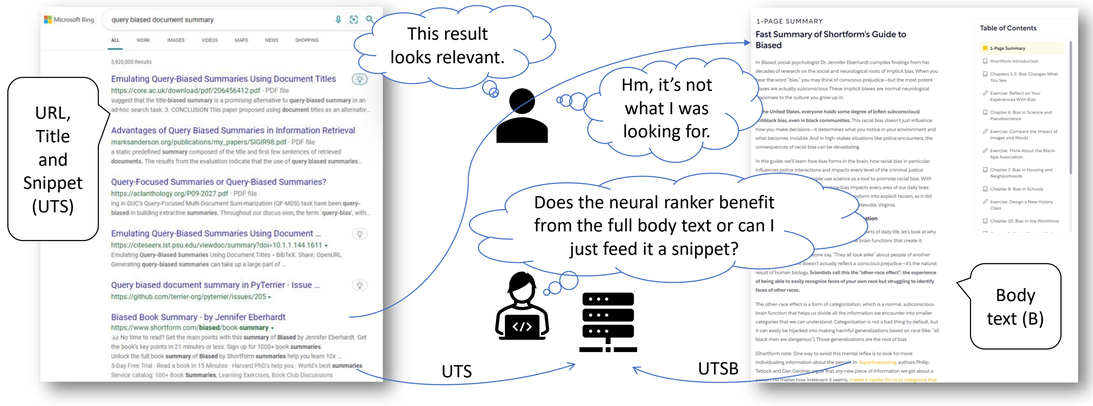
# Less is Less: When are Snippets Insufficient for Human vs Machine Relevance Estimation?

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

Microsoft

## Goal

- When do relevance estimates by a ranking model or a human assessor benefit from the document's full text?
- Do humans and machines benefit from the document's full text in similar ways?




## Experiment setup

- UTS vs UTSB as input to human and machine relevance estimation.
- Measure impact of body text on relevance estimates.

## Human relevance assessment

- Multi-step Human Intelligence Task (HitApp).
- Collect UTS relevance label first.
- Then reveal web page and collect UTSB relevance label.
- Did body text provide benefit? Not needed / Helped to confirm / Revised label.

## Neural ranker based relevance estimates

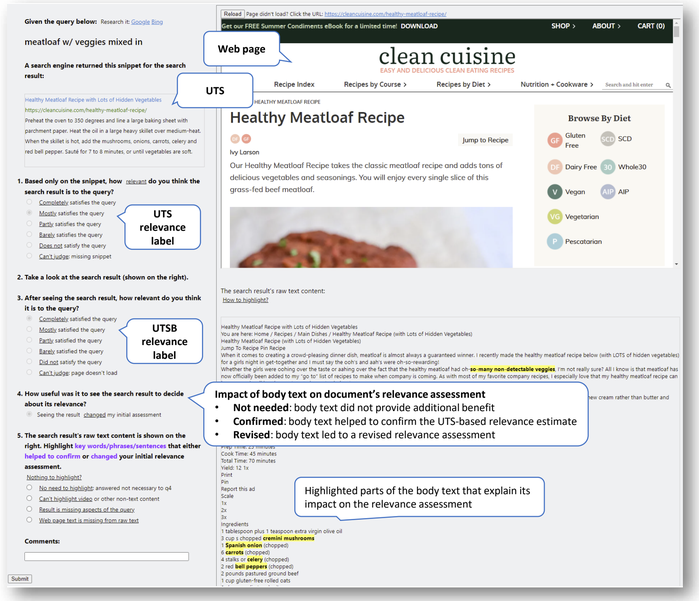
- Separate UTS and UTSB trained models, starting with pre-trained BERT-style model.
- Sentence A is query, Sentence B is UTS or UTSB (512 tokens).

## Data

- 1k random Bing queries -> 12k QU pairs.
- Query and document properties in Table 1.

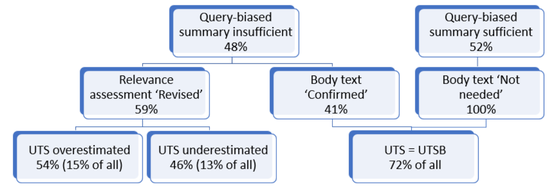
**Table 1.** Query and document features

| 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 |




## Impact of seeing body text on human relevance assessments

- Body text helped in 48% cases.
- For 28%, seeing the body lead to revised label.
- Body text helped predictably poor performing, long, tail, not-navigational, and question type queries (p<0.01).




## Impact of seeing body text on neural ranker performance

- Body text helps ranker: UTSB model outperforms UTS (ΔRBP>0).
- Benefit more evident at the top ranks (ΔRBP@3>ΔRBP@10).
- Body text can degrade performance for some queries.
- Improved queries are long, tail, not-navigational and of question type, while degraded queries are short, head and navigational, and the documents long (p<0.01).

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 |

## When does body impact human and machine relevance estimates

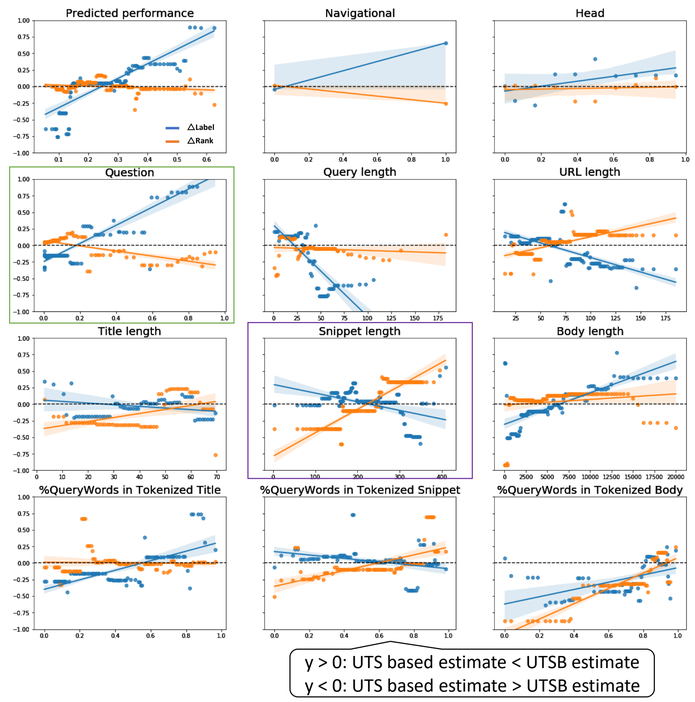
- Trained two EBM regression models with query/document properties as features and ΔLabel and ΔRank as targets.
- Only common feature is body length among top 5 features that explain ΔLabel and ΔRank, resp., see Table 4.
- Humans and machine react to body text for different query/document types.

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) |

## Per feature analysis

- **Question type**: Humans (blue) underestimate relevance based on UTS for question queries. ⇔The ranker's (orange) predicted relevance decreases with body text for question queries.
- **Snippet length**: The neural model's relevance estimate decreases with body text for short snippets and increases for long snippets. ⇔ The longer the snippet, the more likely that humans overestimate relevance.
- Humans and machines react to body text in fundamentally different ways.




## Conclusions

- Studied when human and machine assessors benefit from the full text of the document to estimate its relevance.
- Both humans and BERT style models benefit from the body text in similar cases (for long, not navigational, tail and question type queries), but full text impacts their relevance estimations in very different ways.
- The BERT model's performance improves or degrades with the full text depending on query property. E.g., performance degrades for navigational queries (ΔRBP@3 of -1.07).
- Different types of queries (e.g., head v tail) require models to be optimized differently.