Understanding the User
An Intent-Based Ranking Dataset

Anand, Abhijit; Leonhardt, Jurek; V., Venktesh; Anand, Avishek

DOI
10.1145/3627673.3679166

Publication date
2024

Document Version
Final published version

Published in
CIKM '24

Understanding the User: An Intent-Based Ranking Dataset

Abhijit Anand
L3S Research Center
Hannover, Germany
aanand@L3S.de

Jurek Leonhardt
Delft University of Technology
Delft, The Netherlands
L.J.Leonhardt@tudelft.nl

Venktesh V
Delft University of Technology
Delft, The Netherlands
v.Viswanathan-1@tudelft.nl

Avishek Anand
Delft University of Technology
Delft, The Netherlands
avishek.anand@tudelft.nl

ABSTRACT

As information retrieval systems continue to evolve, accurate evaluation and benchmarking of these systems become pivotal. Web search datasets, such as MS MARCO, primarily provide short keyword queries without accompanying intent or descriptions, posing a challenge in comprehending the underlying information need. This paper proposes an approach to augmenting such datasets to annotate informative query descriptions, with a focus on two prominent benchmark datasets: TREC-DL-21 and TREC-DL-22. Our methodology involves utilizing state-of-the-art LLMs to analyze and comprehend the implicit intent within individual queries from benchmark datasets. By extracting key semantic elements, we construct detailed and contextually rich descriptions for these queries. To validate the generated query descriptions, we employ crowdsourcing as a reliable means of obtaining diverse human perspectives on the accuracy and informativeness of the descriptions. This information can be used as an evaluation set for tasks such as ranking, query rewriting, or others.

CCS CONCEPTS

• **Information systems → Information retrieval**.

KEYWORDS

Intent Dataset; Ad-hoc retrieval; Ranking; User Intents; Web Search; Diversity; Data collection

ACM Reference Format:
Abhijit Anand, Jurek Leonhardt, Venktesh V, and Avishek Anand. 2024. Understanding the User: An Intent-Based Ranking Dataset. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management (CIKM '24), October 21–25, 2024, Boise, ID, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3627673.3679166

1 INTRODUCTION

In information retrieval (IR), a core challenge in building ranking models is to explicitly or implicitly *aligning* the actual user intent with the machine intent, i.e., the intent as understood by the ranker.

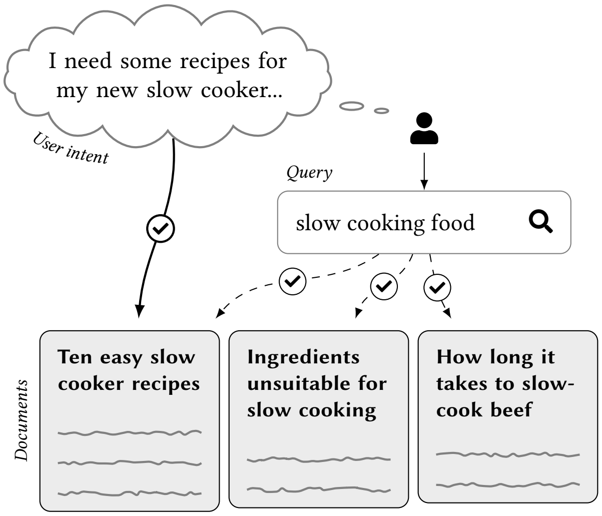

I need some recipes for my new slow cooker...
User intent
Query
slow cooking food
Documents
Ten easy slow cooker recipes
Ingredients unsuitable for slow cooking
How long it takes to slow-cook beef

Figure 1: An illustration of a user querying a search engine. The user has a specific intent in mind, but formulates the query in a more ambiguous way. As a result, there is a discrepancy between the documents relevant to the query and the documents relevant to the actual user intent.

This misalignment stems from the inherent complexity and variability in how users articulate their information needs versus how these needs are interpreted and processed by retrieval systems. This misalignment might be due to multiple reasons – ambiguity, poorly formulated queries, complex queries, or a retrieval set that lacks relevant documents [3, 11].

Most current research on ranking models in IR is based on training parameterized models over large training datasets from MS MARCO [13]. However, to the best of our knowledge, there exist no recent datasets that attempt to measure the chasm between user intent and machine intent. The current practice of measuring ranking performance is through sparsely [13] or densely annotated ad-hoc ranking test sets [5–9] that provide queries and corresponding relevance annotations. While these test sets allow for determining the overall effectiveness of a ranker, they fail to provide a way of measuring the extent to which the ranking models understand the true intent of the user. For example, consider the query "*what are the three countries in 1984*". While the intent—to identify the three countries mentioned in George Orwell's novel "1984"—seems clear,

CIKM '24, October 21–25, 2024, Boise, ID, USA
© 2024 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-0436-9/24/10.
https://doi.org/10.1145/3627673.3679166

it remains difficult to rank effectively because it requires specific contextual knowledge that may not be directly available in the retrieved documents. Another example is the query "*slow cooking food*" (cf. Fig. 1). Although this query appears to be straightforward, it can have multiple intents. This multiplicity of potential intents complicates the ranking process, as the system needs to correctly infer and prioritize the user's actual intent to provide relevant results. Knowing the user's intent allows the model to retrieve and rank documents most relevant to that intent, thereby addressing a critical challenge in handling ambiguous queries.

In this paper, we specifically focus on a subset of these challenges: *queries that contain multiple intents*. We propose a new dataset named DL-MIA (**M**S MARCO **I**ntent **A**nnotations), which is a derivative of the TREC-DL test sets. The DL-MIA dataset contains 2655 tuples of `(query, intent, passage, label)` over a small yet challenging set of 24 queries from the TREC-DL '21 and '22 datasets. To construct DL-MIA, the key challenge was to accurately formulate user intents, as only queries are available in the TREC-DL test sets. Toward this, we used a combination of LLM-generated query-specific intents and sub-intents that are post-processed through a carefully designed crowd-sourcing process to ensure human supervision and quality control. DL-MIA mainly aims at measuring the gap between user intent and query by *fine-grained intent annotation*, but can be used in multiple ranking scenarios, such as re-ranking, diversification, intent coverage, or query suggestion tasks.

Our contributions are twofold – first, we introduce a comprehensive dataset DL-MIA that meticulously documents the variations and complexities of user intent; second, we provide an analysis of this dataset's impact on ranking performance by applying it to several baseline models. DL-MIA is publicly available at https://zenodo.org/doi/10.5281/zenodo.11471482.

2 RELATED WORK

Several ranking datasets have been published that consider the concept of what we refer to as *user intents*. Most notably, the data provided for the TREC-Web track [3] customarily includes topics (queries) along with *topic descriptions* as well as, in many cases, *subtopics*. These subtopics represent various distinct aspects that each topic may have. The data further includes relevance judgments for documents from the ClueWeb collections w.r.t. the topics and subtopics. However, the TREC-Web track has been discontinued after 2014, and ClueWeb corpora are not freely available. Our dataset is similar, as the subtopics are essentially user intents.

The MS MARCO ranking dataset [13], which has emerged as one of the most widely used collections for IR-related tasks in recent years, contains a large number of training and evaluation queries. Furthermore, the TREC-DL track [5, 8] provides annotated test sets of queries and corresponding relevance annotations. More recently, the second version of the MS MARCO corpus, which is significantly larger than the first version, was released to be used in the TREC-DL 2021 track and onward [6, 7, 9].

Mackie et al. [11] showed that queries (topics) within TREC-DL vary with respect to their complexity (and, hence, difficulty) and released the DL-HARD dataset. Along with relevance annotations, this dataset assigns *intent categories* to each query. Similarly, *intent taxonomies* have been proposed for web search in general [2] as well as legal case retrieval [18]. The difference compared to our work is that we annotate specific user intents rather than categories.

Another related line of work deals with the *reformulation* of complex queries. Mackie et al. [12] recently released the CODEC collection for document and entity ranking, which also contains query reformulations. Salamat et al. [16] showed that the way queries are worded has an impact on their corresponding ranking performance. Our proposed user intents can be seen as reformulations that focus on specific aspects of the original query.

3 THE DL-MIA DATASET

In this section, we introduce the DL-MIA dataset by outlining the creation and annotation process and presenting some statistics.

3.1 Dataset Creation

The process of creating the dataset comprises several key stages: generating candidate intents using an LLM (Section 3.1.1), clustering and manual refinement of intents (Section 3.1.2), crowd-sourcing annotations (Section 3.1.3), merging similar intents (Section 3.1.4) and QRel creation (Section 3.1.5). This process is illustrated in Fig. 2.

3.1.1 Generating Candidate User Intents. For all queries in the TREC-DL-21 and '22 test sets, we retrieve all relevant passages using their respective QRel files. We then cluster similar passages per query. To achieve this, we first obtain passage embeddings using Sentence-BERT [14] and then group passages into the same cluster if their pairwise cosine similarity exceeds a threshold of 0.8. In the next step, we select the query and passages from the clusters to give to the LLM to generate five distinct intents relevant to the query-passage pairs. We employ the GPT-4 model with the prompt given below. We use a temperature value of 0.6 to control randomness which helps in getting diverse intents.

> **LLM Prompt**: Intent candidate generation
>
> A person wants to find out distinct intention behind the question **{query}**. Give five descriptive (max. 15 words) distinct intentions which are easy to understand. Consider all documents in your response. Response should be in this format:
> **Intention**:: <intention> , **Doc_list**::<list of documents with the intention>
> **Documents**: **{list of input documents}**

3.1.2 Clustering and Intent Selection. After generating intents, we cluster similar intents using the SBERT embedding and cosine similarity approach as described above. We group intents that are similar in meaning if their pairwise cosine similarity exceeds a threshold of 0.9. This clustering process helps in reducing redundancy and coming up with distinct intents. After clustering, we do manual selection, where we examine the clustered intents and choose the most relevant ones for each cluster. We do this to remove irrelevant intents or hallucinated text by the LLM. If any intents are found to be incomplete or poorly written, they are manually rewritten to improve their clarity and comprehensiveness. This ensures that the intents are well-defined and useful for the next stages of the

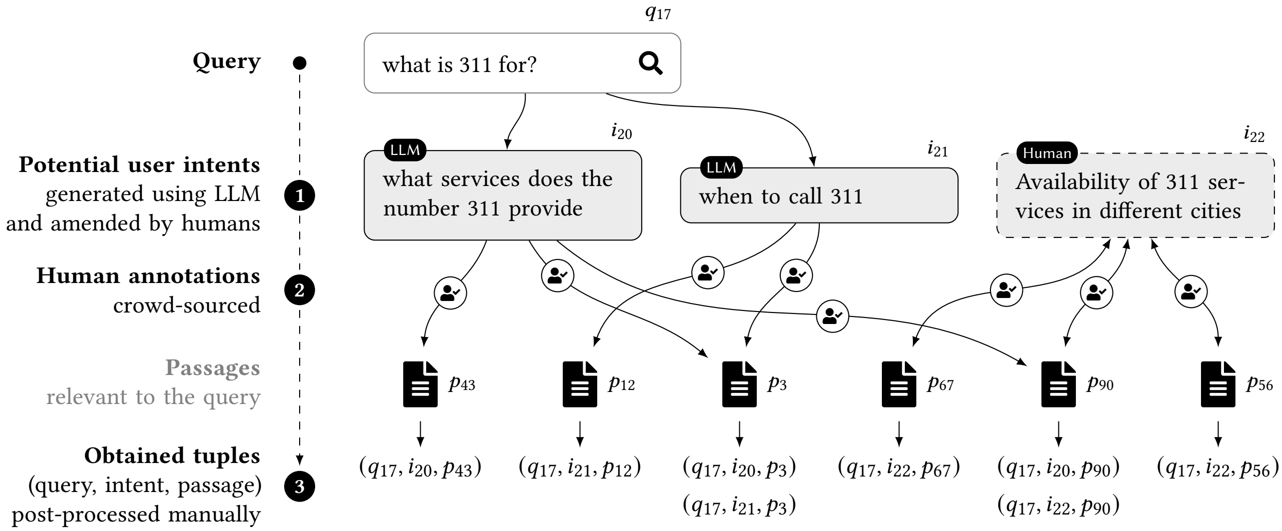
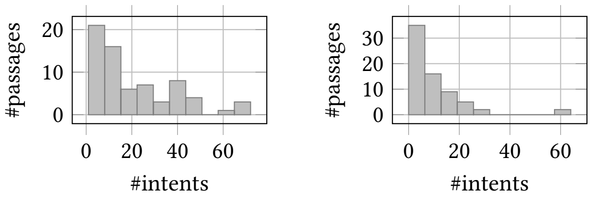


Figure 2: A high-level overview of how DL-MIA is created: Given a query, an LLM is used to generate candidate user intents. The query and its relevant passages (according to the original QRels), along with the candidate intents, are presented to human annotators, who can add, modify, or remove candidate intents and assign passages to them.

dataset creation process. After this process, only queries with 2 or more intents were selected which resulted in 26 queries.

3.1.3 Crowdsourcing Annotation. The next step involves crowdsourcing to annotate the intents with the relevant passages. Our pool of annotators comprises volunteers who are computer scientists and graduate school students familiar with ranking tasks for search. Annotators are presented with a query and a passage and are asked to determine which of the provided intents the passage satisfies. Additionally, annotators are given the option to add or modify intents if they find that the existing ones do not capture the passage's intent. To manage queries with a large number of relevant passages (more than 30), the passages are divided into smaller chunks of 30. This division creates subqueries, making the annotation process more manageable for the annotators. Each subquery is annotated separately, ensuring that the workload is distributed and the annotators can focus on a smaller set of passages. In total, 22 sets of annotations are done by 16 distinct annotators and each set consist of 5 rounds (queries or subqueries), such that each query is annotated at least twice.

3.1.4 Manual Review and Merging of Intents. In order to improve data quality and avoid redundancy, we conduct a manual review and merge intents. We evaluate the intents suggested by the annotators and integrate them into the existing set of intents where appropriate. E.g., in Fig. 2 we merge "when to call 311" and "when to call 311 rather than 911" into a single intent. Any passage-intent pair which does not have at least two annotators is dropped to ensure that the final set of intents reflects a consensus among multiple annotators. The merging process also helps in consolidating similar intents and removing any redundant or less relevant ones. After this process, we end up with 24 queries.



(a) rel ≥ 1

(b) rel = 2

Figure 3: Histograms illustrating the number of relevant passages per intent for (a) all relevant passages and (b) only passages with relevance label 2.

3.1.5 Scoring and Creating QRel File. Finally, we score the intent-passage pairs and create a QRel file for ranking. The scoring is based on the annotations provided by the participants. Each intent-passage pair is scored as follows: a score of 0 is assigned if no annotator marked the intent, a score of 1 is assigned if at least one annotator marked the intent, and a score of 2 is assigned if all annotators marked the intent. These scores reflect the level of agreement among the annotators and the relevance of the intent to the passage. The final query-intent-passage-score mappings are compiled into a QRel file, which is used for ranking. This QRel file serves as ground-truth for evaluating information retrieval systems, ensuring that the dataset can be effectively used for further research and application.

3.2 Statistics

Initially, the dataset included 118 queries from TREC-DL-21 and '22. Through a process of clustering and intent selection, 26 queries were identified as suitable for annotations, as these queries had two or more distinct intents (69 in total). After annotation (Sec. 3.1.3), a manual review and merging of intents were performed (Sec. 3.1.4). This process was necessary because the number of intents increased from 69 to 171 due to annotators adding custom intents. Hence, this review process was crucial in refining the dataset and ensuring the accuracy and clarity of the intents. After this rigorous review, 24 queries and 69 intents were finalized for inclusion in the dataset with **2655 relevance annotations** present in the final QRel file. The distribution of relevant passages per intent is shown in Fig. 3.

Because annotators were able to add custom intents, computing established agreement measures is difficult as the intents annotated by humans may have different granularities; however, the relevance scores we obtained in Sec. 3.1.5 are determined by the overlap of judgments and can therefore be seen as an indication of agreement among annotators.

3.3 Tasks and Evaluation

The DL-MIA dataset can be used for several tasks, such as:

Intent-based ranking aims at improving the document ranking by understanding different user intents and ensuring that the returned documents are relevant to the intent. This can be evaluated using metrics like nDCG@10.

Diversity of search results aims at ensuring that document rankings provide diverse sets of responses that cover various aspects of the query to satisfy users information needs, evaluated using metrics like α-nDCG@10.

Intent-based summarization aims at generating a summary that covers multiple intents of a query, evaluated using metrics such as ROUGE or BLEU.

User and machine intent alignment aims at bridging the gap between user and machine intent through query rewriting to fully specify the intent [1]. DL-MIA aids in training generative models that can generate intents more aligned with real-world user intents.

4 EXPERIMENTS

In order to demonstrate the utility of DL-MIA, we conduct experiments using a number of simple baselines: **BM25** [15] is a lexical model which is also used as a first-stage retriever for re-rankers. **BERT** [10] is a cross-attention re-ranker (BERT-base, 12 layers). The input length is restricted to a maximum of 512 tokens. The model is trained on MS MARCO passage data using a pointwise ranking loss objective with a learning rate of 1e-5. **ColBERTv2** [17] is a multi-vector late-interaction re-ranking model that computes token-wise representations for the query and document and estimates relevance using the MaxSim operation.

4.1 Results

We report results on two of the tasks outlined in Section 3.3, namely *intent-based ranking* and *diversity of search results*. We present these results in Table 1. Note that we evaluate two settings: First, we use the original queries, but evaluate using the user intent-based QRels (i.e., assuming that the user had one specific intent in mind). Second, we treat the user intents as queries directly. The results show that, unsurprisingly, specifying the actual user intent as the query results in better performance than using the (more general) original queries. We additionally demonstrate the diversity ranking performance of various models using the α-nDCG@10 metric. To achieve this in the second setting (where user intents are treated as queries), we employ reciprocal rank fusion [4] with $k = 60$. This technique is applied to the intent-based rankings to generate a unified ranking for the original query. Overall, ColBERTv2 shows the best performance. Finally, we closely examine the ranking performance corresponding to each user intent in Fig. 4. The results are in line with Table 1.

	Intent Ranking	**Diversity**
	nDCG@10	α-nDCG@10
Original queries, user intent QRels		
BM25	0.073	0.144
BERT	0.060	0.114
User intents as queries, user intent QRels		
BM25	0.116	0.250
BERT	0.169	0.375
ColBERTv2	**0.261**	**0.532**

Table 1: DL-MIA ranking performance. Best performing models are in bold. Re-rankers use the corresponding BM25 runs. Diversity is calculated at query level in both cases.

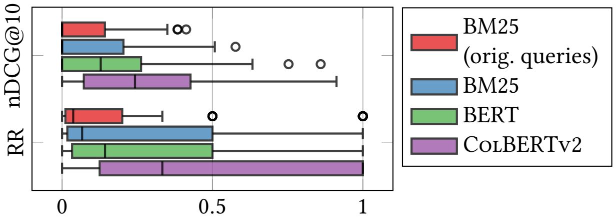


Figure 4: Performance comparison on a per-intent level. The boxplots show the distribution of the ranking performance of individual intents.

The key takeaway from these results is the necessity of specifying concrete user intents; in other words: if a user has a specific information need, it is necessary to provide that intent as a clear, unambiguous query to a search engine.

5 CONCLUSION

In this paper, we have created the DL-MIA dataset to understand user intents, thereby satisfying information needs more effectively. We have used queries from TREC-DL-21 and TREC-DL-22, generated intents using an LLM, and crowd-sourced relevance annotations. DL-MIA can be used for a variety of tasks; we present performance of different models on ranking and diversity tasks, showing the importance of this dataset for fulfilling user information needs. For future work, we plan to extend DL-MIA to include queries from TREC-DL-19, TREC-DL-20, and DL-HARD.

ACKNOWLEDGMENTS

This work is partially supported by German Research Foundation (DFG), under the Project IREM with grant No. AN 996/1-1.

REFERENCES

[1] Avishek Anand, Venktesh V, Abhijit Anand, and Vinay Setty. 2023. Query Understanding in the Age of Large Language Models. arXiv:2306.16004 [cs.IR]

[2] B. Barla Cambazoglu, Leila Tavakoli, Falk Scholer, Mark Sanderson, and Bruce Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (Canberra ACT, Australia) *(CHIIR '21)*. Association for Computing Machinery, New York, NY, USA, 85–94. https://doi.org/10.1145/3406522.3446027

[3] Kevyn Collins-Thompson, Craig Macdonald, Paul N Bennett, Fernando Diaz, and Ellen M Voorhees. 2014. TREC 2014 Web Track Overview.. In *TREC*, Vol. 13. 1–15.

[4] Gordon V Cormack, Charles LA Clarke, and Stefan Buettcher. 2009. 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*. 758–759.

[5] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. arXiv:2102.07662 [cs.IR]

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Jimmy Lin. 2022. Overview of the TREC 2021 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2021-deep-learning-track/

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2023. Overview of the TREC 2022 deep learning track. In *Text REtrieval Conference (TREC)*. NIST, TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2022-deep-learning-track/

[8] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. arXiv:2003.07820 [cs.IR]

[9] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Hossein A. Rahmani, Daniel Campos, Jimmy Lin, Ellen M. Voorhees, and Ian Soboroff. 2024. Overview of the TREC 2023 Deep Learning Track. In *Text REtrieval Conference (TREC)*. NIST, TREC. https://www.microsoft.com/en-us/research/publication/overview-of-the-trec-2023-deep-learning-track/

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *CoRR* abs/1810.04805 (2018). http://arxiv.org/abs/1810.04805

[11] Iain Mackie, Jeffrey Dalton, and Andrew Yates. 2021. How Deep is your Learning: the DL-HARD Annotated Deep Learning Dataset. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Virtual Event</city>, <country>Canada</country>, </conf-loc>) *(SIGIR '21)*. Association for Computing Machinery, New York, NY, USA, 2335–2341. https://doi.org/10.1145/3404835.3463262

[12] Iain Mackie, Paul Owoicho, Carlos Gemmell, Sophie Fischer, Sean MacAvaney, and Jeffrey Dalton. 2022. CODEC: Complex Document and Entity Collection. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (<conf-loc>, <city>Madrid</city>, <country>Spain</country>, </conf-loc>) *(SIGIR '22)*. Association for Computing Machinery, New York, NY, USA, 3067–3077. https://doi.org/10.1145/3477495.3531712

[13] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A Human Generated MAchine Reading COmprehension Dataset. (November 2016). https://www.microsoft.com/en-us/research/publication/ms-marco-human-generated-machine-reading-comprehension-dataset/

[14] Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084* (2019).

[15] Stephen Robertson and Hugo Zaragoza. 2009. The Probabilistic Relevance Framework: BM25 and Beyond. *Found. Trends Inf. Retr.* 3, 4 (apr 2009), 333–389. https://doi.org/10.1561/1500000019

[16] Sara Salamat, Negar Arabzadeh, Shirin Seyedsalehi, Amin Bigdeli, Morteza Zihayat, and Ebrahim Bagheri. 2023. Neural Disentanglement of Query Difficulty and Semantics. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management* (<conf-loc>, <city>Birmingham</city>, <country>United Kingdom</country>, </conf-loc>) *(CIKM '23)*. Association for Computing Machinery, New York, NY, USA, 4264–4268. https://doi.org/10.1145/3583780.3615189

[17] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. ColBERTv2: Effective and Efficient Retrieval via Lightweight Late Interaction. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Marine Carpuat, Marie-Catherine de Marneffe, and Ivan Vladimir Meza Ruiz (Eds.). Association for Computational Linguistics, Seattle, United States, 3715–3734. https://doi.org/10.18653/v1/2022.naacl-main.272

[18] Yunqiu Shao, Haitao Li, Yueyue Wu, Yiqun Liu, Qingyao Ai, Jiaxin Mao, Yixiao Ma, and Shaoping Ma. 2023. An Intent Taxonomy of Legal Case Retrieval. *ACM Trans. Inf. Syst.* 42, 2, Article 62 (dec 2023), 27 pages. https://doi.org/10.1145/3626093