How Deep is your Learning: the DL-HARD Annotated Deep Learning Dataset

Iain Mackie
University of Glasgow
Glasgow, Scotland, UK
i.mackie.1@research.gla.ac.uk

Jeffrey Dalton
University of Glasgow
Glasgow, Scotland, UK
jeff.dalton@glasgow.ac.uk

Andrew Yates
Max Planck Institute for Informatics
Saarbrücken, Germany
ayates@mpi-inf.mpg.de

ABSTRACT

Deep Learning Hard (DL-HARD) is a new annotated dataset designed to more effectively evaluate neural ranking models on complex topics. It builds on TREC Deep Learning (DL) topics by extensively annotating them with question intent categories, answer types, wikified entities, topic categories, and result type metadata from a commercial web search engine. Based on this data, we introduce a framework for identifying challenging queries. DL-HARD contains fifty topics from the official DL 2019/2020 evaluation benchmark, approximately half of which are newly and independently assessed. We perform experiments using the official submitted runs to DL on DL-HARD and find substantial differences in metrics and the ranking of participating systems. Overall, DL-HARD is a new resource that promotes research on neural ranking methods by focusing on challenging and complex topics.

CCS CONCEPTS

• **Information systems** → **Information retrieval**.

KEYWORDS

Deep learning dataset; neural ranking models; semantic query annotation

ACM Reference Format:
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 (SIGIR '21), July 11–15, 2021, Virtual Event, Canada.* ACM, New York, NY, USA, 7 pages. https://doi.org/10.1145/3404835.3463262

1 INTRODUCTION

The development of new machine learning models for ranking is an important area of Information Retrieval research, with a recent emphasis on neural language models [24]. These language models are state-of-the-art for both retrieval [13, 15, 16] and natural language understanding tasks [2, 10, 19]. They are used by leading commercial web search engines to improve ranking and question answering (QA) effectiveness.[1] The focus of this resource is to support the measurement of progress on challenging ranking topics where these new classes of models fail.

The MS MARCO leaderboard is a leading benchmark for both passage and document ranking. It uses real web queries that are candidates from Bing's web QA system. Neural language models have made significant improvements on these types of question-intent queries due to their longer natural language nature.

MS MARCO contains many queries with sparse relevance labels, whereas the TREC Deep Learning track provides a smaller subset assessed more deeply by professional assessors. The large volume of sparse MS Marco data for training, and the high-quality NIST judgments for evaluation, result in DL being a significant step forward for the community. As part of the publicly available DL-HARD dataset, we augment DL with rich manual and automatic query annotations on all four hundred queries (assessed and not assessed). These rich annotations include Question Intent Types [3], our own specially developed answer types, result types from a leading web search engine, coarse topic categories, and automatic and gold entity mentions linked to Wikipedia (Figure 1). Such annotations enable developing new methods for identifying 'hard' queries to inform future benchmark construction.

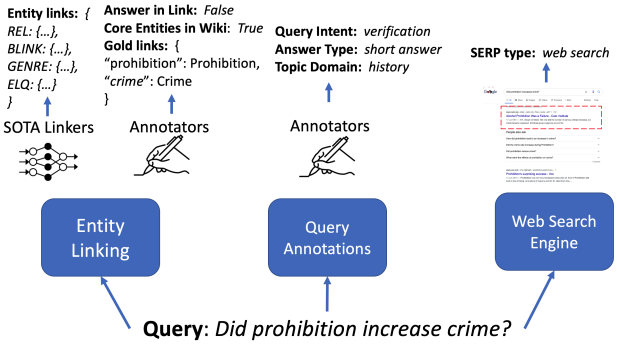

Figure 1: DL-HARD annotation process overview.

Using DL to test modern entity-centric and neural ranking algorithms for long documents is challenging. First, the reported system effectiveness is relatively high, even for existing baseline systems, with a median mean reciprocal ranking above 0.8 for both the DL 2020 document ranking and passage ranking tasks. This appears to show that current deep learning methods leave little headroom for improvement. This work demonstrates that this is not the case in practice, motivating the need for a resource that builds on proven

[1] https://blog.google/products/search/search-language-understanding-bert/

SIGIR '21, July 11–15, 2021, Virtual Event, Canada

ACM ISBN 978-1-4503-8037-9/21/07…$15.00
https://doi.org/10.1145/3404835.3463262

DL data and provides the headroom required for modern systems. Second, the DL queries vary in terms of difficulty, intent, answer type, etc. What are the 'right' queries to focus on when evaluating state-of-the-art neural models? Current commercial web search engines are already tuned to rapidly answer many diverse queries. Based upon studying web search engine behaviour on DL, we find that a significant proportion of the queries are 'solved'. For example, many are factoid questions that can be answered from Knowledge Graphs or with lookups from structured data sources.

To address these issues, we introduce the DL-HARD dataset[2] that consists of 'hard' DL topics from 2019 and 2020. DL-HARD provides annotations on the full four hundred queries and a benchmark consisting of the fifty most difficult queries across both years. These include approximately twenty-five previously assessed queries and twenty-five queries with new sparse judgments annotated at a passage and document level (Figure 2).

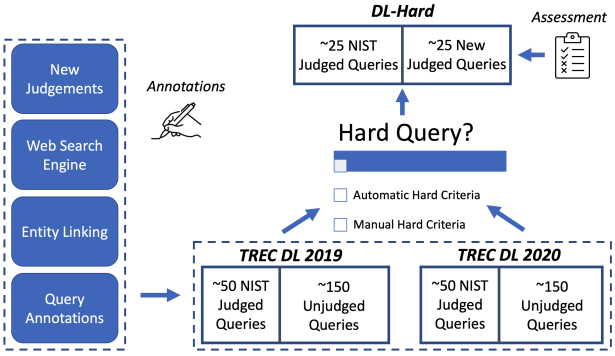

Figure 2: DL-HARD dataset overview. NIST judged/unjudged DL query counts are approximated for simplicity (see track overview for specifics [6]). Due to the differences in DL task querysets, DL-HARD provides 25 new document judgments and 27 new passage judgments.

We perform an empirical evaluation of DL-HARD topics on all the submitted runs to DL 2020. We find that DL-HARD topics are substantially more difficult and lead to swaps in the ranking of systems. When considering the twenty-five queries assessed by DL, each system moved on average 4.6 places in the overall system ranking. For the twenty-five queries with new judgments, this results in even larger changes to the relative ordering of systems. We note that DL-HARD queries are more complex and contain a higher fraction of list and long answer results than DL.

The key contributions of this resource include:

- Diverse manual and automatic annotations for all DL topics: intent type, answer type, search engine result type, topic category, and wikified entities.
- A new DL-HARD benchmark for both document and passage tasks. Approximately half the queries are newly assessed.
- A semi-automatic method for identifying challenging queries that leverages evidence from commercial web search systems, query intent, and other metadata.
- A study of the behavior of official DL submissions on DL-HARD.

[2]DL-HARD is available at https://github.com/grill-lab/DL-HARD

2 RELATED WORK

The MS MARCO passage and document collections [14] consist of queries, web passages or documents, and sparse relevance judgments between them. This dataset derives passage-level relevance judgments from the MS MARCO Question Answering dataset by treating any passage containing a correct answer for the query to be relevant. These passage-level judgments are transferred to the document level by labelling any document containing a relevant passage to also be relevant. While this label transfer approach requires little manual effort and enables the creation of a large dataset, it has several issues that may make it artificially easy. First, all queries have an associated answer from the QA dataset. Second, all queries have a single passage answer. Finally, the set of documents is limited in scope to ones that are a passage candidate for one of the queries.

The queries in MS MARCO are longer than typical web queries (5.8 words on average across DL), which are more challenging to handle than short keyword queries [11]. However, many are factoid or begin with 'wh-' words, which may be easier than more open-ended long queries.

The TREC Deep Learning (DL) track [6, 7] ran tasks using the MS MARCO passage and document collections. Assessments by NIST annotators address the label sparsity issue by providing judgments pooled to a greater depth. We address the query difficulty issue by building upon DL and identifying the most complex queries. In this work, we define web query answer types with a new taxonomy developed bottom-up for MS MARCO to help categorise challenging and interesting topics. The developed 'hard criteria' helps to select these topics systematically.

Cambazoglu et al. [3] study the types of queries in MS MARCO. They create a taxonomy of intents for questions, the types of named entities present, the types of question words used, and the answer's expected granularity. We manually annotate all DL queries with intents from their taxonomy and additionally introduce a complementary schema that is more fine-grained. Similarly, the task of determining a question's answer type has been widely studied in QA. For example, by associating salient terms in the question, such as 'wh-' words, with answer types identified in a corpus [18].

Our dataset also complements previous work in QA, which generally refers to identifying a relevant text span in response to a question. The TREC Question Answering track [22] constructed a series of QA benchmarks beginning in the late 1990s. More recently, the SQuAD [20] and Natural Questions [12] benchmarks each provide over 100,000 crowdsourced questions and associated answers. In contrast to seeking factoid or short answers, DL-HARD dataset focuses on complex answers that can be long and multi-faceted.

3 TASK AND JUDGMENTS

We now describe the resource task and relevance assessment for the DL-HARD dataset.

Task Definition. The task is an information-seeking passage and document ranking task that follows the one described in the TREC Deep Learning track [7]. Because the use case for DL-HARD emphasizes challenging ad-hoc retrieval, the query intents are more likely long descriptions, multiple answers, a list, or require reasoning. The criteria used to filter queries include: spelling errors, incomplete, ambiguous, or target a specialized structured vertical,

i.e. calculator, maps, weather, or dictionary. Additionally, since factoid QA is already a well-studied area in TREC [22] and the NLP community [20], DL-HARD queries are primarily non-factoid.

For the experimental setup, we provide five pre-defined folds that can be used for k-fold cross-validation. The dataset can also be used purely as a test set. DL-HARD can evaluate end-to-end retrieval or re-ranking leveraging provided baseline runs.

Relevance Assessment Process. The resource uses the full NIST assessments for previously judged topics. There are also new passage and document level judgments provided for unjudged queries from DL.

We perform relevance assessment on a graded scale using the same guidelines as the original track for the new judgments. We assess passages returned in the MS MARCO QA corpus and the documents they are drawn from. Unlike the MS MARCO sparse judgments, which generally include only one relevant passage per query, we assess all of the top ten responses. Experienced IR researchers (the authors) perform the annotations.

To calculate agreement with the NIST assessors, we additionally judge the top QA passage responses for 24 queries from DL (12 from each year). We find Krippendorff's alpha is 0.47 on the passage judgments for these queries and 0.43 on the document judgments, which indicates moderate agreement. Krippendorff's alpha drops to 0.12 when transferring passage assessments to documents, illustrating the difficulty of automatically transferring passage assessments. For this reason, we adopt document-level relevance judgments for the official DL-HARD document ranking task.

4 ANNOTATIONS

We detail the annotations provided within the resource. We use these annotations in Section 5 to develop the criteria for selecting hard topics for the DL-HARD dataset.

4.1 Question Intent Annotation

We apply the question intent taxonomy developed for MS MARCO web questions [3]. In contrast to other taxonomies, this has a more fine-grained taxonomy developed bottom-up for MS MARCO. We use their *Query Intent Categories* and guidelines to annotate all official DL queries. At least one author performs each annotation, and ambiguous instances resolved by majority vote. To our knowledge, this is the first resource to make these annotations publicly available.

The distribution of the query intents on the complete DL queryset as well as DL-HARD is shown in Table 1. The most notable difference is the increase in List intents from DL (10.2% across 2019 and 2020) to DL-HARD (34.7%). The annotators note that list queries are harder as the user seeks multiple entities or facts that could span many documents. The proportion of Quantity intents in DL-HARD is much lower as most of these queries are either simple factoid-QA questions ('hydrogen is a liquid below what temperature') or highly underspecified and should be clarified (i.e. 'cost of interior concrete flooring'). DL-HARD also filters out Language and Weather intents.

4.2 SERP Result Types

To retrieve the Search Engine Results Page (SERP), we manually issue every query on a Desktop browser to an English language Google search engine from the United Kingdom in 'incognito mode'. The authors inspect the results and save the raw HTML content to include in the resource. We note queries with potential localization issues (location, region-specific language, or time). Based on the criteria described in Section 5, we exclude these queries because they are unanswerable without local context that is not provided.

Table 1: Query Intent Categories for DL and DL-HARD (document ranking).

Intent Category	DL-2019	DL-2020	DL-HARD
Attribute	1	5	1
Description	21	20	20
Entity	3	4	3
Language	0	2	0
List	7	2	17
Process	1	1	0
Quantity	5	6	3
Reason	3	4	4
Verification	1	1	2
Weather	1	0	0

For each query, we annotate the type of rich results returned in the SERP and whether the Knowledge Graph [17] is used (the raw HTML shows the schema elements). Although many possible types of rich results may be present in a SERP, the ones highlighted below are the most prevalent for DL queries:

- *Spell correct or suggestion*: Shows a suggested spelling correction or alternative query.
- *Knowledge Graph (KG)*: Returns a specific answer entity, list of entities, or their attributes from structured entity data. This includes media structured results for television, movie, and music entity information.
- *Dictionary*: Provides a dictionary definition of one or more words.
- *Weather*: Shows the weather forecast for a locale via an embedded panel.
- *Map*: Shows a Maps vertical result, optionally with possible driving directions.
- *Web Short Answer*: Shows a specific string short answer, possibly with a separate supporting evidence passage from a web result.
- *Web Passage*: Shows a passage (or portion of a list or table) from a web result. It may highlight possible answers.
- *Web Search*: Shows a standard list of '10 blue links'.

The distribution of the response types for DL assessed and DL-HARD topics is shown in Table 2. By far, the most frequent response type is a Web Passage, which is unsurprising given that the queries are questions originally used for QA. It shows that over 20% of the queries are answered directly with short factoid answers, with 12.5% of results from a structured source. Although the Google answer quality is not explicitly assessed, we observe only 2 instances of clear failure due to imprecise and/or ambiguous queries. This indicates that existing models (neural or otherwise) can adequately satisfy these 'easy' factoid queries.

Table 2: SERP result types distribution for DL Track and DL-HARD (document ranking).

SERP Result	DL-2019	DL-2020	DL-HARD
Dictionary	1	3	1
KG	1	5	2
Weather	1	0	0
Web Passage	24	25	28
Web Search	12	9	18
Web Short Answer	4	3	1

We evaluate whether the SERP answer type is a good heuristic for systematically identifying hard queries by mapping the SERP annotations onto 2019/2020 DL runs. The expectation is that queries within the Web Search category should be a reasonable proxy for a hard query, i.e. either (1) the search engine could not find an answer for the query, or (2) the query could not be satisfied by a short passage or entity. Based on an analysis of assessed DL queries on the best DL systems, defined as those with above-median NDCG@10, this trend holds (Table 3). Statistical analysis shows much lower NDCG@10 and Recall@100 with negative Pearson correlation coefficients. Supporting Web Search answer type as a primary feature in the 'automatic hard criteria' in Section 5.

Table 3: SERP result type on DL 2019/20 document ranking systems (systems above median). Mean and Pearson Correlation Coefficient (PCC) across DL assessed queries.

	NDCG@10		Recall@100	
SERP Result	**Mean**	**PCC**	**Mean**	**PCC**
KG	0.577	-0.05	0.794	0.11
Dictionary	0.748	0.13	0.722	0.04
Weather	0.735	0.05	0.464	-0.06
Web Passage	0.647	0.13	0.684	0.04
Web Search	0.535	-0.20	0.581	-0.16
Web Short Answer	0.621	0.00	0.731	0.05

4.3 Answer Type Annotations

Previous manual and automatic annotations focus on the type of question intent or the SERP result type. Therefore, we create a new target answer type for MS MARCO web queries. The manual answer type labels are from all authors with a majority vote resolution. To develop the types, we follow a bottom-up multi-round curation similar to that used for query intents [3]. The answer types are:

- *Definition* - A single passage precisely and completely answers the information need. These are most commonly associated with the Description and Language query intents.
- *Factoid* - A specific short fact answer to a question. These are often associated with Entity, Attribute, Quantity, and Location intent types.
- *Short answer* - A short passage (approximately a sentence) generally satisfies most information needs. These are usually associated with Description and other factoid-like intents.

Table 4: Answer Type distribution for DL Track and DL-HARD (document ranking).

Answer Type	DL-2019	Dl-2020	DL-HARD
Comparison	3	2	0
Definition	9	7	7
Factoid	12	24	5
Guide	0	1	0
List	9	0	15
Long Answer	6	10	13
Multi-Answer	0	1	0
Short Answer	3	0	9
Short Description	0	0	1
Weather	1	0	0

- *Long answer* - A long passage or full document is needed to answer the query. These are associated with Description, List, and Process intents.
- *List* - More than one answer, passage, or entity with justification is needed to answer the query.
- *Maps* - A structured map answer is needed; this is associated with Location and local Calculation intents.
- *Weather* - A structured weather result; corresponds to the Weather intent type.
- *Comparison* - A comparison of two or more entities. These are associated with Description intent types.
- *Guide* - A guide answer is a long semi-structured answer to satisfy the Process intent.

The answer types have strong associations with query intent types. However, we find that the Description intent is often quite general and does not provide guidance on the type of information needed for the answer. This is important because these answer types are useful features for topic complexity (see Section 5).

Table 4 shows the answer type breakdown for the assessed DL and DL-HARD topics. Compared with DL topics, it is clear that there are fewer Factoid responses and more List answers within DL-HARD.

4.4 Query Entity Annotation

Entity linking [5] and semantic parsing [1] of question queries is an important component of modern QA systems. However, the existing DL queries do not have standard automatic or manual annotations. We provide both as part of DL-HARD.

We include four state-of-the-art entity linkers developed for documents and queries: REL [21], Blink [23], Genre [4], and ELQ [23]. We run these annotators with high-recall score thresholds, preserving score information for downstream applications, which is important for entity-based retrieval models [8]. Based upon the automatic results, we create gold entity links to Wikipedia and metadata about the entities, i.e. (1) whether the query entity is in Wikipedia and (2) whether the Wikipedia entity satisfies the query.

4.5 Coarse Topic Categories

Following TREC Conversational Assistance Track (CAsT) [9], we provide a breakdown of topics by coarse subject domain in Table 6. For 2019 we observe frequent DL topic categories to be Health,

Table 5: Automatic hard criteria for categorising hard queries (88 labelled DL assessed document ranking topics).

Include		Exclude			
SERP	**Query Intent**	**Query Intent**	**Precision**	**Recall**	**F1**
Web Search			0.428	0.360	0.391
	List, Reason		0.588	0.400	0.476
	List, Reason, Entity		0.500	0.480	0.490
Web Search	List, Reason		0.486	0.680	0.566
Web Search	List, Reason, Entity		0.450	**0.720**	0.553
Web Search	List, Reason, Entity	Quantity, Weather, Language	0.529	**0.720**	0.610
Web Search	List, Reason	Quantity, Weather, Language	**0.586**	0.680	**0.630**

Table 6: Topic domain category for DL Track and DL-HARD (document ranking).

Topic Domain	DL-2019	DL-2020	DL-HARD
Business & Finance	1	6	3
Entertainment & Celebrity	0	9	0
Food & Travel	5	3	4
Health	10	4	20
History & Education	5	6	8
Language & Literature	1	4	2
Law & Politics	2	2	2
Local	1	0	1
Mathematics & Science	13	6	10
Sports	1	3	1
Technology	4	2	0

Science, and History. In 2020 there is a shift to more Entertainment, Business and Finance topics, and less Science and Health. There is also an increase in Language topics, with predominately definition queries. The largest category for DL-HARD is Health, a challenging category that often requires long answer responses.

5 HARD CRITERIA

New challenging and complex benchmark topics are required to differentiate system performance of neural ranking models. Because manually judging all candidate queries is time-consuming, we develop an 'automatic hard criteria' to generate candidate queries scalably. Each candidate topic is then manually labelled using the 'manual hard criteria' by multiple assessors. This process identifies 50 'hard' topics within the 400 DL topics.

5.1 Automatic Hard Criteria

Given that manually reviewing or creating judgments for all candidate topics is time-consuming, we explore the use of annotated metadata to generate a hard dataset.

For simple and explainable criteria, we test explicit rule-based inclusion and exclusion filters. We measure their agreement with the human labels and the effectiveness of existing systems on the 88 assessed DL queries (25 labelled 'hard' and 63 labelled 'not hard'). We present precision, recall and F1 results in Table 5. We observe that the most effective rule uses Google's SERP answer type (Web Search) as a base with additional List and Reason query intents added to improve recall. We exclude intent types matching Quantity, Weather, and Language (mostly dictionary lookups). We see that adding Entity queries improves recall, but these queries also include several 'easy' factoid questions.

Although results were relatively encouraging, particularly for identifying potential hard queries, we require the 'manual hard criteria' to ensure only the optimal queries are selected. Thus, the additional 25 unassessed DL-HARD queries (from a possible 312) combined automatic and manual criteria for labelling. More advanced methods for automatically selecting hard queries is an area for additional future exploration.

5.2 Manual Hard Criteria

Hard queries are those where current models are not effective. However, not all queries where systems fail are challenging for 'interesting' reasons; it could be due to missed stopwords or tokenization issues, i.e. 'why did the us volunterilay enter ww1'. Additionally, under-specified queries are hard for assessors and search engines to answer definitively, i.e. 'who is robert gray' (multiple Robert Grays) or 'cost of interior concrete flooring' (local and ambiguous).

The authors consider both when and how systems struggle. We consider behavior in a first pass candidate retrieval (candidate recall) and second pass re-ranking (retrieval in top ranks). Queries with either type of failure are candidates for inclusion in DL-HARD. Each candidate topic is individually labelled and resolved across all annotators. These discussions inform the guidelines developed:

- *Non-Factoid* - The query should not be answerable by a single short answer, possibly from a KG.
- *Beyond single passage* - The query should require more than a simple definition or Wikipedia short description.
- *Answerable* - The topic should be answerable solely from the provided query because additional long description or narratives are not provided. Queries depending on external context should also be removed (i.e. location, temporal, etc.).
- *Text-focused* - Queries that require non-text answers or calculation of quantities should be handled by specialized components and excluded.
- *Mostly well-formed* - The query should not contain clear spelling or language errors that would be filtered by an initial query rewriting step.
- *Possibly Complex* - A query is desirable for inclusion if it references multiple entities, seeks a comparison, requires complex reasoning, or has multiple answers.

Table 7: Top 20 systems' effectiveness on DL-HARD compared with DL for the 2020 document ranking task.

	NDCG@10			Reciprocal Rank (RR)			Recall@1000		
System	**DL-HARD**	**DL**	**% Diff**	**DL-HARD**	**DL**	**% Diff**	**DL-HARD**	**DL**	**% Diff**
ICIP_run1	0.452	0.662	-21.1%	0.510	0.736	-22.7%	0.484	0.692	-20.8%
d_d2q_duo	0.449	0.693	-24.5%	0.472	0.734	-26.2%	0.690	0.842	-15.3%
fr_doc_roberta	0.442	0.640	-19.9%	0.524	0.733	-20.9%	0.641	0.788	-14.6%
d_d2q_rm3_duo	0.438	0.690	-25.2%	0.479	0.735	-25.6%	0.664	0.860	-19.6%
mpii_run2	0.432	0.613	-18.1%	0.468	0.677	-20.9%	0.484	0.692	-20.8%
bcai_bertb_docv	0.431	0.628	-19.6%	0.416	0.739	-32.3%	0.581	0.760	-17.9%
ICIP_run3	0.431	0.653	-22.2%	0.536	0.755	-21.9%	0.484	0.692	-20.8%
bigIR-DTH-T5-F	0.425	0.591	-16.5%	0.559	0.681	-12.1%	0.581	0.736	-15.5%
d_rm3_duo	0.424	0.679	-25.5%	0.467	0.733	-26.6%	0.622	0.826	-20.4%
ndrm3-full	0.415	0.616	-20.1%	0.448	0.716	-26.8%	0.609	0.780	-17.1%
ICIP_run2	0.413	0.632	-21.9%	0.489	0.733	-24.4%	0.484	0.692	-20.8%
ndrm3-re	0.409	0.616	-20.7%	0.455	0.713	-25.9%	0.484	0.692	-20.8%
roberta-large	0.408	0.629	-22.2%	0.465	0.739	-27.4%	0.484	0.692	-20.8%
mpii_run1	0.407	0.602	-19.4%	0.494	0.696	-20.2%	0.484	0.692	-20.8%
bigIR-DTH-T5-R	0.407	0.603	-19.6%	0.507	0.697	-19.0%	0.484	0.692	-20.8%
bigIR-DH-T5-F	0.404	0.573	-16.9%	0.572	0.659	-8.8%	0.581	0.736	-15.5%
ndrm3-orc-full	0.402	0.625	-22.3%	0.486	0.719	-23.3%	0.611	0.784	-17.3%
ndrm3-orc-re	0.397	0.622	-22.5%	0.419	0.692	-27.3%	0.484	0.692	-20.8%
TUW-TKL-2k	0.396	0.585	-18.9%	0.462	0.689	-22.7%	0.484	0.692	-20.8%
TUW-TKL-4k	0.393	0.575	-18.2%	0.486	0.694	-20.8%	0.484	0.692	-20.8%
Mean	**0.419**	**0.626**	**-20.8%**	**0.486**	**0.714**	**-22.8%**	**0.545**	**0.736**	**-19.1%**

6 RESOURCE EXPERIMENTS

We measure official TREC 2020 document run submissions on DL-HARD and compare to the original DL Track to (1) determine whether the dataset differs in system behavior and (2) measure differences in system rankings (swaps) on this dataset. For binary metrics, we consider labels of two or greater to be relevant.

The 2020 system effectiveness for DL Track, DL-HARD and the relative differences is shown in Table 7. On an average relative basis, DL-HARD NDCG@10 is 21.1% lower, RR is 23.2% lower, and Recall@1000 is 19.6% lower. There are similar findings when evaluating the 2019 document task and shows headroom for system improvement.

Additionally, many system swaps occur when comparing the DL system rankings to DL-HARD rankings. This includes a new top system ('ICIP_run1'), and each system changes on average 4.6 places, with some systems changing as many as 12 places. This is supported by Kendall's Tau coefficients of 0.696 (2019) and 0.641 (2020) when comparing TREC DL Track and DL-HARD system rankings. This large number of swaps supports that removing easier queries allows for greater differentiation and more precise system comparison.

Similarly, we evaluate the 2019 and 2020 DL systems on the 25 new sparse annotations using the official runs. These results cannot be directly compared to the DL Track as these queries have new judgments. Nonetheless, the top 10 systems only have an NDCG@10 of 0.314 and RR of 0.452, indicating DL-HARD topics with new judgments are challenging for modern systems.

7 CONCLUSION

We introduce the DL-HARD dataset resource for evaluating modern deep learning ranking models. It provides a challenging set of topics with new annotations: question intent types, answer types, categories, entity links, and metadata from Google SERPs. We contribute new judgments for queries not previously assessed by NIST. All of the annotations and assessments are publicly and freely available for use, and all data is non-personal and anonymized.

DL-HARD develops automatic and manual criteria for categorising complex queries, which is applicable when constructing future datasets. We use DL-HARD to compare the overall system effectiveness of systems in the TREC 2020 DL track. We find significant differences in system ordering and an overall reduction in effectiveness (headroom for future research). This resource represents an important step towards more challenging datasets for passage and document ranking.

8 FUTURE WORK

The authors plan future work to remove duplicates found in the MS MARCO collection, add explicit long descriptions to topics to remove ambiguity, and add an entity ranking task to complement the current document/passage ranking tasks.

9 ACKNOWLEDGEMENTS

This work is supported by the Engineering and Physical Sciences Research Council grant EP/V025708/1, the 2019 Bloomberg Data Science Research Grant, and the TensorFlow Research Cloud.

REFERENCES

[1] Jonathan Berant, A. Chou, Roy Frostig, and Percy Liang. 2013. Semantic Parsing on Freebase from Question-Answer Pairs. In *EMNLP*.

[2] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 1877–1901. https://proceedings.neurips.cc/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf

[3] B. B. Cambazoglu, Leila Tavakoli, F. Scholer, M. Sanderson, and B. Croft. 2021. An Intent Taxonomy for Questions Asked in Web Search. *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval* (2021).

[4] Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2021. Autoregressive Entity Retrieval. In *International Conference on Learning Representations*. https://openreview.net/forum?id=5k8F6UU39V

[5] M. Cornolti, P. Ferragina, Massimiliano Ciaramita, Stefan Rüd, and Hinrich Schütze. 2019. SMAPH: A Piggyback Approach for Entity-Linking in Web Queries. *ACM Trans. Inf. Syst.* 37 (2019), 13:1–13:42.

[6] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2021. Overview of the TREC 2020 deep learning track. In *Text REtrieval Conference (TREC)*. TREC.

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, Daniel Campos, and Ellen M. Voorhees. 2020. Overview of the TREC 2019 deep learning track. In *Text REtrieval Conference (TREC)*.

[8] Jeffrey Dalton, Laura Dietz, and James Allan. 2014. Entity query feature expansion using knowledge base links. *Proceedings of the 37th international ACM SIGIR conference on Research development in information retrieval* (2014).

[9] Jeffrey Dalton, Chenyan Xiong, and J. Callan. 2020. TREC CAsT 2019: The Conversational Assistance Track Overview. abs/2003.13624 (2020).

[10] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[11] Samuel Huston and W Bruce Croft. 2010. Evaluating verbose query processing techniques. In *Proceedings of the 33rd international ACM SIGIR conference on Research and development in information retrieval*. 291–298.

[12] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: a Benchmark for Question Answering Research. *Transactions of the Association of Computational Linguistics* (2019).

[13] Canjia Li, Andrew Yates, Sean MacAvaney, Ben He, and Yingfei Sun. 2020. PARADE: Passage representation aggregation for document reranking. *arXiv preprint arXiv:2008.09093* (2020).

[14] 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. In *CoCo@ NIPS*.

[15] Rodrigo Nogueira and Kyunghyun Cho. 2019. Passage Re-ranking with BERT. *arXiv:1901.04085* (2019).

[16] Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. 2020. Document Ranking with a Pretrained Sequence-to-Sequence Model. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*. 708–718.

[17] Natasha Noy, Yuqing Gao, A. Jain, A. Narayanan, A. Patterson, and J. Taylor. 2019. Industry-scale Knowledge Graphs: Lessons Challenges. *Queue* 17 (2019), 48–75.

[18] John Prager, Jennifer Chu-Carroll, and Krzysztof Czuba. 2002. Statistical Answer-Type Identification in Open-Domain Question Answering. In *Proceedings of the Second International Conference on Human Language Technology Research (HLT '02)*.

[19] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer. *Journal of Machine Learning Research* 21 (2020), 1–67.

[20] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas, 2383–2392.

[21] Johannes M. van Hulst, F. Hasibi, K. Dercksen, K. Balog, and A. D. Vries. 2020. REL: An Entity Linker Standing on the Shoulders of Giants. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (2020).

[22] Ellen M. Voorhees. 2001. Overview of the TREC 2001 Question Answering Track. In *Proceedings of the Tenth Text REtrieval Conference (TREC 2001)*. Gaithersburg, Maryland, 42–51.

[23] Ledell Yu Wu, F. Petroni, Martin Josifoski, Sebastian Riedel, and Luke Zettlemoyer. 2020. Zero-shot Entity Linking with Dense Entity Retrieval. In *EMNLP*.

[24] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained Transformers for Text Ranking: BERT and Beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1154–1156.