Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering

Sebastian Hofstätter, Markus Zlabinger, Mete Sertkan, Michael Schröder and Allan Hanbury
first.last@tuwien.ac.at
TU Wien, Vienna, Austria

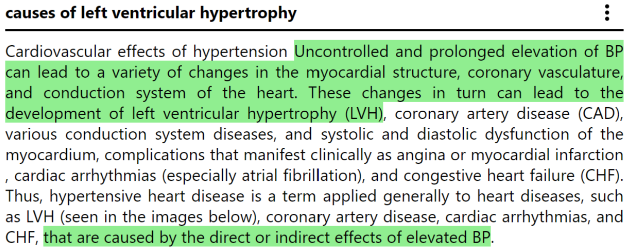
ABSTRACT

There are many existing retrieval and question answering datasets. However, most of them either focus on ranked list evaluation or single-candidate question answering. This divide makes it challenging to properly evaluate approaches concerned with ranking documents and providing snippets or answers for a given query. In this work, we present FiRA: a novel dataset of Fine-Grained Relevance Annotations. We extend the ranked retrieval annotations of the Deep Learning track of TREC 2019 with passage and word level graded relevance annotations for all relevant documents. We use our newly created data to study the distribution of relevance in long documents, as well as the attention of annotators to specific positions of the text. As an example, we evaluate the recently introduced TKL document ranking model. We find that although TKL exhibits state-of-the-art retrieval results for long documents, it misses many relevant passages.

ACM Reference Format:
Sebastian Hofstätter, Markus Zlabinger, Mete Sertkan, Michael Schröder and Allan Hanbury. 2020. Fine-Grained Relevance Annotations for Multi-Task Document Ranking and Question Answering. In *Proceedings of the 29th ACM International Conference on Information and Knowledge Management (CIKM '20), October 19–23, 2020, Virtual Event, Ireland.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3340531.3412878

1 INTRODUCTION

Currently, there is a substantial disconnect between the way neural retrieval models operate on full text and the relevance labels that guide learning, evaluation and analysis. Those labels, such as in the TREC Deep Learning Track 2019 (TREC-DL) [7], cover a long document, with thousands of tokens, as a whole. For traditional models, which collapse document statistics to a bag of words, this label granularity is sufficient. However, recently introduced neural ranking models operate on a more fine-grained level. They hierarchically build up their score from interactions of all words to a single score for the full document [1, 12, 34]. Therefore, the task of neural retrieval models shifts towards a combination of document-level relevance and word-level question answering or information extraction. The models are simultaneously tasked to detect relevant spans of information and, based on these, make a scoring decision that can be used to rank many candidate documents. However, a considerable bottleneck in their development and evaluation is the lack of datasets covering both sub-tasks equally well. Current datasets either focus on retrieval results with dense judgements across ranked lists [7, 11], or question answer selections in single candidate texts that can retroactively be converted to retrieval collections, but lead to incomplete retrieval judgements [27].



causes of left ventricular hypertrophy ⋮

Cardiovascular effects of hypertension Uncontrolled and prolonged elevation of BP can lead to a variety of changes in the myocardial structure, coronary vasculature, and conduction system of the heart. These changes in turn can lead to the development of left ventricular hypertrophy (LVH), coronary artery disease (CAD), various conduction system diseases, and systolic and diastolic dysfunction of the myocardium, complications that manifest clinically as angina or myocardial infarction , cardiac arrhythmias (especially atrial fibrillation), and congestive heart failure (CHF). Thus, hypertensive heart disease is a term applied generally to heart diseases, such as LVH (seen in the images below), coronary artery disease, cardiac arrhythmias, and CHF, that are caused by the direct or indirect effects of elevated BP.

Figure 1: This screenshot of our FiRA tool shows the task of selecting fine-grained relevance spans (or answers) within a relevant document snippet.

In this work, we present FiRA: a new dataset of **Fi**ne-grained **R**elevance **A**nnotations with word-level relevance selections and passage-level graded relevance labels for all relevant documents of the TREC 2019 Deep Learning Document dataset. We split the documents into snippets and displayed query & document snippet pairs to annotators. In Figure 1, we show an example of an annotated document-snippet. We ensure a high quality by employing at least 3-way majority voting for every candidate and continuous monitoring of quality parameters during our annotation campaign, such as the time spent per annotation. Furthermore, by requiring annotators to select relevant text spans, we reduce the probability of false-positive retrieval relevance labels. We conducted our annotation campaign with 87 computer science students using a custom annotation tool, which provides a seamless workflow between selecting relevance labels and relevant text spans.

The FiRA dataset contains 24,199 query & document-snippet pairs of all 1,990 relevant documents for the 43 queries of TREC-DL. We chose to extend the TREC-DL document ranking test collection and use the ranked retrieval data as a starting point, because 1) it allows us to narrow our annotation task, as fine-grained annotations are very time-consuming and 2) we are able to compare our non-expert annotators to TREC's expert annotators. FiRA is not specific to any model, but incorporates the full pool of TREC annotations.

Our novel FiRA dataset enables a variety of scenarios:

- **Training:** FiRA can be used for standalone and multi-task retrieval & QA fine-tuning. The density and coverage over the query-document pairs allow for granular fine-tuning using word-level loss functions.

CIKM '20, October 19–23, 2020, Virtual Event, Ireland

ACM ISBN 978-1-4503-6859-9/20/10…$15.00
https://doi.org/10.1145/3340531.3412878

- **Evaluation:** FiRA augments the TREC-DL evaluation, with more fine-grained relevance labels and the ability to evaluate the answer passages extracted from the full documents. Combined with the labeled non-relevant documents of TREC-DL, FiRA allows to evaluate rankings and answer selections.
- **Analysis:** Neural document ranking models that output relevance scores for a full document can be thoroughly analyzed. Especially, their partial results and attention regions can be quantitatively inspected with FiRA.

The human subjectivity of the relevance task is a recurring topic in IR research [4]. When using human-annotated datasets, it is important to know about the noise and uncertainty that different human relevance judgements bring. For this, we employ 3-way majority voting throughout our data collection. Additionally, we deliver our FiRA dataset with a quantitative analysis of the question:

RQ1 How much subjectivity do our fine-grained relevance annotations exhibit?

To that end, we selected 10 pairs to be annotated by all our annotators, so that we can study their subjectivity with a dense distribution that reduces outliers. We found, that on a 2-level-graded relevance scale our annotators agree strongly, whereas on the 4-level-graded relevance and the text selections the subjectivity increases.

We utilize our new dataset and present a thorough study of the positional bias in both relevance across a document and inside a single snippet. By providing annotators snippets in random order, so that there is no bias of attention towards the beginning, we aim to answer our research question:

RQ2 Does there exist a position-based relevance bias in long documents?

Our FiRA annotations show that there is indeed an increased likelihood of encountering relevant answers at the beginning of documents, although we also observe that there is a considerable number of relevant snippets found in later parts of the documents.

Furthermore, we studied the behavior of our annotators concerning the selection of relevant words inside a single snippet. An analysis of the MSMARCO-QA [3] dataset showed a strong bias towards the beginning of snippets. Therefore, using our FiRA annotations, we answer the question:

RQ3 Are annotators biased towards the beginning of a snippet?

For this we created a control-group and a treatment-group, in which we rotated the first and second halves of the texts. Neither group shows a position bias towards the start. The rotation-group actually amplifies the importance of context around a relevant span selection, as split sentences were less likely to be selected.

To showcase the usefulness of FiRA, we use the fine-grained judgements to analyze how well the recently introduced state-of-the-art TKL neural document re-ranking model [12] reaches its potential by utilizing TKL's interpretability functions. The TKL model provides information about which three regions with 30 words of a document it used compute the score.

RQ4 How many of FiRA's relevant spans does the TKL model use to compute the document score?

We find that the TKL model trained on TREC-DL data [7] has a disproportionate focus on the beginning of documents as a ranking signal.

The observations made using FiRA can inspire the next generation of neural document ranking models. Our FiRA dataset, including raw annotations, documentation and all accompanying helper scripts, is freely available at:
https://github.com/sebastian-hofstaetter/fira-trec-19-dataset

2 METHODOLOGY

We describe how we transformed the TREC-DL dataset to prepare it for fine-grained annotations; our annotation task description; followed by timeline analysis of the FiRA annotation exercise that we conducted among computer science students.

2.1 Data Preparation

Our dataset is based on the document ranking task of TREC's Deep Learning track 2019. This guarantees a broad and diverse pool of documents in our annotation campaign. TREC used the following graded relevance labels: *Not Relevant* (0), *Relevant* (1), *Highly Relevant* (2), and *Perfect* (3). For the 43 queries of the DL task we gathered all documents marked as *Highly Relevant* (1149 query-document pairs) and *Perfect* (841 query-document pairs) by the TREC experts. We concatenated the title and body text, and did not visually indicate the title to the annotators.

The documents are long with an average 3,039 words and a median of 1,171 words. We assumed that if we want exact and exhaustive word-level relevance annotations, we need to divide the documents into smaller pieces, as this has been shown to be an effective task simplification [10]. We therefore split the documents into snippets. We aimed to create semantically coherent snippets of complete sentences of approximately 120–130 words,[1] as this is the maximum number of words that fit on a typical smartphone screen without the need to scroll. We used the following rules to create our snippets:

- Split a document based on sentences using the BlingFire library.[2]
- Add sentences to a snippet until the maximum length of 130 words is reached. If it overflows, we start with a new snippet starting with this sentence.
- If a sentence does not fit (e.g. if a document does not contain any punctuation), we split by whitespace, so we could concatenate tokens back together, to fit the desired snippet length.

We created a maximum of 30 document snippets per document, resulting in a cap of roughly 4,000 tokens, as this was the maximum capacity of our annotation resources. We opted to prioritize complete majority voting coverage over more depth.

This procedure resulted in a total of 24,199 document snippets, with an average of 12 snippets per query-document pair.

2.2 FiRA Annotation Task Guidelines

The guiding principle of this annotation campaign was to divide the task into the smallest possible pieces and allow for desktop as well as mobile annotations, to allow annotation to take place in the widest possible types of settings. Our aim was to create a

[1] We split by whitespace, other tokenization methods may result in more tokens.
[2] https://github.com/microsoft/BlingFire

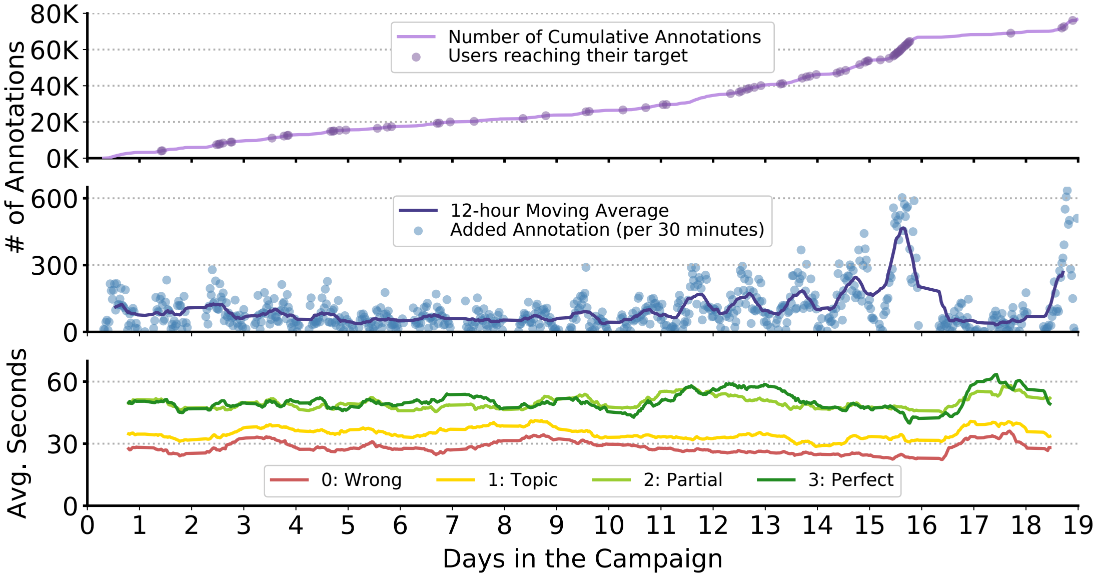


Figure 2: Monitoring of our annotation campaign. The topmost plot shows the total annotation count; the middle plot the added annotations in a 30 minute window; and the bottom plot shows the daily-average time spent per annotation class. Note: The first deadline was set to the end of day 16 and later extended by three days.

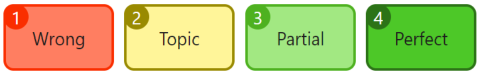


Figure 3: Screenshot of the label-selection UI of the FiRA annotation tool. The numbers indicate keyboard shortcuts.

simple and reduced user interface, which included condensing the naming schema of the relevance classes to fit on a single line on a smartphone. We kept the graded relevance from TREC, however we re-named the relevance labels and defined them as follows:

Wrong The document has nothing to do with the query, and does not help in any way to answer it.

Topic The document talks about the general area or topic of a query, might provide some background info, but ultimately does not answer it.

Partial The document contains a partial answer, but you think that there should be more to it.

Perfect The document contains a full answer: easy to understand and it directly answers the question in full.

To make it easier for our non-expert annotators to distinguish the classes, we displayed them with strong visual cues as background colors of the voting buttons, as shown in Figure 3. Our annotation guidelines stated: "For Partial and Perfect grades, you need to select the text spans that are in fact the relevant text parts to the questions." We purposefully kept the definition of "relevant" ambiguous to let the annotators decide for themselves what they deem relevant and to receive realistic relevance labels for non-expert web search engine users. Furthermore, we emphasised in our guidelines that if an answer requires domain specific knowledge and the connection to the query is not clear in the snippet, then that snippet is not relevant. Finally, we concluded our annotation guidelines with one example per relevance class.

2.3 Annotation Campaign

We conducted our annotation campaign among 87 computer science students. We set a target of 500 annotations per student and incentivized our students with bonus points if they continued after reaching their target. On average, each student annotated 890 samples.[3]

Previous work highlighted the difficulty of working with student annotators [26], therefore we took a number of steps to monitor the quality of the annotations: **1)** Besides a 3-way majority voting for every pair, we also let every student annotate the same set of 10 pairs spread throughout their workload; **2)** We monitored the time per annotation, to detect "cheating" by students who just immediately select any class; **3)** We asked the students for feedback twice (after the first 20 annotations and after they completed their target).

Our campaign occurred over the course of 19 days as an exercise with a clear deadline for our students. It is generally believed that academics and students alike are especially motivated closer to a deadline. To better understand this phenomenon and to provide quality control, we plot the cumulative and running annotation trajectories, as well as the daily-average time per label over the

[3] We have to note that the annotation campaign was conducted during a COVID-19 lockdown, which might have helped us gather more annotations than we would have in a normal situation.

Table 1: FiRA dataset statistics. The # of labels is after majority voting.

# of queries	43
# of documents in collection	3,213,835
# of FiRA annotated query-document pairs	1,990
+ TREC-DL # of query-document pairs	16,258
# of judged document snippets	24,197
# of total judgements	78,340
# of 0: Wrong	13,565 (56 %)
# of 1: Topic	6,201 (26 %)
# of 2: Partial	3,145 (13 %)
# of 3: Perfect	1,286 (5 %)
# of annotators	87
avg. # per annotator	890

course of our campaign in Figure 2. In the top plot, we see that some students finished their target earlier, however, most students finished close to the first deadline. In the second plot, we can clearly observe the deadline motivation phenomenon, with increased activity starting 5 days before the deadline. Even though we observe this increased activity, the average time per annotation only decreases gradually by a few seconds. The *Partial* and *Perfect* times are naturally higher than the other two classes, as the annotators were required to select text in addition to choosing a label.

Overall our students gave us very positive feedback for the question: *How did you like to work with FiRA so far?* 38% said *Very good*, 45% *Good*, 12% *Decent*, and only 5 % selected *Don't like* as the rating feedback.[4] We see this as a positive confirmation of our task design. Many students also gave written feedback, which helps us tremendously to understand the limitations of our dataset. The main problem among annotators was the distinction between the *Wrong* and *Topic* labels. We believe that this problem arose from the fact that we only annotate documents that have at least somewhere in them a relevant part, making it more likely that the unrelated parts stay on topic. Therefore, when using the annotations, we should assume both classes to be non-relevant. Interestingly, if we consider the average time per annotation from Figure 2 we observe that the *Topic* label took a few second longer to classify than *Wrong*, which leads us to believe that the two classes, while challenging and noisy to distinguish in single examples, make sense overall.

2.4 FiRA Dataset Processing

After the annotation campaign we post-processed the raw annotations to form the final published FiRA dataset. In Table 1, we give an overview of the number of judged pairs, total annotations, and label distribution of the FiRA dataset. The combination of TREC-DL and FiRA creates a very densely annotated retrieval and QA-snippet dataset, which can be used by any approach investigating fine-grained relevance aspects of retrieval and QA.

We transform the raw annotations into standard-format qrels for retrieval, as well as a specialized but simple format for indicating character-level selection of relevant regions. We purposefully stay at a character-level to be independent of tokenization methods, however we deliver simple helper scripts to easily map our character ranges to tokens of any tokenization method.

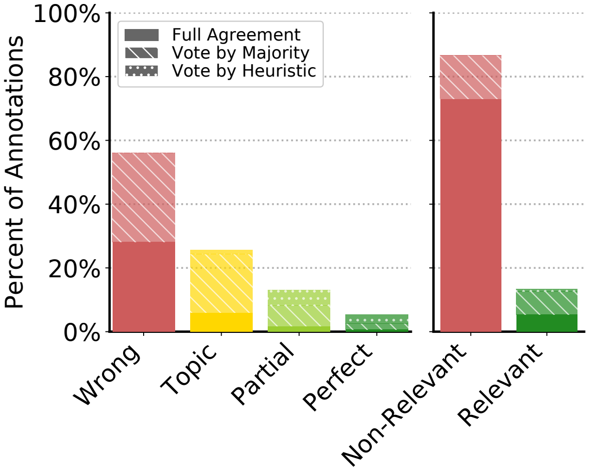


Figure 4: FiRA's distribution of labels, and ratio of heuristic & majority voting.

To distill our raw annotations to voted judgements per query-document snippet, we relied on the following procedure: If a full agreement between all annotators or a majority for a single class exists, we take the majority class. If there exists a split between two or more labels, we employ the heuristic to take the highest order of relevance class.[5] We apply this heuristic with the assumption that if an annotator took a higher class, they spent more time on the decision on average (as seen in Figure 2) and the two relevant classes require a selection of text, which reduces the risk of false positives.

In Figure 4, we show the distribution of classes in the dataset as well as the distribution of the judgement voting type per class. We rarely see a full agreement of all judges. This shows that our approach of prioritizing 3-way majority voting coverage over more pairs to annotate was warranted.

We have to emphasize that we only annotated documents judged to be overall relevant by TREC. We observe that even though some part of a document is relevant, most parts are not relevant to the query. We investigate this observation further in Section 4.

3 AGREEMENT & SUBJECTIVITY

When we treat retrieval collections solely as a given universal ground truth to measure retrieval metrics, we are prone to overlook the inherent subjectivity of their assignment and the annotators creating the dataset. As part of FiRA, we selected 10 pairs to be graded by all annotators, evenly distributed throughout the task. We selected the 10 query-document snippet pairs to contain (in our opinion) 3 *Perfect*, 3 *Partial*, 2 *Topic* and 2 *Wrong* instances. We conducted this experiment to obtain evidence towards answering RQ1: *How much subjectivity do our fine-grained relevance annotations exhibit?*

[4] The feedback was not anonymous, so that we could follow up on the written feedback.

[5] We take the descending order of: Perfect, Partial, Topic, Wrong.

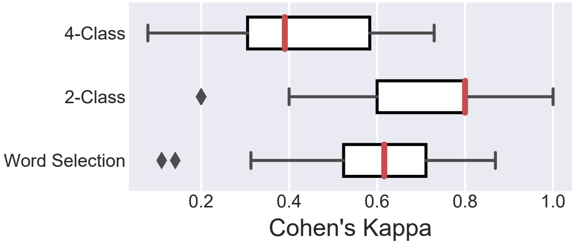


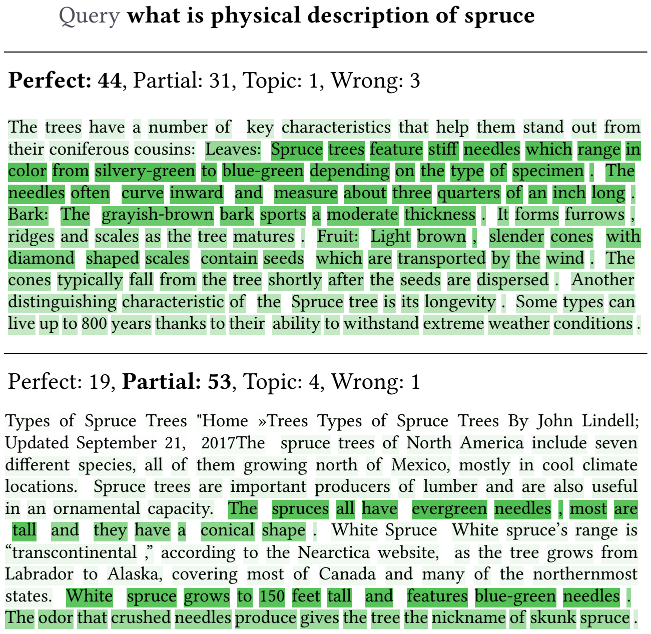
Figure 5: Cohen's Kappa agreement between individual students and the aggregated annotations via majority voting.

We analyze the subjectivity based on the inter-annotator agreement (IAA) between the student annotators. We compute the agreement via Cohen's Kappa, a standard metric to compare two sets of annotations. As data basis, we consider the 10 pairs that were annotated by all students. Based on these 10 pairs, we compute an aggregated annotation by majority vote; then, we compare each student's annotation with the aggregated one.

The Kappa scores for 2-classes,[6] 4-classes, and word-based relevance assignment are presented in Figure 5. For the 2-class relevance assignment and the labeling of the relevant word phrases, substantial Kappa agreements are reached, ranging from 0.5 to 0.8. For the more difficult task of differentiating between all four relevance classes, a mediocre agreement ranging from 0.3 to 0.6 is measured. Our results align with Alonso and Mizzaro [2] who report similar Kappa agreements for the relevance labeling of TREC collections when employing non-expert annotators.

The inter-annotator agreement is a metric for the quality in data annotation project. A low agreement indicates that the annotators failed to agree on common judgements. However, for the task of relevance feedback assignment for IR, also the subjectivity of individual workers is a source of disagreement [4]: For example, one annotator might find a given text phrase relevant whereas a second annotator does not find that the exact same text phrase sufficiently answers the query. Therefore, for relevance judgement tasks, a certain level of disagreement can be expected since the task subjectivity needs to be taken into account as well.

As a case based example for subjectivity, we present in Figure 6 the distribution of fine-grained word-level annotations on two document snippets for the same query. We would classify the snippet on top as *Perfect* and the lower as *Partial.* The majority of our annotators also selected the respective labels, although we see a strong overlap between the two relevant classes. We can observe that in some instances, such as the second snippet, it is very easy to select the relevant regions, and most annotators agree on two sentences, whereas in the first snippet we see a greater variability with the second and third sentence at its core. In our opinion, relevance is context dependent, especially in the second example of Figure 6, as the selected answers are correct and fulfill the information need: if you seek a short and concise answer it would be *Perfect*, however if someone seeks an exhaustive description of spruce trees it would only be *Partial.*



Query **what is physical description of spruce**

Perfect: 44, Partial: 31, Topic: 1, Wrong: 3

The trees have a number of key characteristics that help them stand out from their coniferous cousins: Leaves: Spruce trees feature stiff needles which range in color from silvery-green to blue-green depending on the type of specimen . The needles often curve inward and measure about three quarters of an inch long . Bark: The grayish-brown bark sports a moderate thickness . It forms furrows , ridges and scales as the tree matures . Fruit: Light brown , slender cones with diamond shaped scales contain seeds which are transported by the wind . The cones typically fall from the tree shortly after the seeds are dispersed . Another distinguishing characteristic of the Spruce tree is its longevity . Some types can live up to 800 years thanks to their ability to withstand extreme weather conditions .

Perfect: 19, **Partial: 53**, Topic: 4, Wrong: 1

Types of Spruce Trees "Home »Trees Types of Spruce Trees By John Lindell; Updated September 21, 2017The spruce trees of North America include seven different species, all of them growing north of Mexico, mostly in cool climate locations. Spruce trees are important producers of lumber and are also useful in an ornamental capacity. The spruces all have evergreen needles , most are tall and they have a conical shape . White Spruce White spruce's range is "transcontinental ," according to the Nearctica website, as the tree grows from Labrador to Alaska, covering most of Canada and many of the northernmost states. White spruce grows to 150 feet tall and features blue-green needles . The odor that crushed needles produce gives the tree the nickname of skunk spruce .

Figure 6: Comparison of two completely judged document snippets. The background heatmap for each word displays the number of times this word was part of a relevant region selection. The darker the green the more often annotators selected this word as being relevant.

4 POSITION BIAS STUDY

In this study, we look at two different position bias scenarios. First, a bias of snippet relevance based on the location in the document, which is determined by a combination of how humans write documents, and the broad and general nature of the questions asked in the TREC-DL dataset. Our second focus is studying whether annotators pay insufficient attention and prematurely decide on the relevance of a snippet after reading only its beginning. To this end, we conducted an experiment during the annotation process.

4.1 In-Document Position Importance

Previous studies repeatedly highlighted the greater importance to an overall document relevance assessment of some passages, such as introductory paragraphs in news articles [5, 31] and first-k term models in TREC-DL 2019 [13]. In our study, we aimed to manually confirm this observation in a web search context and answer our RQ2: *Does there exist a position-based relevance bias in long documents?* To prevent an inherent bias against later positions, we showed our annotators snippets of documents in random order. With this approach, we are confident that our annotation results reflect true relevance differences and not user-interface based annotator bias.

Our results are detailed in Figure 7, where we show the proportion of relevant annotations per snippet location for documents split by TREC classes 2 (Highly Relevant) and 3 (Perfect). Our main observation here is a confirmation of these earlier studies that highlight the importance of earlier passages in a document. However, we also see that this distribution of relevance across longer

[6] We combine *Wrong* and *Topic*, as well as *Partial* and *Perfect*

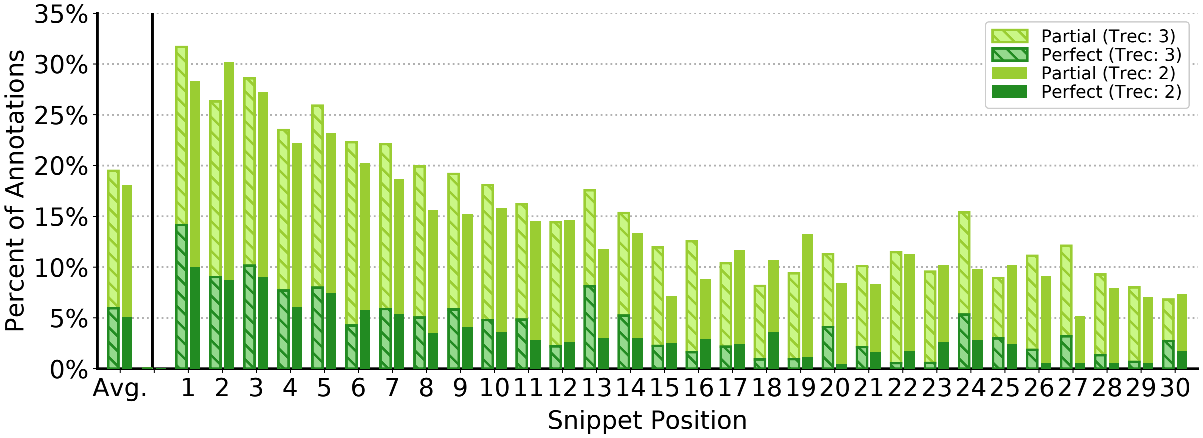


Figure 7: Relative amount of relevant annotations per position of the snippets in the document, for both TREC labels.

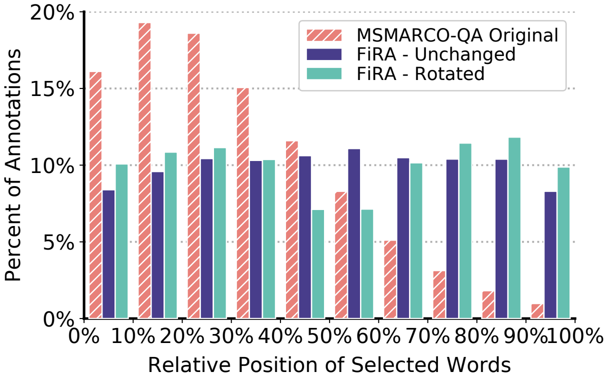


Figure 8: Relative distribution of selected words in passages of MSMARCO-QA and FiRA.

documents is not very strongly skewed and a sizeable amount of relevant information can be found in later parts. This shows that for high-recall targets, also inside documents, it is vital to operate on the full documents.

Our next observation based on Figure 7 is that our fine-grained annotations only show a small difference for the TREC classes 2 (Highly Relevant) and 3 (Perfect) in terms of the likelihood of *Perfect* and *Partial* relevant snippets. The biggest difference occurs in the first snippet position, where TREC-3 has 4 % more *Perfect* annotations. This result can be the output of different annotation guidelines and setups, as we broke down and simplified the task, or just shows — as discussed in Section 3 — that we have to accept noise in relevance annotations due to human subjectivity. This result suggests dampening the gain values of relevance labels used in nDCG to be less different than 2 and 3, as we cannot observe a strong difference between the two relevant labels.

4.2 In-Snippet Annotator Bias

Before starting our annotation campaign, we discovered an imbalance in the distribution of selected answer words[7] in the MSMARCO-QA dataset, which is the initial source of queries and websites later used in TREC-DL and FiRA.[8] This imbalance is visible in Figure 8, where the MSMARCO-QA answers clearly favor earlier words in the passages. Our assumption is that this strong bias favoring earlier positions comes from annotators prematurely deciding that a passage is not relevant if the answer is not contained in the beginning. A major difference between MSMARCO-QA and FiRA is the annotation workflow: MSMARCO displayed a ranked list of results, whereas we displayed one document snippet at a time.

To study RQ3 (*Are annotators biased towards the beginning of a snippet?*) we rotated snippets half of the time before showing them to annotators. The rotation was conducted by cutting a snippet into two halves by the word count and concatenating the two halves in reverse order. Then, after a judgement, we mapped the selection back to its original position in the document. Our results for the distribution of selected relevant words are shown in Figure 8.

Using our FiRA approach, in both the unchanged and rotated case we do not observe an imbalance of selected relevant words towards the beginning. In the unchanged passages, we see that the first and last 10 % (roughly the first and last sentence) received fewer annotations. In the rotated case, we observe this dip in selected passages around the cutting point. Those words moved to the beginning and end of the user interface.

Overall the average percentage of words in a snippet selected by an annotator is 31 % with a standard deviation of 24 % and a mean number of 43 selected words in total. This shows that our even position distribution does not come from annotators selecting every word.

Our annotators were not biased towards the beginning of passages, and in fact rotating the passages is not needed. We disproved our hypothesis, however we made another interesting discovery worth our attention in future work: Our annotators, using the FiRA tool, were less likely to select words from the first and last sentence of a snippet. We hypothesise that this is due to missing context information, cut off by the snippet creation. For future annotation campaigns, we plan to incorporate context sentences, connecting the current passage to the preceding and following ones, to further improve the uniform balance of selected answers.

[7] MSMARCO-QA does not contain exact answer spans, therefore we matched answer words to words in their respective training passages (matched: 32 % or 176,013 answers).

[8] The query passage pairs differ between MSMARCO-QA and FiRA, but the query distribution and domain remain the same

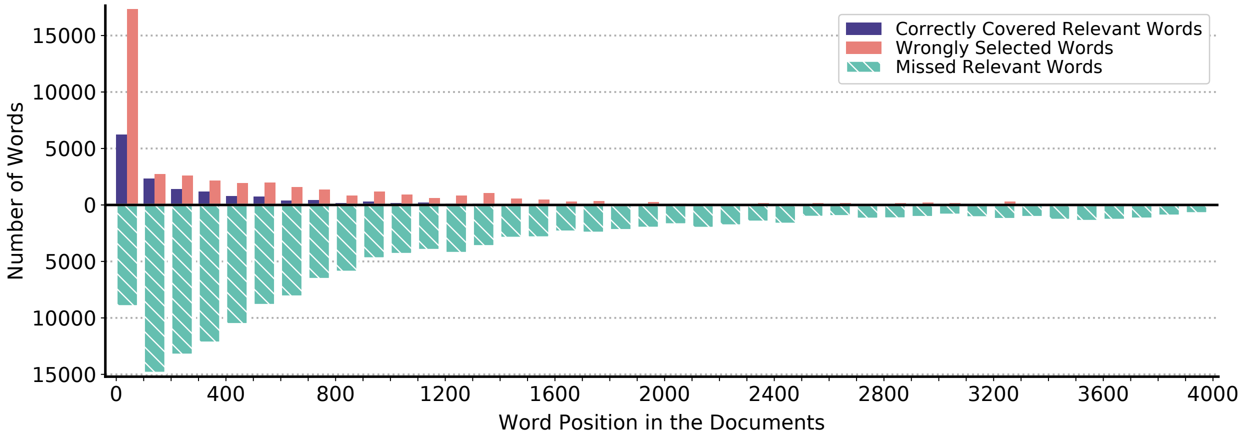


Figure 9: Analysis of TKL's relevant positions of retrieved relevant documents using FiRA. On top are the correctly (true-positive) and wrongly (false-positive) selected spans, whereas on the bottom is the "iceberg" of missed relevant spans selected in FiRA, but missed by TKL.

5 NEURAL DOCUMENT RANKING

To showcase the usefulness of our FiRA dataset, we now turn to the analysis of neural re-ranking models. We focus in particular on the recently introduced TKL document ranking model [12], as it offers solid interpretation capabilities, namely it outputs detected relevant regions inside a long document.

The TKL model scores a query-document pair by first contextualizing the two sequences independently with Transformers, where the document sequence is split into overlapping chunks to accommodate long input sequences. The TKL model then creates a cosine match-matrix between every query and document term representation. This match matrix is scanned with a sliding saturation function, to detect the most relevant regions of a document. In addition to its ranking score, TKL outputs the position of the three top regions of a document, which were used as a basis to calculate the score. Now with FiRA, we can analyze, on a word-level, if the model focused on relevant regions or if the model learned artefacts that differ from human judgement. Hence we can answer our RQ4: *How many of FiRA's relevant spans does the TKL model use to compute the document score?*

In Figure 9, we analyze the word-level relevance prediction of the TKL model on the FiRA dataset. We plot the number of words that have been correctly marked as relevant, the number of words wrongly selected as relevant by TKL, and the number of missed relevant words that have not been captured by TKL. With FiRA data we can now visualize the "iceberg" of TKL's missed relevant words. The TKL model learns a strong bias towards the beginning of a document, as a ranking signal, although FiRA shows that much relevant information is available later in the documents, where TKL rarely selects a region.

6 RELATED WORK

Our work relates to datasets for question answering and retrieval as well as the analysis of document structure and passage importance in IR. In addition, we also highlight studies that could benefit from the usage of FiRA.

Datasets. The TREC document task [7] builds on re-crawled data from the MSMARCO-QA collection [3]. The MSMARCO-QA collection differs from our work in that only single passages have been binary-annotated and the answers are natural language text written by the annotators.

The 2007 edition of WebCLEF featured a snippet selection evaluation, where annotators where asked to judge returned snippets from the participating systems [18]. However, due to the low participation count, and non-exhaustive evaluation the collection has been shown to not be reusable for future systems [25]. In contrast, we conducted an exhaustive and model agnostic annotation campaign.

In the landscape of Question Answering datasets, the SQuAD [27] QA dataset, was annotated by finding questions to fit an answer present in a given Wikipedia passage. The SearchQA [9] dataset expanded snippets with Google search result snippets. For larger context QA and advanced reasoning, Clark et al. [6] published ARC and Mihaylov et al. [21] published the OpenBookQA. Kwiatkowski et al. [20] used real Google queries and annotated answers selected from Wikipedia articles in the Natural Questions dataset. Those datasets are focused on different aspects of QA, without necessary relevance annotations for a broader set of documents. Therefore, the transformation to IR collections results in very sparse relevance labels. FiRA tries to bridge this gap by extending TREC-DL ranking annotations with selected snippets.

Fine-grained annotations of long documents have already been the focus of other NLP domains such information extraction [16] and relation extraction [33]. Fine-grained annotations are valuable in specialised domains, such as medical tagging [28] or detecting proposition types in argumentation [19].

Relevance Analysis. Our work studies relevance distributions across a document using our new annotations as well as the subjectivity of the relevance task. To better understand the inherent meaning of judging relevance, Inel et al. [15] studied topical relevance and various settings of the ranking task, including different document granularities. Wu et al. [31] studied the influence of passage-level relevance on full-document relevance for Chinese news articles. They found the first passage to be a very strong indicator of overall relevance. Hofstätter et al. [14] created the Neural IR-Explorer to help us better understand single word-interactions in the score aggregation of neural re-ranking models.

Neural Models. We identified numerous works on neural models published in recent years that would benefit from an evaluation and analysis based on FiRA. A common approach to utilizing the large-scale pre-trained BERT model [8] in document ranking is to apply BERT to passages [30], overlapping windows [32], or single sentences [1]. In all these cases, BERT produces partial results that need to be aggregated externally to produce a final ranking score that could be compared with traditional full-document judgements. With FiRA, these models could be evaluated on their respective granularity before aggregating scores. Apart from BERT-based approaches, Jiang et al. [17] proposed a semantic text matching for long-form documents model, and Zheng et al. [34] created RLTM for long document ranking. With modifications, both could also be analyzed with FiRA.

In addition to retrieval focused models, the research community also proposed multi-task models for the QA-at-scale or sometimes referred to open-world-QA task, where systems need to retrieve candidates and select answers. Models such as the Retriever-Reader model from Ni et al. [22], the multi-task Retrieve-and-Read models [23, 24], or models based on phrase indices [29], are mainly evaluated on re-purposed QA-collections that lack dense retrieval judgements. FiRA presents a perfect fit for the evaluation of such models, as it can cover both retrieval and question answering aspects equally well.

7 CONCLUSION

With the release of FiRA, we provide the research community with a high-quality dataset that can be employed in a range of diverse settings. We showcased the usage of FiRA with a position bias and distribution study, as well as fine-grained analysis of a neural document ranking model. We look forward to seeing novel approaches by the research community, from both retrieval and question answering fields, to evaluate and analyse their models with FiRA.

Acknowledgements. We want to thank Patrick Kerschbaum for implementing our custom annotation tool as well as our students of the Advanced Information Retrieval course in the summer term of 2020 for annotating the data and being so patient and motivated in the process.

REFERENCES

[1] Zeynep Akkalyoncu Yilmaz, Wei Yang, Haotian Zhang, and Jimmy Lin. 2019. Cross-Domain Modeling of Sentence-Level Evidence for Document Retrieval. In *Proc. of EMNLP.*

[2] Omar Alonso and Stefano Mizzaro. 2012. Using crowdsourcing for TREC relevance assessment. *Information Processing & Management* 48, 6 (2012), 1053–1066.

[3] Payal Bajaj, Daniel Campos, Nick Craswell, et al. 2016. MS MARCO : A Human Generated MAchine Reading COmprehension Dataset. In *Proc. of NIPS.*

[4] Pia Borlund. 2003. The concept of relevance in IR. *Journal of the American Society for IST* (2003).

[5] Matteo Catena, Ophir Frieder, Cristina Ioana Muntean, Franco Maria Nardini, Raffaele Perego, and Nicola Tonellotto. 2019. Enhanced News Retrieval: Passages Lead the Way!. In *Proc. of SIGIR.*

[6] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457* (2018).

[7] Nick Craswell, Bhaskar Mitra, Emine Yilmaz, and Daniel Campos. 2019. Overview of the TREC 2019 deep learning track. In *In Proc. of TREC.*

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).

[9] Matthew Dunn, Levent Sagun, Mike Higgins, V Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. Searchqa: A new q&a dataset augmented with context from a search engine. *arXiv preprint arXiv:1704.05179* (2017).

[10] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. 2013. Keep it simple: Reward and task design in crowdsourcing. In *Proc. of CHItaly.*

[11] Helia Hashemi, Mohammad Aliannejadi, Hamed Zamani, and W Bruce Croft. 2020. ANTIQUE: A non-factoid question answering benchmark. In *Proc. of ECIR.*

[12] Sebastian Hofstätter, Hamed Zamani, Bhaskar Mitra, Nick Craswell, and Allan Hanbury. 2020. Local Self-Attention over Long Text for Efficient Document Retrieval. In *Proc. of SIGIR.*

[13] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2019. TU Wien @ TREC Deep Learning '19 – Simple Contextualization for Re-ranking. In *Proc. of TREC.*

[14] Sebastian Hofstätter, Markus Zlabinger, and Allan Hanbury. 2020. Neural-IR-Explorer: A Content-Focused Tool to Explore Neural Re-Ranking Results. In *Proc. of ECIR.*

[15] Oana Inel, Giannis Haralabopoulos, Dan Li, Christophe Van Gysel, Zoltán Szlávik, Elena Simperl, Evangelos Kanoulas, and Lora Aroyo. 2018. Studying topical relevance with evidence-based crowdsourcing. In *Proc. of CIKM.*

[16] Sarthak Jain, Madeleine van Zuylen, Hannaneh Hajishirzi, and Iz Beltagy. 2020. SciREX: A Challenge Dataset for Document-Level Information Extraction. arXiv:cs.CL/2005.00512

[17] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic text matching for long-form documents. In *Proc. of WWW.*

[18] Valentin Jijkoun and Maarten de Rijke. 2008. Overview of WebCLEF 2007, Carol Peters, Valentin Jijkoun, Thomas Mandl, Henning Müller, Douglas W. Oard, Anselmo Peñas, Vivien Petras, and Diana Santos (Eds.).

[19] Yohan Jo, Elijah Mayfield, Chris Reed, and Eduard Hovy. 2020. Machine-Aided Annotation for Fine-Grained Proposition Types in Argumentation.

[20] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, et al. 2019. Natural Questions: a Benchmark for Question Answering Research. *Trans. of the ACL* (2019).

[21] Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proc. of EMNLP.*

[22] Jianmo Ni, Chenguang Zhu, Weizhu Chen, and Julian McAuley. 2018. Learning to attend on essential terms: An enhanced retriever-reader model for open-domain question answering. *arXiv preprint arXiv:1808.09492* (2018).

[23] Kosuke Nishida, Kyosuke Nishida, Masaaki Nagata, Atsushi Otsuka, Itsumi Saito, Hisako Asano, and Junji Tomita. 2019. Answering while summarizing: Multi-task learning for multi-hop qa with evidence extraction. *arXiv preprint arXiv:1905.08511* (2019).

[24] Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. 2018. Retrieve-and-read: Multi-task learning of information retrieval and reading comprehension. In *Proc. of CIKM.*

[25] Arnold Overwijk, Dong Nguyen, Claudia Hauff, Rudolf Berend Trieschnigg, Djoerd Hiemstra, and Franciska MG de Jong. 2008. On the Evaluation of Snippet Selection for Information Retrieval.. In *Proc. of CLEF.*

[26] Joao Palotti, Guido Zuccon, Johannes Bernhardt, Allan Hanbury, and Lorraine Goeuriot. 2016. Assessors agreement: A case study across assessor type, payment levels, query variations and relevance dimensions. In *Proc. of CLEF.*

[27] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. of EMNLP.*

[28] Kirk Roberts, Sonya E Shooshan, Laritza Rodriguez, Swapna Abhyankar, Halil Kilicoglu, and Dina Demner-Fushman. 2015. The role of fine-grained annotations in supervised recognition of risk factors for heart disease from EHRs. *Journal of Bio.Inf.* (2015).

[29] Minjoon Seo, Jinhyuk Lee, Tom Kwiatkowski, Ankur P Parikh, Ali Farhadi, and Hannaneh Hajishirzi. 2019. Real-time open-domain question answering with dense-sparse phrase index. *arXiv preprint arXiv:1906.05807* (2019).

[30] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Jingtao Zhan, Yukun Zheng, Min Zhang, and Shaoping Ma. 2020. Leveraging Passage-level Cumulative Gain for Document Ranking. In *Proc. of WWW.*

[31] Zhijing Wu, Jiaxin Mao, Yiqun Liu, Min Zhang, and Shaoping Ma. 2019. Investigating Passage-level Relevance and Its Role in Document-level Relevance Judgment. In *Proc. of SIGIR.*

[32] Ming Yan, Chenliang Li, Chen Wu, Bin Bi, Wei Wang, Jiangnan Xia, and Luo Si. 2019. IDST at TREC 2019 Deep Learning Track. In *Proc. of TREC.*

[33] Yuan Yao, Deming Ye, Peng Li, Xu Han, et al. 2019. DocRED: A Large-Scale Document-Level Relation Extraction Dataset. In *Proc. of ACL.*

[34] Chen Zheng, Yu Sun, Shengxian Wan, and Dianhai Yu. 2019. RLTM: an efficient neural IR framework for long documents. *arXiv preprint arXiv:1906.09404* (2019).