2. Almeida, T., Matos, S.: Calling attention to passages for biomedical question answering. In: Jose, J.M., Yilmaz, E., Magalhães, J., Castells, P., Ferro, N., Silva, M.J., Martins, F. (eds.) Advances in Information Retrieval. pp. 69–77. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_9
3. Cormack, G.V., Clarke, C.L.A., Buettcher, S.: 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. p. 758–759. SIGIR '09, Association for Computing Machinery, New York, NY, USA (2009). https://doi.org/10.1145/1571941.1572114, `https://doi.org/10.1145/1571941.1572114`
4. Guo, J., Fan, Y., Ai, Q., Croft, W.B.: A deep relevance matching model for ad-hoc retrieval. Proceedings of the 25th ACM International on Conference on Information and Knowledge Management (Oct 2016). https://doi.org/10.1145/2983323.2983769

5. Jo, S., Choi, W., Lee, K.: CBNU at TREC 2018 precision medicine track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the Twenty-Seventh Text REtrieval Conference, TREC 2018, Gaithersburg, Maryland, USA, November 14-16, 2018. NIST Special Publication, vol. 500-331. National Institute of Standards and Technology (NIST) (2018), https://trec.nist.gov/pubs/trec27/papers/cbnu-PM.pdf
6. Liu, X., Li, L., Yang, Z., Dong, S.: SCUT-CCNL at TREC 2019 precision medicine track. In: Voorhees, E.M., Ellis, A. (eds.) Proceedings of the Twenty-Eighth Text REtrieval Conference, TREC 2019, Gaithersburg, Maryland, USA, November 13-15, 2019. NIST Special Publication, vol. 1250. National Institute of Standards and Technology (NIST) (2019), https://trec.nist.gov/pubs/trec28/papers/CCNL.PM.pdf
7. Řehůřek, R., Sojka, P.: Software Framework for Topic Modelling with Large Corpora. In: Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks. pp. 45–50. ELRA, Valletta, Malta (May 2010), http://is.muni.cz/publication/884893/en
8. Robertson, S., Zaragoza, H.: The probabilistic relevance framework: Bm25 and beyond. Found. Trends Inf. Retr. **3**(4), 333–389 (Apr 2009). https://doi.org/10.1561/1500000019
9. Tsatsaronis, G., Balikas, G., Malakasiotis, P., Partalas, I., Zschunke, M., Alvers, M., Weißenborn, D., Krithara, A., Petridis, S., Polychronopoulos, D., Almirantis, Y., Pavlopoulos, J., Baskiotis, N., Gallinari, P., Artieres, T., Ngonga Ngomo, A.C., Heino, N., Gaussier, E., Barrio-Alvers, L., Paliouras, G.: An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. BMC Bioinformatics **16**, 138 (04 2015). https://doi.org/10.1186/s12859-015-0564-6

A Remaining visualisation

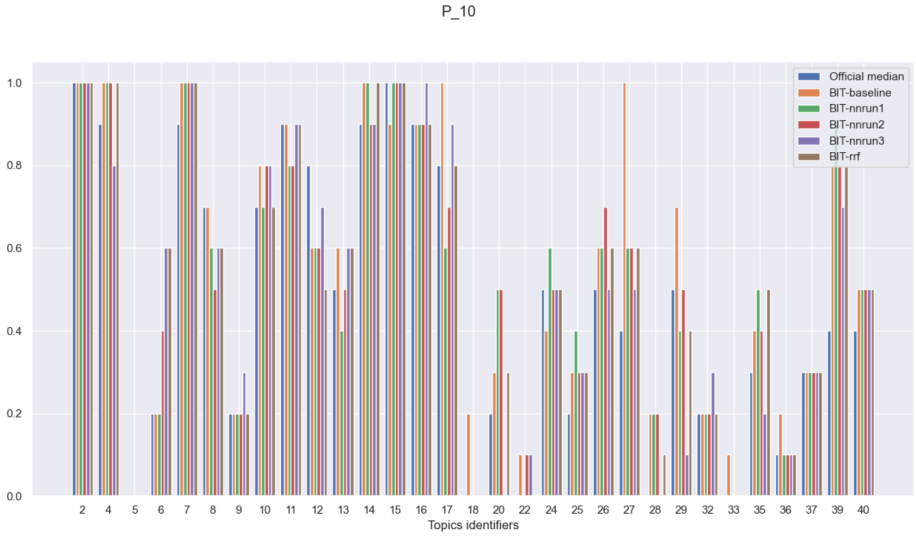

Fig. B1: Phase 1: P@10 of all the submitted runs comparable to the official median.

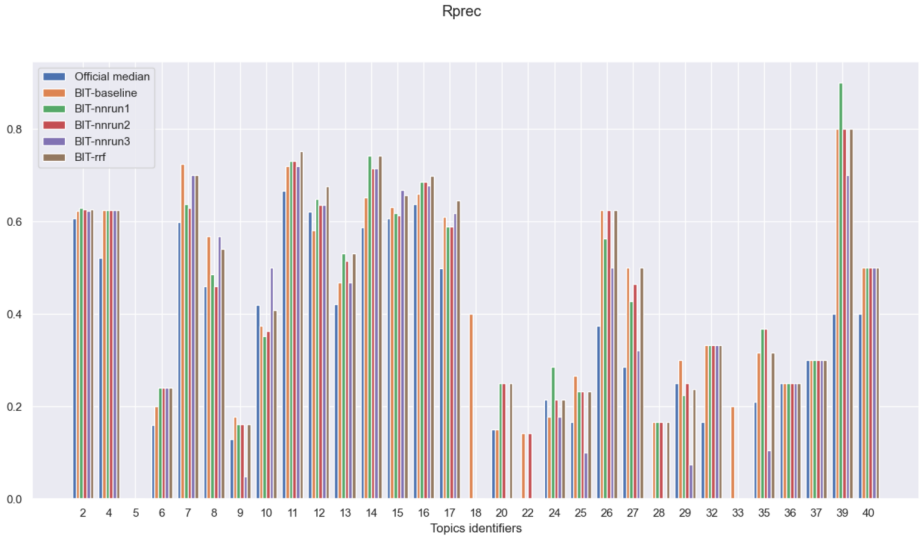

Fig. B2: Phase 1: R-prec of all the submitted runs comparable to the official median.

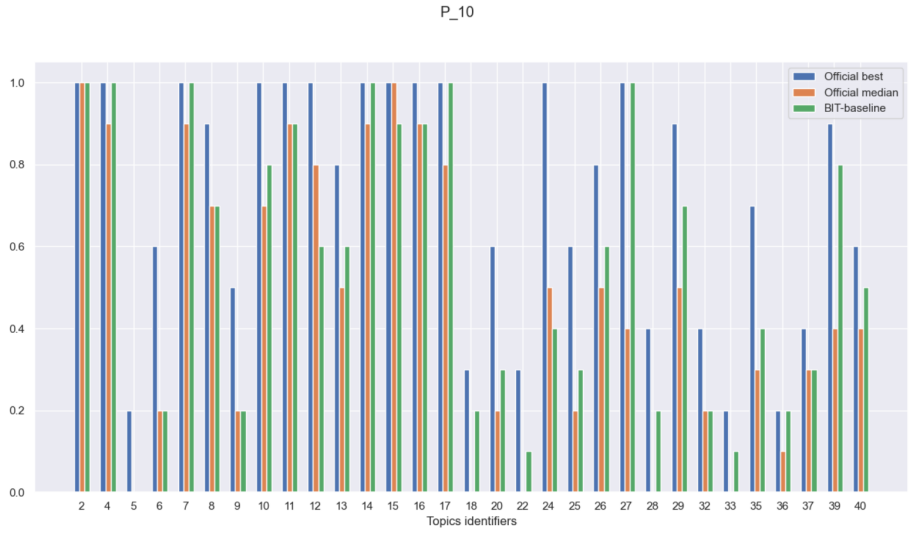

Fig. B3: Phase 1: P@10 of our best run against the official median and best values for each topic.

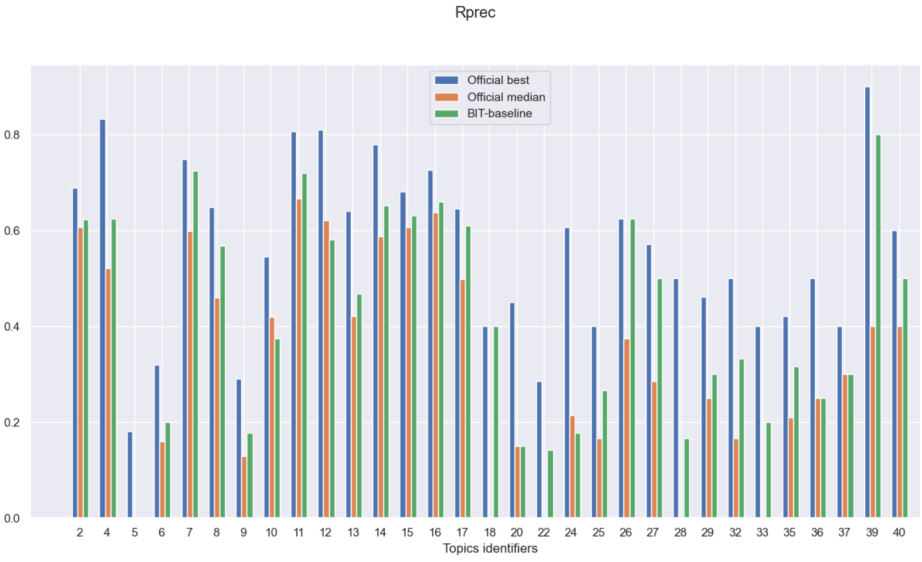

Fig. B4: Phase 1: R-prec of our best run against the official median and best values for each topic.

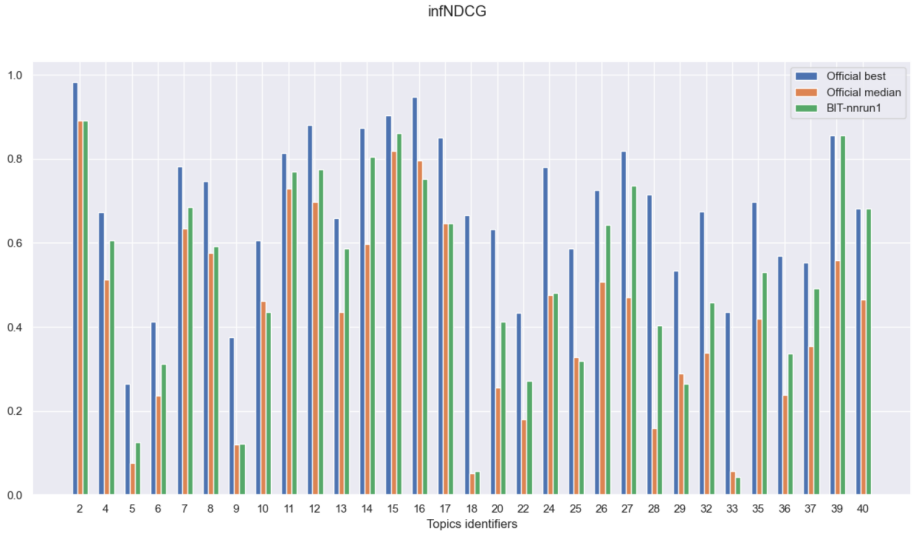

Fig. B5: nDCG of our best **neural** run against the official median and best values for each topic.

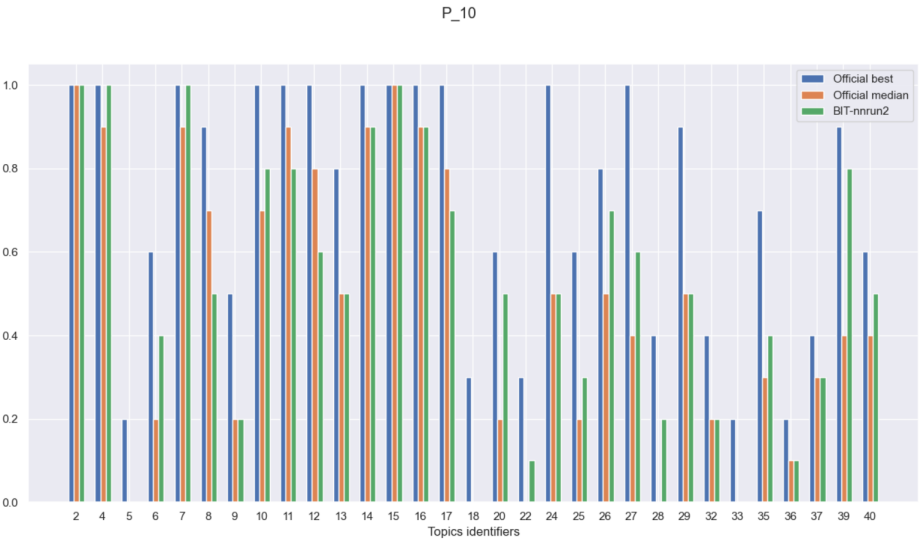

Fig. B6: Phase 1: P@10 of our best **neural** run against the official median and best values for each topic.

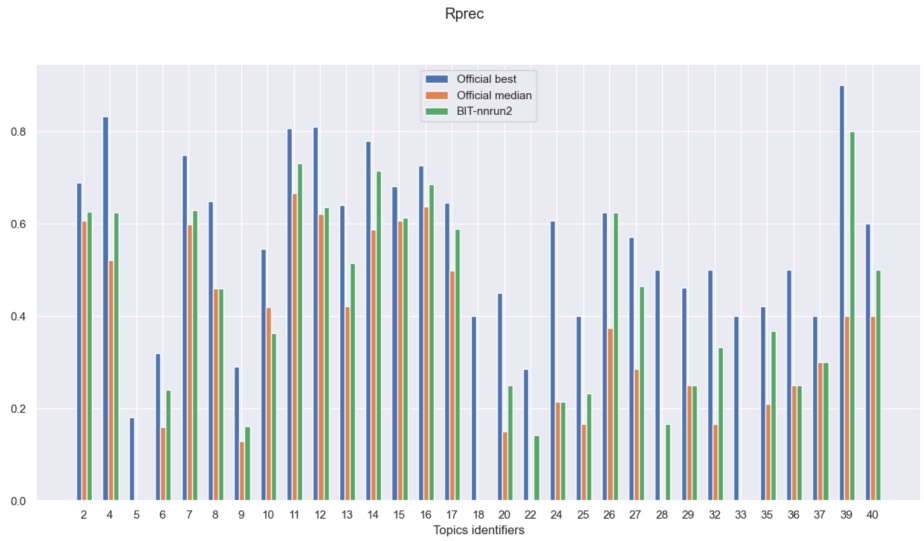

Fig. B7: Phase 1: R-prec of our best **neural** run against the official median and best values for each topic.

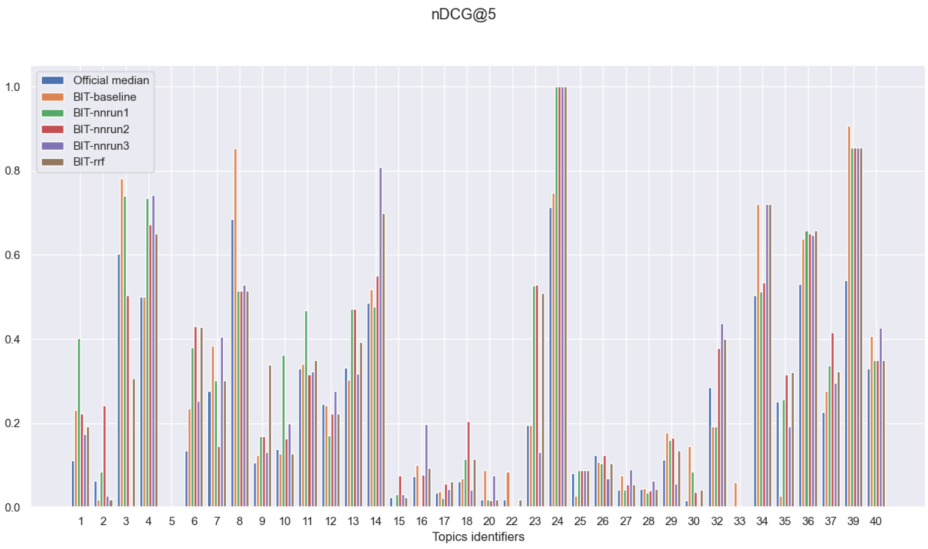

Fig. B8: Phase 2: nDCG@5 of all the submitted runs comparable to the official median.

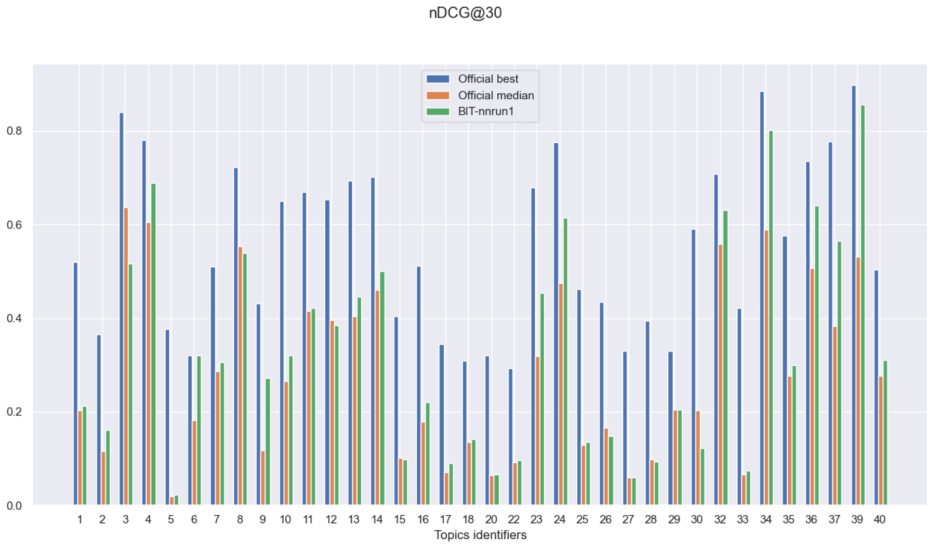

Fig. B9: Phase 2: nDCG@30 of our best **neural** run against the official median and best values for each topic.

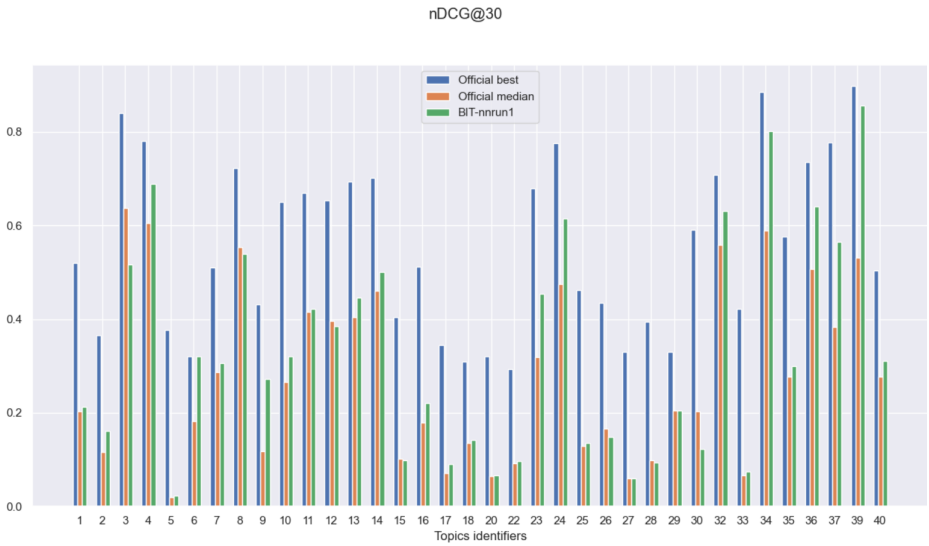

Fig. B10: Phase 2: nDCG@5 of our best **neural** run against the official median and best values for each topic.