Less is Less: When Are Snippets Insufficient for Human vs Machine Relevance Estimation?

Gabriella Kazai, Bhaskar Mitra, Anlei Dong, Nick Craswell, and Linjun Yang

Microsoft, One Microsoft Way, Redmond, WA, USA
{gkazai,bmitra,anldong,nickcr,linjya}@microsoft.com

Abstract. Traditional information retrieval (IR) ranking models process the full text of documents. Newer models based on Transformers, however, would incur a high computational cost when processing long texts, so typically use only snippets from the document instead. The model's input based on a document's URL, title, and snippet (UTS) is akin to the summaries that appear on a search engine results page (SERP) to help searchers decide which result to click. This raises questions about when such summaries are sufficient for relevance estimation by the ranking model or the human assessor, and whether humans and machines benefit from the document's full text in similar ways. To answer these questions, we study human and neural model based relevance assessments on 12k query-documents sampled from Bing's search logs. We compare changes in the relevance assessments when only the document summaries and when the full text is also exposed to assessors, studying a range of query and document properties, e.g., query type, snippet length. Our findings show that the full text is beneficial for humans and a BERT model for similar query and document types, e.g., tail, long queries. A closer look, however, reveals that humans and machines respond to the additional input in very different ways. Adding the full text can also hurt the ranker's performance, e.g., for navigational queries.

Keywords: Relevance Estimation · Crowdsourcing · Neural IR.

1 Introduction

In adhoc retrieval, ranking models typically process text from the URL, title and body of the documents. While the URL and title are short, the body may include thousands of terms. Recently, Transformer-based ranking models have demonstrated significant improvements in retrieval effectiveness (Lin et al., 2020), but are notoriously memory and compute intensive. Their training and inference cost grows prohibitively with long input. A common solution is to estimate document relevance based only on sub-parts of the document, e.g., query-biased snippets. Such approaches are motivated by the *scope hypothesis* (Robertson et al., 2009), which states that the relevance of a document can be inferred by considering only its most relevant parts. Several neural approaches, *e.g.*, Hofstätter et al. (2021); Yan et al. (2019), have operationalized this hypothesis in their model design. Document summaries based on URL, title and query-biased snippet (UTS) are also typically presented on SERPs to searchers. While the model uses UTS to estimate relevance when ranking, the human searcher uses UTS to estimate relevance when deciding whether to click a result. These scenarios motivate us to study when snippets are sufficient replacements of the full body text for relevance estimation by humans and machines. Concretely, by collecting human relevance assessments and relevance rankings from a machine-learned model both for UTS only and UTS plus body text inputs, we study whether humans and machines benefit from the document's full text under similar conditions and in similar ways or if humans and machines respond to the additional input differently.

2 Related work

Automatic document summarization dates as far back as the foundational work by Luhn (1958) and Edmundson (1964). In the context of search, several early user studies (Tombros and Sanderson, 1998; Sanderson, 1998; White et al., 2003) demonstrated the usefulness of query-biased snippets for assessing document relevance. Demeester et al. (2012, 2013) studied how well the document's relevance can be predicted based on the snippet alone in federated search. Unlike these prior works, our goal is to study the differences in human and machine relevance assessments when only document summaries or when also the body texts are inspected. Past studies have also employed diverse measures of snippet quality based on manual assessment (Kaisser et al., 2008), eye-tracking studies (Lagun and Agichtein, 2012; Cutrell and Guan, 2007), view-port analysis (Lagun and Agichtein, 2011), historical clickthrough data (Clarke et al., 2007; Yue et al., 2010), and A/B testing (Savenkov et al., 2011), but did not try to understand when and why human and model assessments differ.

The application of passage-based document views for adhoc document ranking have been explored in the context of traditional retrieval methods (Bendersky and Kurland, 2008; Salton et al., 1993), but gained more attention recently (Nogueira and Cho, 2019; Yan et al., 2020; Hofstätter et al., 2020, 2021; Li et al., 2020) in the context of Transformer-based (Vaswani et al., 2017) neural ranking models. While these models typically evaluate several passages per document, single query-biased summaries can be applied under stricter efficiency concerns. Our work helps to understand the feasibility of estimating document relevance based on just the UTS information.

Finally, our work is similar to Bolotova et al. (2020) in the sense that we too study humans and a BERT model, but while Bolotova et al. (2020) focused on attention, we study changes in relevance estimation due to input change.

3 Experiment design

To answer our research questions, we collect both human and neural model based relevance assessments in two conditions: 1) when the human/machine assessor is only shown the query-biased summary, made up of the URL, title and snippet (UTS), and 2) when the body text is also exposed (UTSB). We use snippets returned by Bing's API.

We collect relevance assessments from humans via a Human Intelligent Task (HIT) with multiple judging steps, ensuring that the same person labels both conditions. First, we ask assessors to estimate a search result's relevance to the query based on its UTS information alone (UTS label). We then show assessors the web page and ask them to re-assess its relevance (UTSB label). Both labels use a five point scale. Next, we ask if seeing the web page led to a revised assessment ('Revised'; this is auto-filled), if it helped to confirm the UTS based estimate ('Confirmed') or if the page did not provide further help in the assessment ('Not needed'). Finally, assessors are asked to highlight parts of the body text that explain why the body text provided additional benefit over the UTS. Figure 1 shows the final HIT state. We use UHRS, an internal crowdsourcing platform, to collect judgments from trusted, quality monitored, long-term judges and pay them their standard hourly rate. We obtain an inter-assessor agreement rate of 0.44 for the UTS and 0.53 for the UTSB labels (Krippendorff α).

For our Neural Ranker based relevance estimation, we follow the state-of-the-art neural ranking approach (Nogueira and Cho, 2019) and train a UTS and a UTSB ranker, starting with a pretrained BERT-style (Devlin et al., 2019) model. The model inputs comprise sentence A, which is the query, and sentence B, which is either UTS or

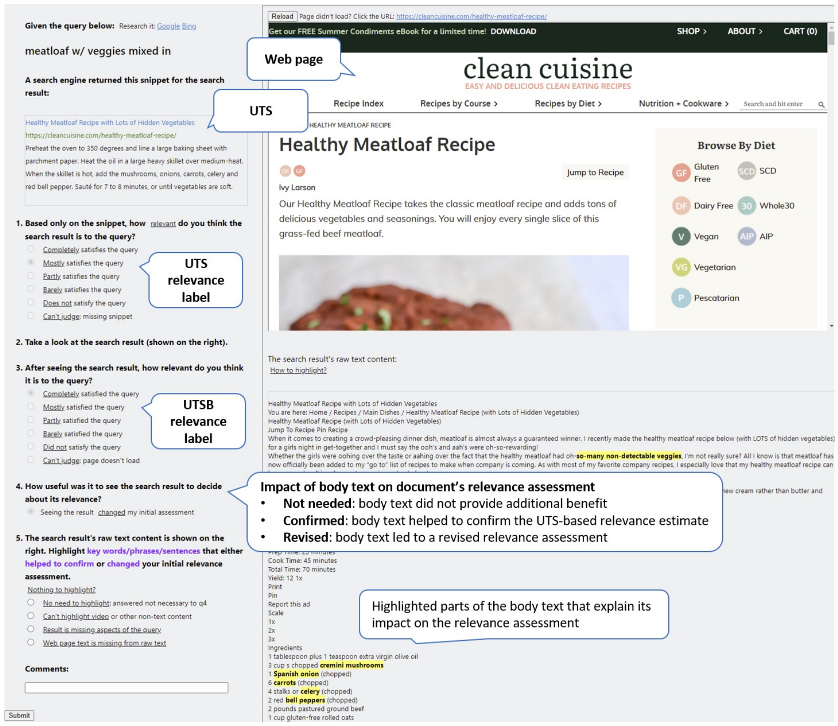

Fig. 1: Human Intelligent Task to collect UTS and UTSB labels from assessors

UTSB, respectively. Query and UTS have an expected length of less than 128 tokens, so we use an input sequence length of 512 tokens in all our experiments, truncating the input if it is longer. This allows the UTSB model to see significantly more document text than is seen from snippet alone, and allows us to observe systematic differences between UTS and UTSB. We use the [CLS] vector as input to a single layer neural network to obtain the probability of the document being relevant. We refer to the probability prediction values as UTS and UTSB ranking scores and use the ranking orders they impose to study whether neural models benefit from the body text.

For our dataset, we sample 1k queries at random from Bing's search logs, then for each query, we scrape the Bing SERP and collect a total of 12k query-URL pairs. We collect human labels for every query-URL and run ranking experiments with our dataset as the test set. For our investigation of when the body text impacts a human/machine assessor, we focus on the query and document properties listed in Table 1.

Table 1: Query and document features

Variable	Description
Performance predictor	Output of a proprietary query performance prediction model ($\in$[0,1])
Query type: Navigational	Classifier output predicting if the query is navigational (1) or not (0)
Query type: Head/tail	Predicted query popularity ($\in[0(tail),1(head)]$)
Query type: Question	If the query is a natural language question ($\in[0(no),1(yes)]$)
Lengths	Query, URL, Title, Snippet, and Body lengths in characters
% of query tokens	The ratio of query tokens that appear in the URL, Title, Snippet, Body

Table 2: The UTSB model's performance improvement over the UTS model, measured using RBP (on a 100 point scale) and either the UTS or UTSB human labels as ground-truth (GT).

	ΔRBP@3	ΔRBP@10
UTS label GT	0.165	0.071
UTSB label GT	0.797	0.587
% improved/degraded	33/31	45/43

Table 3: Reasons when human assessors could not highlight parts of the body text to explain why it was beneficial over the UTS

	UTS>UTSB	UTS<UTSB
Missing term	76%	12%
Other	20%	48%
Video	4%	40%

4 Results and Discussions

Impact of body text on human assessors: We stipulate that UTS alone is insufficient in cases when human assessors either revised their initial assessment upon seeing the body text ('Revised') or when the body text was needed to confirm their UTS label ('Confirmed'). Overall, assessors indicated that UTS alone was insufficient (body text was beneficial) in 48% of the cases. Of these, 'Revised' made up 59% and 'Confirmed' the other 41%. When assessors revised their ratings, they initially overestimated the document's relevance in 54% of cases (UTS>UTSB) and underestimated it in 46% of cases (UTS<UTSB). The higher ratio of overestimates could hint at possible SEO manipulation methods succeeding or assessors exhibiting confirmation bias with UTS. Using statistical analysis (t-test) to compare the sample means of the query document properties (Table 1) across cases where the body text benefited judges or not, we found that the body text was helpful for predictably poor performing, long, not-navigational, tail and question type queries (all stat. sig. $p<0.01$).

Impact of body text on neural ranker: We assume that UTS is insufficient when the UTSB model outperforms the UTS model. We calculate the two models' performance using RBP (Moffat and Zobel, 2008) with both the human UTS and UTSB labels as ground-truths. As it can be seen in Table 2, the UTSB model outperforms the UTS model (ΔRBP>0), where the benefit from body text is more evident at the top ranks (ΔRBP@3>ΔRBP@10). We also see that the ranker learns to make better use of the body text when the training labels also consider the body text (2nd row). Looking at the ratio of queries where the UTSB model outperforms the UTS model (3rd row), we see that there is room for improvement: the percentage of queries that benefit from the body text is just higher than those that body text degrades. Differences in the sample means of the query document properties (Table 1) for the improved and degraded queries reveals that improved queries are long, tail, not-navigational and of question type, while degraded queries are short, head and navigational, and the documents long (all stat. sig. $p<0.01$).

Explanation of body text's impact: We make use of the interpretML framework[1] and train two Explainable Boosting Machine (EBM) glassbox regression models (tree-based, cyclic gradient boosting Generalized Additive Models) (Lou et al., 2013). For each query-URL pair input, we use the properties listed in Table 1 as features and construct the target labels as follows:

- **ΔLabel**: Target label for the EBM model used to explain human assessors' reaction to seeing the body text, mapped as -1 if UTS>UTSB (UTS label overestimated document relevance), 0 if UTS=UTSB, and 1 if UTS<UTSB (UTS underestimated).
- **ΔRank**: To model the neural rankers' reaction we opt to use the ranking position (rp) since the UTS and UTSB scores are not directly comparable (different trained models) and use -1 if UTS rp < UTSB rp (UTSB model's relevance estimation

[1] https://interpret.ml/docs/ebm.html

Table 4: The EBM models' top 5 feature importance scores for human and machine assessors, explaining the delta observed in the human assessors' UTS and UTSB labels (ΔLabel) and the neural models' UTS and UTSB based rankings (ΔRank), respectively.

ΔLabel (UTSB label - UTS label)	**ΔRank** (UTS rp - UTSB rp)
Question (0.2825)	%QueryWords in Tokenized Body (0.2858)
Body length (0.2434)	Snippet length (0.2831)
Performance predictor (0.2418)	Title length (0.2478)
%QueryWords in Tokenized Title (0.2218)	Body length (0.1658)
Query length (0.2141)	%QueryWords in Tokenized Snippet (0.1459)

decreased compared to UTS), 0 if UTS rp = UTSB rp, and 1 if UTS rp > UTSB rp (UTSB's estimate increased compared to UTS).

Table 4 shows the EBM models' top 5 feature importance scores for human and machine assessors, telling us which of the query and document properties explain the delta observed in the human assessors' UTS and UTSB labels (ΔLabel) and the neural models' UTS and UTSB based rankings (ΔRank), respectively. We can see that a change in labels or rankings is explained by very different factors: body length is the only common factor in the top 5. The top explanation of change in the humans' UTS vs UTSB assessments is whether the query is phrased as a question, while the top reason for the ranker is the ratio of query tokens that are present in the body text.

To examine how ΔLabel and ΔRank change with each feature, in Figure 2, we plot EBM's learnt per-feature functions. Each plot shows how a given feature contributes to the model's prediction. For example, the Query length plot shows that for short queries, human assessors (blue line) are more likely to underestimate (y>0) the document's relevance based on UTS alone, while for long queries, they tend to overestimate (y<0). The neural model (orange line) shows a similar but more subtle trend: for short queries, body text increases the ranker's relevance estimate over the UTS estimate, while for long queries the predicted relevance decreases with body text. The Question plot shows that humans tend to underestimate the document's relevance when the query is more likely to be a question. This indicates that document summaries fail to convince searchers that the document answers their question. The ranker's predicted relevance, however, decreases with body text for question type queries. Looking at the Snippet length plot, we see that the neural model is more likely to decrease its estimate of the document's relevance with body text when snippets are short, but increase it for long snippets. This suggests that when snippets include more context, the ranker is more likely to see these as evidence of irrelevance, which is diminished when body text is added. Snippet length has the opposite impact on humans: the longer the snippet, the more likely they overestimate the result's relevance. Overall, we see very little similarities (parallel trends) in the human vs ranker feature plots, indicating that humans and machines react to body text in fundamentally different ways. Human assessors are more likely to overestimate relevance from UTS for long, tail, and not-navigational queries, and underestimate when the query is head, navigational or a question. They also overestimate for long snippets and short documents, and underestimate for long documents and short snippet. Unlike humans, the neural model results in more near-flat plots: the most impact is seen for document (rather than query) properties, e.g., Snippet length and ratio of query tokens in the snippet and body.

Additional considerations: When assessors revised their relevance assessment but were unable to highlight parts of the body text to explain the change (in 72% of overestimates and 22% of underestimates), they were asked to indicate a reason.

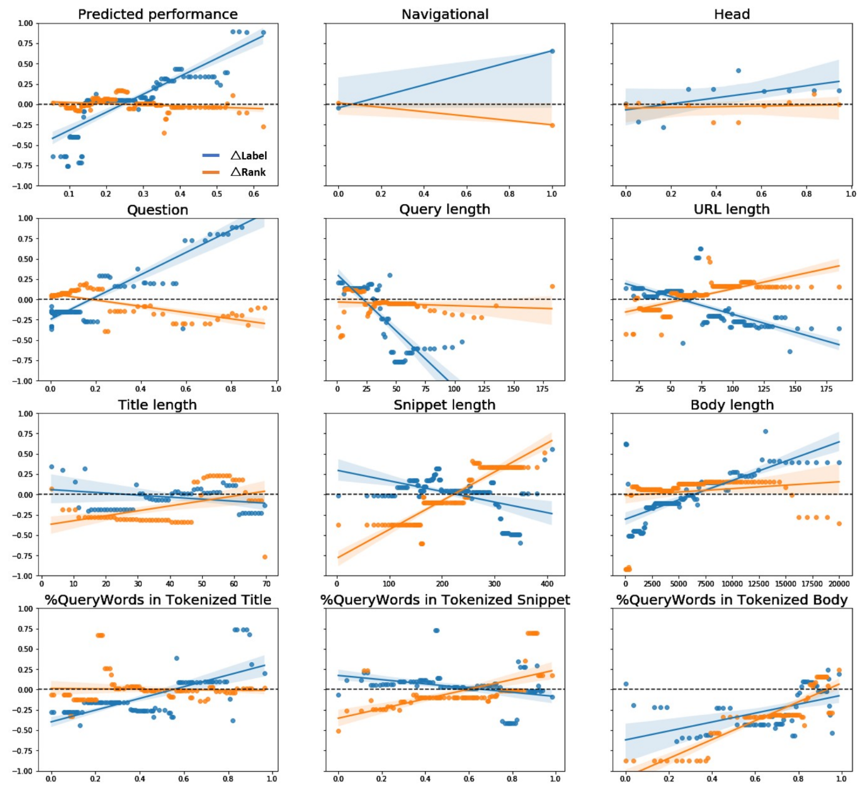

Fig. 2: EBM's learnt feature functions for each query and document feature, explaining the Δ changes: $y > 0$ means that 'seeing' the body text led to an increase in the relevance estimate compared to UTS

Table 3 shows that the absence of query terms in the document was the main reason for overestimates without highlighted text (76%). This suggests that informing users of missing query terms on the SERP is a helpful strategy. On the other hand, a major reason when assessors underestimated a document was when video (or other non-textual content) was present on the page (40%) - an aspect that was not considered by the neural model.

5 Conclusions

We studied when human and machine assessors benefit from the full text of the document to estimate its relevance. We showed that both humans and BERT style models benefit from the body text in similar cases (long, not navigational, tail and question type queries), but that full text impacts their relevance assessments in different ways (e.g., full text increases humans' relevance estimates but decreases the ranker's). In addition, we observe differences in the properties of queries where the BERT model's performance improves or degrades with the full text, e.g., performance degrades for navigational queries (ΔRBP@3 of -1.07). This indicates that more work is necessary on BERT style models when considering full text as input or that different types of queries (e.g., head v tail) require models to be optimized differently. While our findings are a function of the query-biased summaries, the observed differences in human and model reactions to additional information indicate that different mechanisms are needed for human vs machine inputs.

Bibliography

Bendersky, M., Kurland, O.: Utilizing passage-based language models for document retrieval. In: Proc. of ECIR (2008)

Bolotova, V., Blinov, V., Zheng, Y., Croft, W.B., Scholer, F., Sanderson, M.: Do people and neural nets pay attention to the same words: Studying eye-tracking data for non-factoid qa evaluation. In: Proc. of CIKM, p. 85–94 (2020), URL https://doi.org/10.1145/3340531.3412043

Clarke, C.L., Agichtein, E., Dumais, S., White, R.W.: The influence of caption features on clickthrough patterns in web search. In: Proc. of SIGIR, pp. 135–142, ACM (2007)

Cutrell, E., Guan, Z.: What are you looking for? an eye-tracking study of information usage in web search. In: Proc. of the SIGCHI conference on Human factors in computing systems, pp. 407–416 (2007)

Demeester, T., Nguyen, D., Trieschnigg, D., Develder, C., Hiemstra, D.: What snippets say about pages in federated web search. In: Asia Information Retrieval Symposium, pp. 250–261, Springer (2012)

Demeester, T., Nguyen, D., Trieschnigg, D., Develder, C., Hiemstra, D.: Snippet-based relevance predictions for federated web search. In: Proc. of ECIR, pp. 697–700, Springer (2013)

Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: 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), pp. 4171–4186, Association for Computational Linguistics, Minneapolis, Minnesota (Jun 2019), https://doi.org/10.18653/v1/N19-1423, URL https://www.aclweb.org/anthology/N19-1423

Edmundson, H.: Problems in automatic abstracting. Communications of the ACM **7**(4), 259–263 (1964)

Hofstätter, S., Mitra, B., Zamani, H., Craswell, N., Hanbury, A.: Intra-document cascading: Learning to select passages for neural document ranking. In: Proc. SIGIR, ACM, ACM (2021)

Hofstätter, S., Zamani, H., Mitra, B., Craswell, N., Hanbury, A.: Local Self-Attention over Long Text for Efficient Document Retrieval. In: Proc. of SIGIR, ACM (2020)

Kaisser, M., Hearst, M.A., Lowe, J.B.: Improving search results quality by customizing summary lengths. In: Proc. of ACL-08: HLT, pp. 701–709 (2008)

Lagun, D., Agichtein, E.: Viewser: Enabling large-scale remote user studies of web search examination and interaction. In: Proc. of SIGIR, pp. 365–374, ACM (2011)

Lagun, D., Agichtein, E.: Re-examining search result snippet examination time for relevance estimation. In: Proc. of SIGIR, pp. 1141–1142, ACM (2012)

Li, C., Yates, A., MacAvaney, S., He, B., Sun, Y.: Parade: Passage representation aggregation for document reranking. arXiv preprint arXiv:2008.09093 (2020)

Lin, J., Nogueira, R., Yates, A.: Pretrained transformers for text ranking: Bert and beyond. arXiv preprint arXiv:2010.06467 (2020)

Lou, Y., Caruana, R., Gehrke, J., Hooker, G.: Accurate intelligible models with pairwise interactions. In: Proc. of the 19th ACM SIGKDD International Conference on

Knowledge Discovery and Data Mining, p. 623–631, KDD'13, ACM (2013), URL https://doi.org/10.1145/2487575.2487579
Luhn, H.P.: The automatic creation of literature abstracts. IBM Journal of research and development **2**(2), 159–165 (1958)
Moffat, A., Zobel, J.: Rank-biased precision for measurement of retrieval effectiveness. ACM Trans. Inf. Syst. **27**(1) (Dec 2008), URL https://doi.org/10.1145/1416950.1416952
Nogueira, R., Cho, K.: Passage re-ranking with BERT. CoRR **abs/1901.04085** (2019), URL http://arxiv.org/abs/1901.04085
Robertson, S., Zaragoza, H., et al.: The probabilistic relevance framework: Bm25 and beyond. Foundations and Trends® in Information Retrieval **3**(4), 333–389 (2009)
Salton, G., Allan, J., Buckley, C.: Approaches to passage retrieval in full text information systems. In: Proc. of SIGIR, ACM (1993)
Sanderson, M.: Accurate user directed summarization from existing tools. In: Proc. of CIKM, p. 45–51, ACM (1998), URL https://doi.org/10.1145/288627.288640
Savenkov, D., Braslavski, P., Lebedev, M.: Search snippet evaluation at yandex: lessons learned and future directions. In: International Conference of the Cross-Language Evaluation Forum for European Languages, pp. 14–25, Springer (2011)
Tombros, A., Sanderson, M.: Advantages of query biased summaries in information retrieval. In: Proc. of SIGIR, pp. 2–10, ACM (1998)
Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention is all you need. In: Proc. of NeurIPS (2017)
White, R.W., Jose, J.M., Ruthven, I.: A task-oriented study on the influencing effects of query-biased summarisation in web searching. Information Processing & Management **39**(5), 707–733 (2003)
Yan, M., Li, C., Wu, C., Bi, B., Wang, W., Xia, J., Si, L.: Idst at trec 2019 deep learning track: Deep cascade ranking with generation-based document expansion and pre-trained language modeling. In: TREC (2019)
Yan, M., et al.: Idst at trec 2019 deep learning track: Deep cascade ranking with generation-based document expansion and pre-trained language modeling. In: TREC (2020)
Yue, Y., Patel, R., Roehrig, H.: Beyond position bias: Examining result attractiveness as a source of presentation bias in clickthrough data. In: Proc. of the 19th International Conference on World Wide Web, pp. 1011–1018, ACM (2010)