Information Retrieval Galápagos Tortoise
WiSe 23/24 FSU Jena

Report January 28, 2024

In the previous milestone, we used the following retrieval pipeline:

A linear combination of a PL2 scoring [AR02] and a BM25 scoring [RW94] with Bo1 query expansion. The BM25's score is weighted twice as much as the PL2's score. Next, the top-50 results are reranked with a neural reranking approach from [PNL21]. For simplicity we reranked only with the mono-t5 transformer using the monot5-base-msmarco[1] model. For that, we need to split the documents into snippets, since the context length of the model is limited to 512 token. After the reranking process, the scores of the snippets have to be aggregated for each document, to compare the documents again. Options for this process include:

- max passage, where the document gets the maximum score of any of its snippets
- mean passage, where the document gets the mean score of any of its snippets
- k-max average where the document gets the average score of the top-k scores of its snippets.

Based on these scoring options, we formulate the following hypotheses:

Hypotheses

1. In the setting presented above, there is a significant ($\alpha < 0.05$) difference between the nDCG-scores of aggregating with max passage and with mean passage.
2. In the setting presented above, choosing $k \in \{2 \cdot i \mid i \in [1, 10]\}$ such that the nDCG-score of k-max average aggregation is maximized, yields a significantly ($\alpha < 0.05$) better nDCG-score than using max passage as well as mean passage aggregation.

Datasets

The experiments were conducted on two different datasets:

1. `irds:ir-lab-jena-leipzig-wise-2023/validation-20231104-training`
2. `irds:ir-lab-jena-leipzig-wise-2023/leipzig-topics-small-20240119-training`

[1] https://huggingface.co/castorini/monot5-base-msmarco

Results on Dataset 1

Hypothesis 1

The results of our first experiment are summarized in Table 1.

	max passage	mean passage
ndcg	0.273420	0.259885
ndcg@5	0.178241	0.159467
ndcg@10	0.218086	0.196292
ndcg p-value	0.000198	NaN
ndcg@5 p-value	0.000701	NaN
ndcg@10 p-value	0.000018	NaN

Table 1: Results of Hypothesis 1 on Validation Dataset

Using the max passage reranking scheme we obtained an nDCG-score of $\approx$ 0.2734, whereas with mean passage aggregation an nDCG-score of only $\approx$ 0.2599 was reached. The probability of observing this difference when Hypothesis 1 is wrong, that is the p-value, is $0.000198 < 0.05$. Thus, Hypothesis 1 can be confirmed.

Even though our hypothesis did not specifically request it, we decided to furthermore include results for nDCG@5 and nDCG@10 in our table. As one can see, replacing the nDCG-measure in Hypothesis 1 with one of them, still yields a hypothesis that can be confirmed.

The experiments conducted on the second dataset confirm these findings:

	max passage	mean passage
ndcg	0.573446	0.369422
ndcg@5	0.581550	0.255592
ndcg@10	0.525421	0.279602
ndcg p-value	0.004448	NaN
ndcg@5 p-value	0.000489	NaN
ndcg@10 p-value	0.001255	NaN

Table 2: Results of Hypothesis 1 on Leipzig Topics

Hypothesis 2

As for Hypothesis 2, Plot 1 shows the relation between parameter $k \in \{2 \cdot i | i \in [1, 10]\}$ and the nDCG-score obtained with k-max average aggregation. Clearly, increasing k decreases the nDCG-score. Setting $k = 2$ yields the highest nDCG-score (≈ 0.2714), which is, however, not better as the one obtained with max passage aggregation. Thus, the first part of Hypothesis 2, stating a significantly better nDCG-score of k-max average aggregation with optimal k compared to max passage aggregation can be discarded immediately.

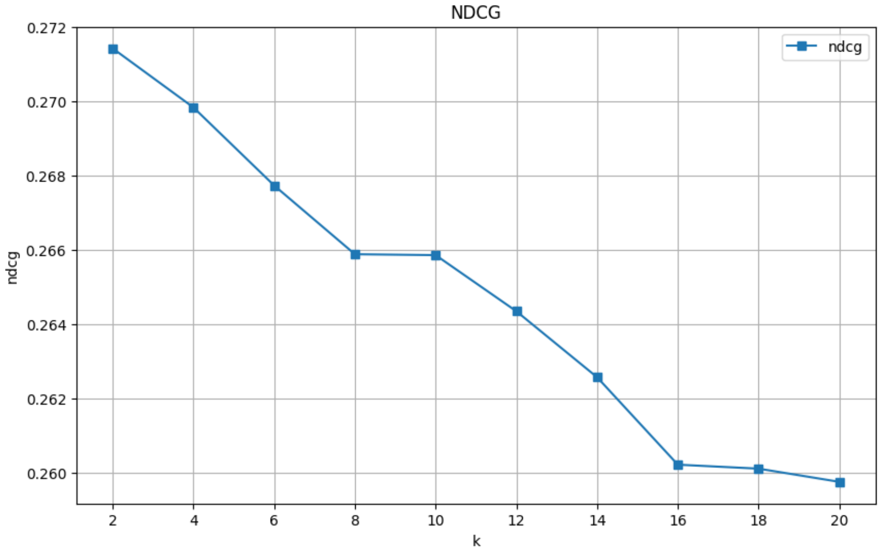

Figure 1: NDCG Plot

To investigate the second part of Hypothesis 2, that is comparing nDCG-scores of 2-max average and mean passage aggregation, have a look at Table 3.

	best kmax pipeline	max passage	mean passage
ndcg	0.271430	0.273420	0.259885
ndcg@5	0.176492	0.178241	0.159467
ndcg@10	0.215375	0.218086	0.196292
ndcg p-value	NaN	0.292945	0.000483
ndcg@5 p-value	NaN	0.547793	0.000671
ndcg@10 p-value	NaN	0.265544	0.000033

Table 3: Results of Hypothesis 2 on Validation Dataset

We already know the nDCG-scores of mean passage aggregation (≈ 0.2599) and 2-max average aggregation (≈ 0.2714). The probability of observing this difference when the second part of Hypothesis 2 is wrong is $0.000483 < 0.05$. Thus, the second part of the hypothesis turns out to be true. Due to the way we formulated the hypothesis (demanding that 2-max average aggregation performs significantly better as both, max and mean passage aggregation) we still need to discard it as a whole. We would have come to the same conclusion if we had examined nDCG@5 or nDCG@10 instead.

Again, the experiments conducted on the second dataset confirm these findings:

	best kmax pipeline	max passage	mean passage
ndcg	0.560287	0.573446	0.507939
ndcg@5	0.539375	0.581550	0.418526
ndcg@10	0.503897	0.525421	0.423454
ndcg p-value	NaN	0.328876	0.002114
ndcg@5 p-value	NaN	0.087884	0.000748
ndcg@10 p-value	NaN	0.252585	0.001160

Table 4: Results of Hypothesis 2 on Leipzig Topics

Conclusion

We were able to confirm that max passage aggregation yields a significantly better nDCG-score than mean passage aggregation. The nDCG-score of k-max average aggregation turned out best when setting $k = 2$ and was significantly better than the one obtained with mean passage aggregation. The nDCG-scores of max passage and 2-max average passage are very similar (which is not surprising since max passage is a special case of k-max average passage with $k = 1$ which is almost the k that got picked).

We are, however, a bit surprised that max passage aggregation performed best globally. Perhaps, this is a result of measuring the relevance of documents in binary: A single relevant passage already renders the whole document relevant, no matter how irrelevant its other passages are. Assessing the relevance in a more fined-grained manner might punish these irrelevant passages, so that max passage aggregation might not outperform the competing reranking schemes. It might be worth investigating this question further. Furthermore, experimenting with varying slide sizes in the monot5-model could be enlightening as well.

References

[RW94] Stephen E. Robertson and Steve Walker. “Some Simple Effective Approximations to the 2-Poisson Model for Probabilistic Weighted Retrieval”. In: *Proceedings of the 17th Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval. Dublin, Ireland, 3-6 July 1994 (Special Issue of the SIGIR Forum)*. Ed. by W. Bruce Croft and C. J. van Rijsbergen. ACM/Springer, 1994, pp. 232–241. DOI: 10.1007/978-1-4471-2099-5_24. URL: https://doi.org/10.1007/978-1-4471-2099-5%5C_24.

[AR02] Gianni Amati and C. J. van Rijsbergen. “Probabilistic models of information retrieval based on measuring the divergence from randomness”. In: *ACM Trans. Inf. Syst.* 20.4 (2002), pp. 357–389. DOI: 10.1145/582415.582416. URL: http://doi.acm.org/10.1145/582415.582416.

[PNL21] Ronak Pradeep, Rodrigo Frassetto Nogueira, and Jimmy Lin. “The Expando-Mono-Duo Design Pattern for Text Ranking with Pretrained Sequence-to-Sequence Models”. In: *CoRR* abs/2101.05667 (2021). arXiv: 2101.05667. URL: https://arxiv.org/abs/2101.05667.