WOWS-EVAL 2025: Efficient Baselines for Automated Relevance Label Transfer

Daria Alexander[1,*], Maik Fröbe[2,*] and Gijs Hendriksen[1,*]

[1] *Radboud University Nijmegen*
[2] *Friedrich-Schiller-Universität Jena*

Abstract

This paper describes our submissions to the 2025 shared task on WOWS-EVAL that aims to automatically estimate the relevance of documents to a query given documents that are already known to be relevant to the query. In the end, we aim to transfer topics and relevance assessments from established test collections to new and evolving datasets. The goal of this paper is to provide a set of efficient and cheap methods that can be used as baselines for more complex relevance assessors (e.g. LLM-as-a-judge). We apply three main approaches: (1) framing this as a retrieval problem and using the retrieval scores with and without query expansion through relevance feedback, (2) transformer-based labeling through `autoqrels`, and (3) weak supervision using Snorkel and handcrafted labeling rules. We evaluate the effectiveness of our labeling approaches by computing the correlation between ground-truth system rankings and the system rankings we obtain when using our generated relevance assessments. While our approaches outperform a basic approach that simply performs BM25 retrieval, our maximum correlation of 0.427 highlights that automated relevance transfer using cheap models is unreliable, and further experiments and approaches should be tried before this could be applied to new collections in practice.

Keywords

IR Evaluation, Relevance Judgments, Longitudinal IR

1. Introduction

In the context of the OpenWebSearch.eu project[1], which aims to build and maintain an Open Web Index [1, 2], evaluation plays a crucial role. In the best case, the Open Web Index supports diverse downstream retrieval applications. Evaluation can help to ensure that the downstream retrieval applications work. However, information retrieval evaluation in the Cranfield paradigm [3] usually requires relevance judgments for a static set of information needs on a static set of documents (some exceptions like the LongEval shared task also handle longitudinal aspects [4, 5, 6]). Since the Open Web Index is constantly updated in the Open Web Search scenario (documents might be updated, removed, or newly created), the underlying document collection evolves. This requires that the relevance judgments for the Open Web Index must also be updated, as this allows continuous evaluation of which retrieval pipelines work well in which retrieval scenarios.

In this work, we describe three approaches for transferring relevance labels from existing test collections. First, we frame the task as retrieval problem, where we use traditional lexical retrieval models with respectively without relevance feedback from the known relevant documents, and then use the min-max normalized retrieval scores as probabilities that documents with unknown relevance are relevant. Second, we ran an OpenWebSearch.eu hackathon with ca. 30 participants in which we collected prompts for the relevance label transfer for the `autoqrels` framework [7]. Third, we combine those approaches with rule-based weak-supervision signals using Snorkel. We performed evaluations on ClueWeb22, to which we transferred the relevance judgments from 13 queries from TREC-style shared tasks that ran before 2022. On ClueWeb22, we judge 1 100 documents manually as ground truth to evaluate our approaches. Our code is available on GitHub.[2]

Second International Workshop on Open Web Search (WOWS 2025)

[*] All authors contributed equally.

[1] https://openwebsearch.eu

[2] https://github.com/OpenWebSearch/wows-code/tree/main/ecir25/baselines

2. Related Work

We review related work on transferring relevance labels, on how to predict relevance with transformer-based models, and on weak supervision as they form the basis of our work.

2.1. Relevance Label Transfer

One of our motivations was the case-study that tried to transfer relevance judgments from the ClueWeb09 to the ClueWeb12 via near-duplicate detection [8]. The idea was that, if a document was relevant for a query in the 2009 crawl, and there exists a near-duplicate of this document in the 2012 crawl, one can transfer the label. However, only very few documents from 2009 had near-duplicates in 2012, so that only 10 % of the ClueWeb09 relevance judgments could be transferred to near-duplicates in the ClueWeb12. This motivates our work, as we try to look if we can transfer more relevance judgments via approaches that go beyond simple near-duplicate detection.

2.2. Relevance Prediction with Transformers

Transformer models can be highly effective for retrieval and re-ranking [9]. Consequently, they have also been used to create relevance judgments. MacAvaney and Soldaini propose `autoqrels` that enables the creation of relevance judgments in a pointwise and a pairwise manner [7]. We use their input and output structure for our task as well. More recently, large language models have been used to create relevance judgments [10, 11, 12, 13, 14]. Still, there is an ongoing discussion to what degree relevance judgments should and could be automated [15, 16]. With our goal to transfer relevance judgments in a longitudinal setting, we try to enable more re-use scenarios for human-generated relevance judgments.

2.3. Weak Supervision

One of the most common problems with successful training of machine learning models is the lack of datasets with high quality annotations. Manual collection of annotations is a costly, tedious and time-consuming process. Academic research institutions often do not have enough funding to gather large-scale annotations, limiting their capabilities of creating high-quality corpora. While LLMs can be used for that task [17, 18, 19], their computational efficiency of might hinder scaling to large collections. More lightweight approaches such as weak supervision can, on the other side, scale to large collections.

Weak supervision is an approach in machine learning where noisy, limited, or imprecise sources are used instead of (or along with) gold labelled data. It became popularized with the introduction of the data programming paradigm [20]. This paradigm enables the quick and easy creation of labelling functions by which users express weak supervision strategies or domain heuristics. Various weak supervision approaches can be represented by labelling functions, such as distant supervision, heuristics or the results of crowd-sourcing annotations. Weak supervision has been successfully applied in various problems in the area of natural language processing and information retrieval [21, 22, 23].

Snorkel is a weak supervision system that enables users to train models using labelling functions without hand labelling any data [24]. It is an end-to-end system for creating labelling functions and training and evaluating the labelling model. Snorkel is initially designed to work with classification or extraction tasks. According to [25], it offers comparable performance to newer and more complex weak supervision systems. Hence, we also use Snorkel for our experiments.

3. Experimental Setup

As our intuition is to use relevance data that was created for old versions of the Open Web Index to evaluate retrieval systems on newer versions of the Open Web Index, we design our experiments so that we transfer relevance judgments from TREC-style shared tasks that ran in the past to the ClueWeb22 collection [26]. We start from four source corpora with TREC-style relevance judgments:

the 2020 and 2021 edition of Touché [27, 28], Robust04 [29], the 2019 and 2020 edition of TREC Deep Learning [30, 31], and the 2009 TREC Web track [32]. From each of those source corpora, we select a set of topics, yielding 13 topics in total that we aim to transfer to ClueWeb22. For each topic, we submit the title and description against ChatNoir [33, 34]. We use pooling on those runs to judge overall 1 100 documents with two relevance assessors. We then create 500 randomized retrieval runs by shuffling the judged documents and stratified sampling the runs so that they cover the nDCG range between 0 and 1 uniformly (for the complete range between 0 and 1, we create buckets of size 0.02 and select from each bucket one run). This gives us a ground-truth system ranking that covers the range of nDCG scores between 0 and 1 (it is possible that no runs fall into a bucket, but we ensure that the 1.0 run is included). For every approach to predict the relevance of a document, we follow the `autoqrels` methodology that a relevance can be a float between 0 and 1, as those floats can be directly passed to the nDCG calculation. We then use the predicted relevance of an approach (a score between 0 and 1) to calculate alternative system rankings obtained from the predicted relevance. Finally, we calculate the correlations between the ground-truth system ranking and the system ranking obtained via the predicted relevance scores. We collected the approaches with TIRA/TIREx [35, 36].

4. Approaches

We next describe our approaches.

4.1. Our BM25 Baseline

To have a reasonable baseline, we used BM25 scores of documents with unknown relevance to the query. For this, we created a PyTerrier [37] index of all documents with unknown relevance. We created a BM25 top-1000 ranking for the query. We normalized the BM25 scores (which could be outside the 0–1 range) using min-max normalization. We assigned documents that were not retrieved a probability of 0. Min-max normalization ensures that the document that was retrieved on the top-position has the highest probability of 1 of being relevant.

4.2. Estimating Relevance Probabilities with Relevance Feedback

A common approach to incorporate relevance information into lexical retrieval is to apply (pseudo-) relevance feedback: expanding the input query based on a set of (pseudo-)relevant documents. We apply relevance feedback for the pairwise relevance label transfer task as follows: (1) we perform RM3 query expansion [38] using the known relevant document with Robust04 [29] as the background collection, (2) we compute the BM25 ranking scores for each unjudged document with the expanded query, and (3) we apply min-max normalization to transform the scores into a probability distribution (as done with the BM25 baseline). We experiment with two settings: one in which RM3 receives a single relevant document, and one in which we supply all of the known relevant documents.

4.3. Estimating Relevance Probabilities with `autoqrels`

We specifically designed pointwise (which uses the query and unknown document as input) and pairwise (which uses the query, known relevant document, and unknown document) `autoqrels` approaches for the relevance transfer. We ran a 60 minute OpenWebSearch.eu hackathon with ca. 30 participants during which we collected 5 prompts to the problem (ca. 30 minutes were dedicated to collaboratively develop prompts, the remaining time for introducing the setup). We executed each prompt with three Flan-T5 variants of different size (small, base, and large). Listing 1 shows the pointwise prompt that we used as starting point for the hackaton, and Listing 2 the pairwise prompt.

```
Instruction: Indicate if the passage answers the question.
###
Example 1:
Question: At about what age do adults normally begin to lose bone mass?
Passage: For most people, bone mass peaks during the third decade of life.
         By this age, men typically have accumulated more bone mass than women.
         After this point, the amount of bone in the skeleton typically begins to decline.
Answer: Perfectly relevant
###
Example 2:
Question: when and where did the battle of manassas take place
Passage: Summary of the Battle of Bull Run. The conflict took place close to Manassas Junction, Virginia.
         Around 35,000 Union soldiers marched from Washing D.C. towards Bull Run (a small river)
         where a 20,000 troop Confederate force was stationed.
Answer: Irrelevant
###
Example 3:
Question: which kind of continental boundary is formed where two plates move horizontally past one another?
Passage: One plate slides horizontally past another.
         The best-known example is the earthquake-prone San Andreas Fault Zone of California,
         which marks the boundary between the Pacific and North America Plates.
Answer: Highly relevant
###
Example 4:
Question: what foods should you stay away from if you have asthma
Passage: Get early and regular prenatal care. The first 8 weeks of your pregnancy are important to your babys development.
         Early and regular prenatal care can boost your chances of having a safe pregnancy and a healthy baby.
         Prenatal care includes screenings, regular exams, pregnancy and childbirth education, and counseling and support.
Answer: Irrelevant
###
Example 5:
Question: what is lbm in body composition
Passage: They also measured the participants body fat (subtracting the body fat weight from the total body weight).
Answer: Relevant
###
Example 6:
Question: QUERY
Passage: DOCUMENT
Answer:
```

Listing 1: The initial pointwise prompt for `autoqrels` for the hackathon.

```
Determine if passage B is as relevant as passage A for the given query.
Passage A: "RELEVANT-DOCUMENT"
Passage B: "UNKNOWN-DOCUMENT"
Query: "QUERY"
Is passage B as relevant as passage A?
```

Listing 2: The initial pairwise prompt for `autoqrels` for the hackathon.

4.4. Estimating Relevance Probabilities with Snorkel

Snorkel has been used for a variety of tasks [39, 21, 40, 41], but, to our knowledge, no one has utilised it for predicting document relevance. The core of Snorkel are labeling functions which assign a certain label to the input data; those labels are then aggregated to predict a final label. The outputs of the labeling functions are usually binary. For example, for a task that would determine whether a mail is spam or not the labels would be: 1 (spam), 0 (not spam) and -1 (abstain).

However, for our task, we need to emit probabilities directly instead of obtaining binary scores. We achieved this by using Snorkel's *LabelModel* and its *predict_proba()* method, which returns the probability distribution over labels for each example in the label matrix.

For pointwise ranking, each query-document pair is assigned a probabilistic relevance score. For pairwise ranking, we aim to estimate the probability that an unknown document is relevant, given a document that is already known to be relevant for the same query. Each instance in the label matrix consists of a pair where one document is labeled as relevant, and the other is an unknown candidate. Instead of making a direct binary comparison, the *LabelModel* aggregates weak signals from multiple labeling functions to infer the probability that the unknown document is also relevant.

By using weak supervision, Snorkel's *LabelModel* learns the reliability of different labeling functions and outputs a probabilistic relevance score for the unknown document. This probability can be used as a ranking signal, which allows to sort documents based on their estimated likelihood of relevance.

4.4.1. Pointwise ranking

In Snorkel, the goal of pointwise ranking is to assign weak relevance labels to individual (query, document) pairs, which are then aggregated into probabilistic labels by Snorkel's *LabelModel*. We created labeling functions based on the following features:

- BM25 score
- Boolean match (the document contains at least one query term)
- Word Levenshtein distance
- TF-IDF cosine similarity
- BERT cosine simlilarity

We use the "all-mpnet-base-v2" BERT model for BERT cosine similarity probability. Also, as Snorkel aggregates multiple weak labels, we boosted highly confident signals, ensuring that Snorkel learns the correct probability distributions.

4.4.2. Pairwise ranking

We define several pairwise labeling functions in Snorkel that provide weak supervision signals for ranking tasks by comparing an unknown document to a known relevant document for the same query. Instead of predicting absolute relevance scores, these functions output probabilistic signals that estimate the likelihood of the unknown document being relevant, given an existing relevant document.

A typical pairwise labeling function takes as input: (1) a query Q, (2) an unknown document D_u, and (3) a known relevant document D_r. The function then computes a pairwise comparison score:

$$f(Q, D_r, D_u) = S(Q, D_r) - S(Q, D_u)$$

where $S(Q, D)$ represents a similarity or relevance score assigned to a document with respect to the query. The function then outputs a weak supervision signal that indicates the degree to which the unknown document is less relevant than the known relevant document. To perform pairwise ranking, we used labeling functions based on the following features:

- BM25 score
- Word Levenstein distance
- Jaccard similarity
- TF-IDF cosine similarity
- BERT cosine simlilarity

Snorkel's *LabelModel* aggregates multiple weak labeling functions to generate a probabilistic ranking signal. The raw difference score serves as a weak preference indicator, and after label aggregation, the model outputs relevance probabilities for the unknown document.

5. Evaluation

Table 1 shows the evaluation results. For each of our approaches, we report the Spearman correlation (ranges from -1 to 1) in how well the system rankings are preserved against the ground-truth system ranking when using the predicted relevance judgments by our approaches. A Spearman correlation of 1 would indicate that the system rankings are identical, which would be the best case as this would allow us to select suitable retrieval models on the transferred relevance judgments. A Spearman correlation of 0 would indicate random correlations (e.g., randomly generating system rankings) whereas -1 would indicate a perfect negative correlation. Our results show that all predictors achieve a positive correlation. Even the unsupervized BM25 baseline achieves a positive correlation of 0.151, though this is too low to be applicable in practice. Our relevance feedback approaches and the Snorkel approach outperform

Approach	Spearman
Flan-T5-large (best prompt)	**0.427**
Flan-T5-base (best prompt)	0.266
Flan-T5-small (best prompt)	0.067
BM25	0.151
RF (All)	0.266
RF (One)	**0.276**
Snorkel (Pointwise)	0.230
Snorkel (Pairwise)	0.035

Table 1
The effectiveness of the different relevance predictors measured as Spearman correlation.

the BM25 baseline and achieve correlations between 0.230 and 0.276. For the `autoqrels` approaches, the backbone model substantially impacts the effectiveness. The smallest model (Flan-T5-small) with the best prompt is even less effective than BM25, whereas our largest model (Flan-T5-large) achieves a correlation of 0.427. Overall, the correlations that our approaches achieve indicate that further improvements are needed before they can be used in practice. The code to build the evaluation and more approaches (that we excluded from our analysis here) are available online.[3]

6. Conclusion and Future Work

We described our approaches to update potentially outdated relevance judgments for document collections that might evolve over time. Our intended use-case is to support evaluations of retrieval systems on the Open Web Index, where new documents might be added, removed, or modified over time. Given a set of documents that are known to be relevant to an information need, we explored different approaches to transfer the relevance judgments to updated versions of the corpus. Our results show that the relevance label transfer is not yet reliable enough to be used for practical applications.

Consequently, there might be several ideas that could be interesting for future work. Taking the topic difficulty into account could be interesting, as the transfer might be easier on easier topics. Alternatively, integrating more information from the topics could also improve the effectiveness of the relevance transfer. Another interesting line could be to incorporate old relevance judgments as contrastive examples into the prompts to large language models.

Acknowledgments

This work has received funding from the European Union's Horizon Europe research and innovation programme under grant agreement No 101070014 (OpenWebSearch.EU, https://doi.org/10.3030/101070014).

References

[1] M. Granitzer, S. Voigt, N. A. Fathima, M. Golasowski, C. Guetl, T. Hecking, G. Hendriksen, D. Hiemstra, J. Martinovic, J. Mitrovic, I. Mlakar, S. Moiras, A. Nussbaumer, P. Öster, M. Potthast, M. S. Srdic, S. Megi, K. Slaninová, B. Stein, A. P. de Vries, V. Vondrák, A. Wagner, S. Zerhoudi, Impact and development of an open web index for open web search, J. Assoc. Inf. Sci. Technol. 75 (2024) 512–520. URL: https://doi.org/10.1002/asi.24818. doi:10.1002/ASI.24818.

[2] G. Hendriksen, M. Dinzinger, S. M. Farzana, N. A. Fathima, M. Fröbe, S. Schmidt, S. Zerhoudi, M. Granitzer, M. Hagen, D. Hiemstra, M. Potthast, B. Stein, The open web index - crawling and

[3] https://github.com/OpenWebSearch/wows-code/blob/main/ecir25/evaluation.ipynb

indexing the web for public use, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part V, volume 14612 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 130–143. URL: https://doi.org/10.1007/978-3-031-56069-9_10. doi:10.1007/978-3-031-56069-9_10.

[3] E. M. Voorhees, The evolution of cranfield, in: N. Ferro, C. Peters (Eds.), Information Retrieval Evaluation in a Changing World - Lessons Learned from 20 Years of CLEF, volume 41 of *The Information Retrieval Series*, Springer, 2019, pp. 45–69. URL: https://doi.org/10.1007/978-3-030-22948-1_2. doi:10.1007/978-3-030-22948-1_2.

[4] R. Alkhalifa, I. M. Bilal, H. Borkakoty, J. Camacho-Collados, R. Deveaud, A. El-Ebshihy, L. E. Anke, G. G. Sáez, P. Galuscáková, L. Goeuriot, E. Kochkina, M. Liakata, D. Loureiro, H. T. Madabushi, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at CLEF 2023, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023, Dublin, Ireland, April 2-6, 2023, Proceedings, Part III, volume 13982 of *Lecture Notes in Computer Science*, Springer, 2023, pp. 499–505. URL: https://doi.org/10.1007/978-3-031-28241-6_58. doi:10.1007/978-3-031-28241-6_58.

[5] R. Alkhalifa, H. Borkakoty, R. Deveaud, A. El-Ebshihy, L. E. Anke, T. Fink, G. G. Sáez, P. Galuscáková, L. Goeuriot, D. Iommi, M. Liakata, H. T. Madabushi, P. Medina-Alias, P. Mulhem, F. Piroi, M. Popel, C. Servan, A. Zubiaga, Longeval: Longitudinal evaluation of model performance at CLEF 2024, in: N. Goharian, N. Tonellotto, Y. He, A. Lipani, G. McDonald, C. Macdonald, I. Ounis (Eds.), Advances in Information Retrieval - 46th European Conference on Information Retrieval, ECIR 2024, Glasgow, UK, March 24-28, 2024, Proceedings, Part VI, volume 14613 of *Lecture Notes in Computer Science*, Springer, 2024, pp. 60–66. URL: https://doi.org/10.1007/978-3-031-56072-9_8. doi:10.1007/978-3-031-56072-9_8.

[6] M. Cancellieri, A. El-Ebshihy, T. Fink, M. Fröbe, P. Galuscáková, G. G. Sáez, L. Goeuriot, D. Iommi, J. Keller, P. Knoth, P. Mulhem, F. Piroi, D. Pride, P. Schaer, Longeval at CLEF 2025: Longitudinal evaluation of IR systems on web and scientific data, in: J. Carrillo-de-Albornoz, A. G. S. de Herrera, J. Gonzalo, L. Plaza, J. Mothe, F. Piroi, P. Rosso, D. Spina, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 16th International Conference of the CLEF Association, CLEF 2025, Madrid, Spain, September 9-12, 2025, Proceedings, volume 16089 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 363–387. URL: https://doi.org/10.1007/978-3-032-04354-2_20. doi:10.1007/978-3-032-04354-2_20.

[7] S. MacAvaney, L. Soldaini, One-shot labeling for automatic relevance estimation, in: H. Chen, W. E. Duh, H. Huang, M. P. Kato, J. Mothe, B. Poblete (Eds.), Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023, ACM, 2023, pp. 2230–2235. URL: https://doi.org/10.1145/3539618.3592032. doi:10.1145/3539618.3592032.

[8] M. Fröbe, J. Bevendorff, L. Gienapp, M. Völske, B. Stein, M. Potthast, M. Hagen, Copycat: Near-duplicates within and between the clueweb and the common crawl, in: F. Diaz, C. Shah, T. Suel, P. Castells, R. Jones, T. Sakai (Eds.), SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021, ACM, 2021, pp. 2398–2404. URL: https://doi.org/10.1145/3404835.3463246. doi:10.1145/3404835.3463246.

[9] J. Lin, R. Nogueira, A. Yates, Pretrained Transformers for Text Ranking: BERT and Beyond, Synthesis Lectures on Human Language Technologies, Morgan & Claypool Publishers, 2021. URL: https://doi.org/10.2200/S01123ED1V01Y202108HLT053. doi:10.2200/S01123ED1V01Y202108HLT053.

[10] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Perspectives on large language models for relevance judgment, in: M. Yoshioka, J. Kiseleva, M. Aliannejadi (Eds.), Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, ICTIR 2023, Taipei, Taiwan, 23 July 2023, ACM, 2023, pp. 39–50. doi:10.1145/3578337.3605136.

[11] G. Faggioli, L. Dietz, C. L. A. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, H. Wachsmuth, Who determines what is relevant? humans or ai? why not both?, Commun. ACM 67 (2024) 31–34. doi:10.1145/3624730.

[12] P. Thomas, S. Spielman, N. Craswell, B. Mitra, Large language models can accurately predict searcher preferences, in: G. H. Yang, H. Wang, S. Han, C. Hauff, G. Zuccon, Y. Zhang (Eds.), Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2024, Washington DC, USA, July 14-18, 2024, ACM, 2024, pp. 1930–1940. doi:10.1145/3626772.3657707.

[13] S. Upadhyay, R. Pradeep, N. Thakur, N. Craswell, J. Lin, UMBRELA: umbrela is the (open-source reproduction of the) bing relevance assessor, CoRR abs/2406.06519 (2024). doi:10.48550/ARXIV.2406.06519. arXiv:2406.06519.

[14] S. Upadhyay, R. Pradeep, N. Thakur, D. Campos, N. Craswell, I. Soboroff, H. T. Dang, J. Lin, A large-scale study of relevance assessments with large language models: An initial look, CoRR abs/2411.08275 (2024). URL: https://doi.org/10.48550/arXiv.2411.08275. doi:10.48550/ARXIV.2411.08275. arXiv:2411.08275.

[15] I. Soboroff, Don't use llms to make relevance judgments, CoRR abs/2409.15133 (2024). doi:10.48550/ARXIV.2409.15133. arXiv:2409.15133.

[16] L. Dietz, O. Zendel, P. Bailey, C. L. A. Clarke, E. Cotterill, J. Dalton, F. Hasibi, M. Sanderson, N. Craswell, Principles and guidelines for the use of LLM judges, in: H. Zamani, L. Dietz, B. Piwowarski, S. Bruch (Eds.), Proceedings of the 2025 International ACM SIGIR Conference on Innovative Concepts and Theories in Information Retrieval, ICTIR 2025, Padua, Italy, 18 July 2025, ACM, 2025, pp. 218–229. URL: https://doi.org/10.1145/3731120.3744588. doi:10.1145/3731120.3744588.

[17] R. Zhang, Y. Li, Y. Ma, M. Zhou, L. Zou, Llmaaa: Making large language models as active annotators, arXiv preprint arXiv:2310.19596 (2023).

[18] M. Pavlovic, M. Poesio, The effectiveness of llms as annotators: A comparative overview and empirical analysis of direct representation, arXiv preprint arXiv:2405.01299 (2024).

[19] H. Zhang, J. Yang, J. Nie, P. Liang, K. Wu, D. Lian, R. Mao, Y. Song, Efficient data labeling by hierarchical crowdsourcing with large language models, in: Proceedings of the 31st International Conference on Computational Linguistics, 2025, pp. 11290–11303.

[20] A. J. Ratner, C. M. De Sa, S. Wu, D. Selsam, C. Ré, Data programming: Creating large training sets, quickly, in: D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, R. Garnett (Eds.), Advances in Neural Information Processing Systems, volume 29, Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper/2016/file/6709e8d64a5f47269ed5cea9f625f7ab-Paper.pdf.

[21] S. Badene, K. Thompson, J.-P. Lorré, N. Asher, Weak supervision for learning discourse structure, in: EMNLP, 2019.

[22] J. A. Fries, P. Varma, V. S. Chen, K. Xiao, H. Tejeda, P. Saha, J. Dunnmon, H. Chubb, S. Maskatia, M. Fiterau, et al., Weakly supervised classification of aortic valve malformations using unlabeled cardiac mri sequences, Nature communications 10 (2019) 1–10.

[23] M. Dehghani, H. Zamani, A. Severyn, J. Kamps, W. B. Croft, Neural ranking models with weak supervision, in: Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR '17, Association for Computing Machinery, New York, NY, USA, 2017, p. 65–74. URL: https://doi.org/10.1145/3077136.3080832. doi:10.1145/3077136.3080832.

[24] A. Ratner, S. H. Bach, H. Ehrenberg, J. Fries, S. Wu, C. Ré, Snorkel: Rapid training data creation with weak supervision, Proc. VLDB Endow. 11 (2017) 269–282. URL: https://doi.org/10.14778/3157794.3157797. doi:10.14778/3157794.3157797.

[25] G. Zheng, G. Karamanolakis, K. Shu, A. H. Awadallah, Walnut: A benchmark on weakly supervised learning for natural language understanding, arXiv preprint arXiv:2108.12603 (2021).

[26] A. Overwijk, C. Xiong, J. Callan, Clueweb22: 10 billion web documents with rich information, in: E. Amigó, P. Castells, J. Gonzalo, B. Carterette, J. S. Culpepper, G. Kazai (Eds.), SIGIR '22: The 45th International ACM SIGIR Conference on Research and Development in Information Retrieval,

Madrid, Spain, July 11 - 15, 2022, ACM, 2022, pp. 3360–3362. URL: https://doi.org/10.1145/3477495.3536321. doi:10.1145/3477495.3536321.

[27] A. Bondarenko, M. Fröbe, M. Beloucif, L. Gienapp, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of touché 2020: Argument retrieval - extended abstract, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 384–395. URL: https://doi.org/10.1007/978-3-030-58219-7_26. doi:10.1007/978-3-030-58219-7_26.

[28] A. Bondarenko, L. Gienapp, M. Fröbe, M. Beloucif, Y. Ajjour, A. Panchenko, C. Biemann, B. Stein, H. Wachsmuth, M. Potthast, M. Hagen, Overview of touché 2021: Argument retrieval, in: K. S. Candan, B. Ionescu, L. Goeuriot, B. Larsen, H. Müller, A. Joly, M. Maistro, F. Piroi, G. Faggioli, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, Virtual Event, September 21-24, 2021, Proceedings, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 450–467. URL: https://doi.org/10.1007/978-3-030-85251-1_28. doi:10.1007/978-3-030-85251-1_28.

[29] E. Voorhees, Overview of the trec 2004 robust retrieval track, in: TREC, 2004.

[30] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, E. M. Voorhees, Overview of the TREC 2019 deep learning track, CoRR abs/2003.07820 (2020). URL: https://arxiv.org/abs/2003.07820. arXiv:2003.07820.

[31] N. Craswell, B. Mitra, E. Yilmaz, D. Campos, Overview of the TREC 2020 deep learning track, CoRR abs/2102.07662 (2021). URL: https://arxiv.org/abs/2102.07662. arXiv:2102.07662.

[32] C. L. A. Clarke, N. Craswell, I. Soboroff, Overview of the TREC 2009 web track, in: E. M. Voorhees, L. P. Buckland (Eds.), Proceedings of The Eighteenth Text REtrieval Conference, TREC 2009, Gaithersburg, Maryland, USA, November 17-20, 2009, volume 500-278 of *NIST Special Publication*, National Institute of Standards and Technology (NIST), 2009. URL: http://trec.nist.gov/pubs/trec18/papers/WEB09.OVERVIEW.pdf.

[33] J. H. Merker, J. Bevendorff, M. Fröbe, T. Hagen, H. Scells, M. Wiegmann, B. Stein, M. Hagen, M. Potthast, Web-scale retrieval experimentation with chatnoir-pyterrier, in: C. Hauff, C. Macdonald, D. Jannach, G. Kazai, F. M. Nardini, F. Pinelli, F. Silvestri, N. Tonellotto (Eds.), Advances in Information Retrieval - 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 6-10, 2025, Proceedings, Part V, volume 15576 of *Lecture Notes in Computer Science*, Springer, 2025, pp. 96–104. URL: https://doi.org/10.1007/978-3-031-88720-8_17. doi:10.1007/978-3-031-88720-8_17.

[34] J. Bevendorff, B. Stein, M. Hagen, M. Potthast, Elastic chatnoir: Search engine for the clueweb and the common crawl, in: G. Pasi, B. Piwowarski, L. Azzopardi, A. Hanbury (Eds.), Advances in Information Retrieval - 40th European Conference on IR Research, ECIR 2018, Grenoble, France, March 26-29, 2018, Proceedings, volume 10772 of *Lecture Notes in Computer Science*, Springer, 2018, pp. 820–824. URL: https://doi.org/10.1007/978-3-319-76941-7_83. doi:10.1007/978-3-319-76941-7_83.

[35] M. Fröbe, M. Wiegmann, N. Kolyada, B. Grahm, T. Elstner, F. Loebe, M. Hagen, B. Stein, M. Potthast, Continuous Integration for Reproducible Shared Tasks with TIRA.io, in: J. Kamps, L. Goeuriot, F. Crestani, M. Maistro, H. Joho, B. Davis, C. Gurrin, U. Kruschwitz, A. Caputo (Eds.), Advances in Information Retrieval. 45th European Conference on IR Research (ECIR 2023), Lecture Notes in Computer Science, Springer, Berlin Heidelberg New York, 2023, pp. 236–241. doi:10.1007/978-3-031-28241-6_20.

[36] M. Fröbe, J. Reimer, S. MacAvaney, N. Deckers, S. Reich, J. Bevendorff, B. Stein, M. Hagen, M. Potthast, The Information Retrieval Experiment Platform, in: H.-H. Chen, W. Duh, H.-H. Huang, M. Kato, J. Mothe, B. Poblete (Eds.), 46th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 2023), ACM, 2023, pp. 2826–2836. doi:10.1145/3539618.3591888.

[37] C. Macdonald, N. Tonellotto, S. MacAvaney, I. Ounis, Pyterrier: Declarative experimentation in python from BM25 to dense retrieval, in: G. Demartini, G. Zuccon, J. S. Culpepper, Z. Huang, H. Tong (Eds.), CIKM '21: The 30th ACM International Conference on Information and Knowledge Management, Virtual Event, Queensland, Australia, November 1 - 5, 2021, ACM, 2021, pp. 4526–4533. URL: https://doi.org/10.1145/3459637.3482013. doi:10.1145/3459637.3482013.

[38] N. Abdul-Jaleel, J. Allan, W. B. Croft, F. Diaz, L. Larkey, X. Li, M. D. Smucker, C. Wade, Umass at trec 2004: Novelty and hard, in: Proceedings of TREC-13, 2004.

[39] S. H. Bach, D. Rodriguez, Y. Liu, C. Luo, H. Shao, C. Xia, S. Sen, A. Ratner, B. Hancock, H. Alborzi, et al., Snorkel drybell: A case study in deploying weak supervision at industrial scale, in: Proceedings of the 2019 International Conference on Management of Data, 2019, pp. 362–375.

[40] S. Dua, I. Baldini, D. A. Katz-Rogozhnikov, E. van der Veen, A. Britt, P. Mangalath, L. B. Kleiman, C. D. V. Fitz, Biomedical corpus filtering: A weak supervision paradigm with infused domain expertise., in: SDU@ AAAI, 2021.

[41] D. Alexander, W. Kusa, A. P. de Vries, Orcas-i: queries annotated with intent using weak supervision, in: Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2022, pp. 3057–3066.