TREC CAR Y3 Results Analysis

Laura Dietz; UNH

Ben Gamari; Well-Typed

John Foley; Smith College

jjfoley@smith.edu

Methods Submitted (Summary)

- BERT-based re-ranking
- BM25 in various flavors (Lucene, Anserini, ???)
- CombMNZ of Terrier ranking models (IRIT)
- Neural Models (BiLSTM, etc.) for re-ranking
- Some query expansion, e.g., BM25+RM3
- Document-Entity models (UNH)

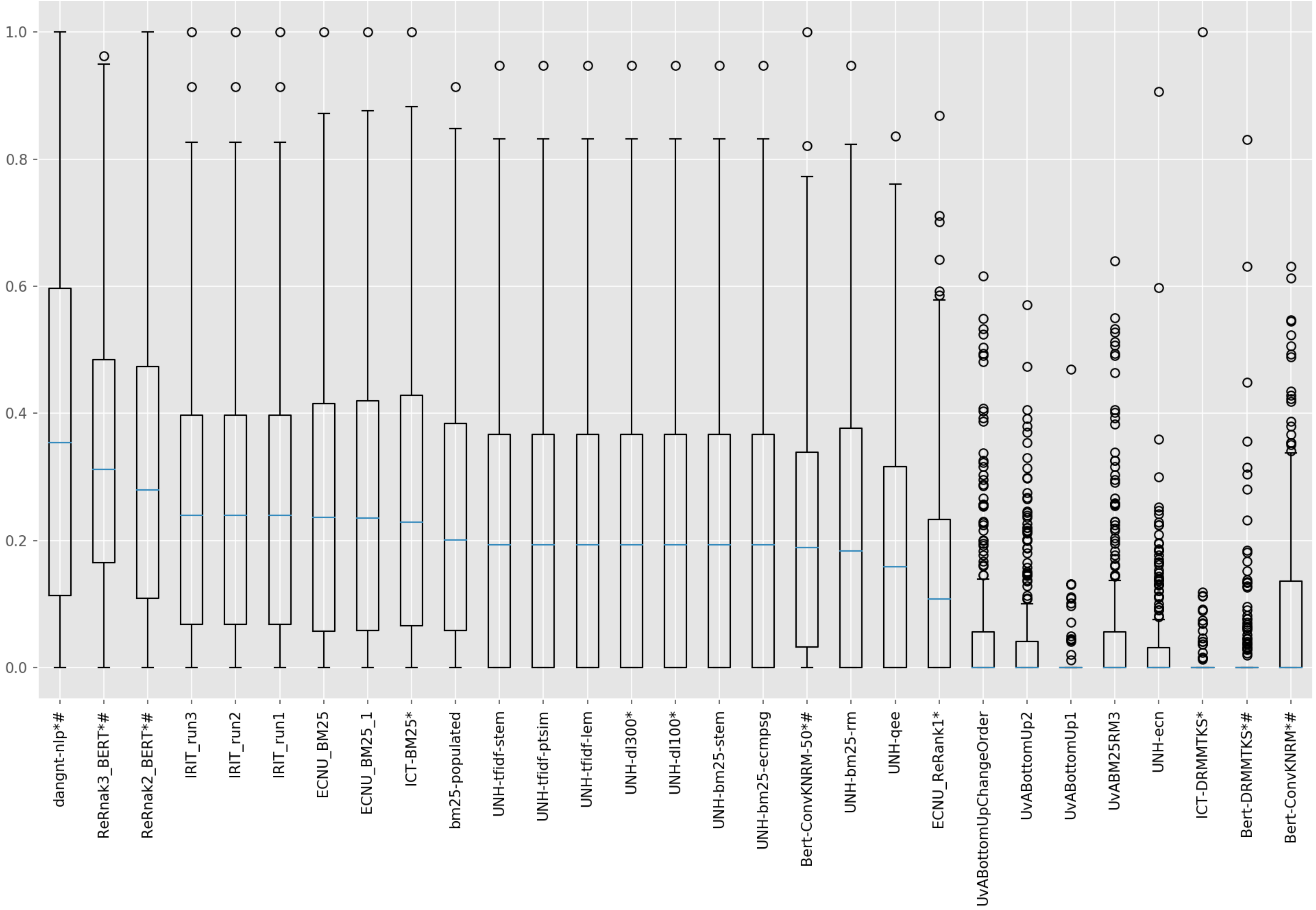
Methods Submitted (Summary)

- BERT-based re-ranking
- BM25 in various flavors (Lucene, Anserini, ???)
- CombMNZ of Terrier ranking models (IRIT)
- Neural Models (BiLSTM, etc.) for re-ranking
- Some query expansion, e.g., BM25+RM3
- Document-Entity models (UNH)

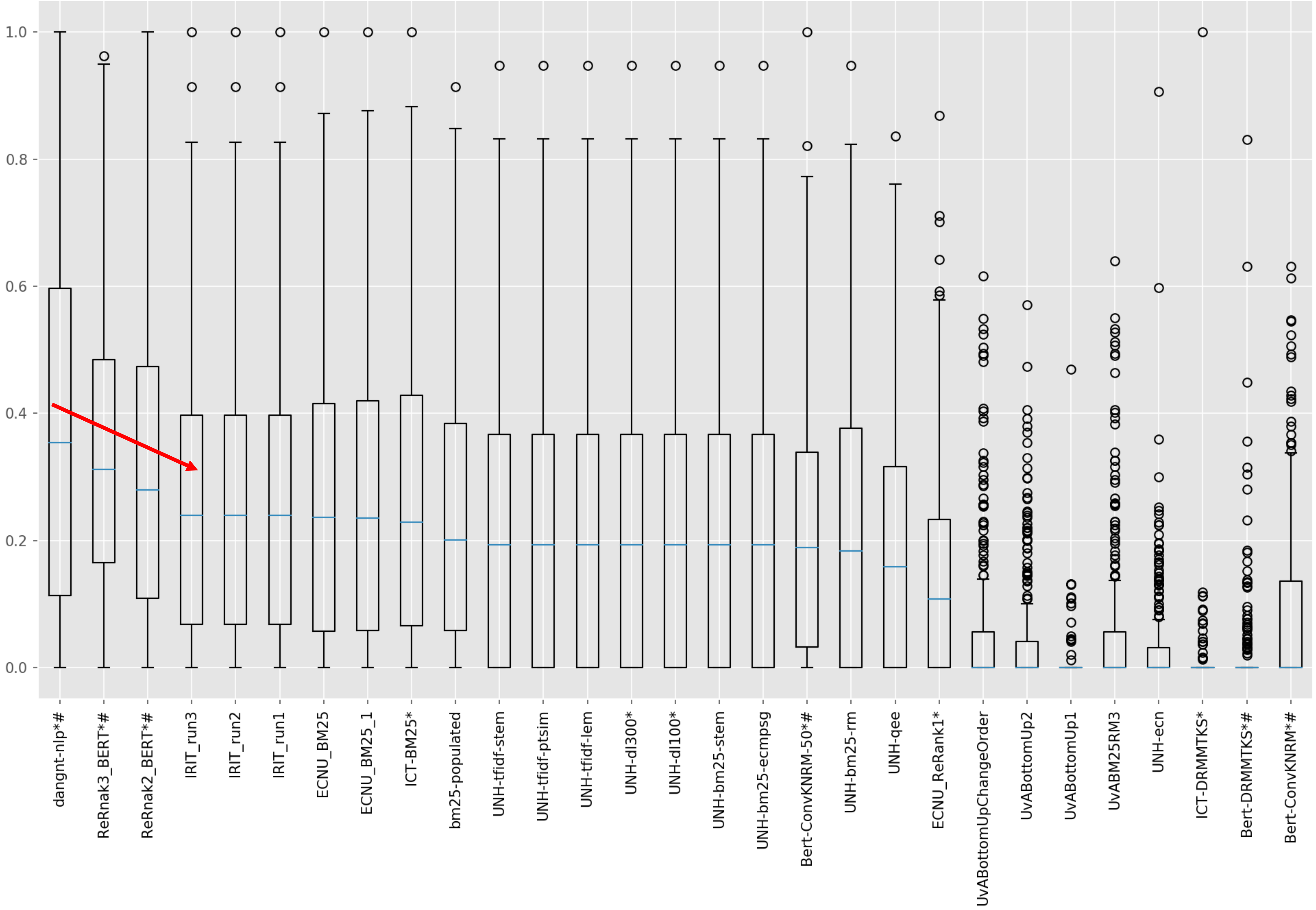
Per-Heading Evaluation

- Each paragraph was linked to a topic by assessors
  - Not-Relevant = 0
  - Can-Label = 1
  - Should-Label = 2
  - Must-Label = 3
- Use NDCG to take advantage of graded relevance.
- Use paragraph_origins from runs to identify queries
  - Evaluating the task like that of prior years
- Pairwise Randomization Test applied to select pairs.

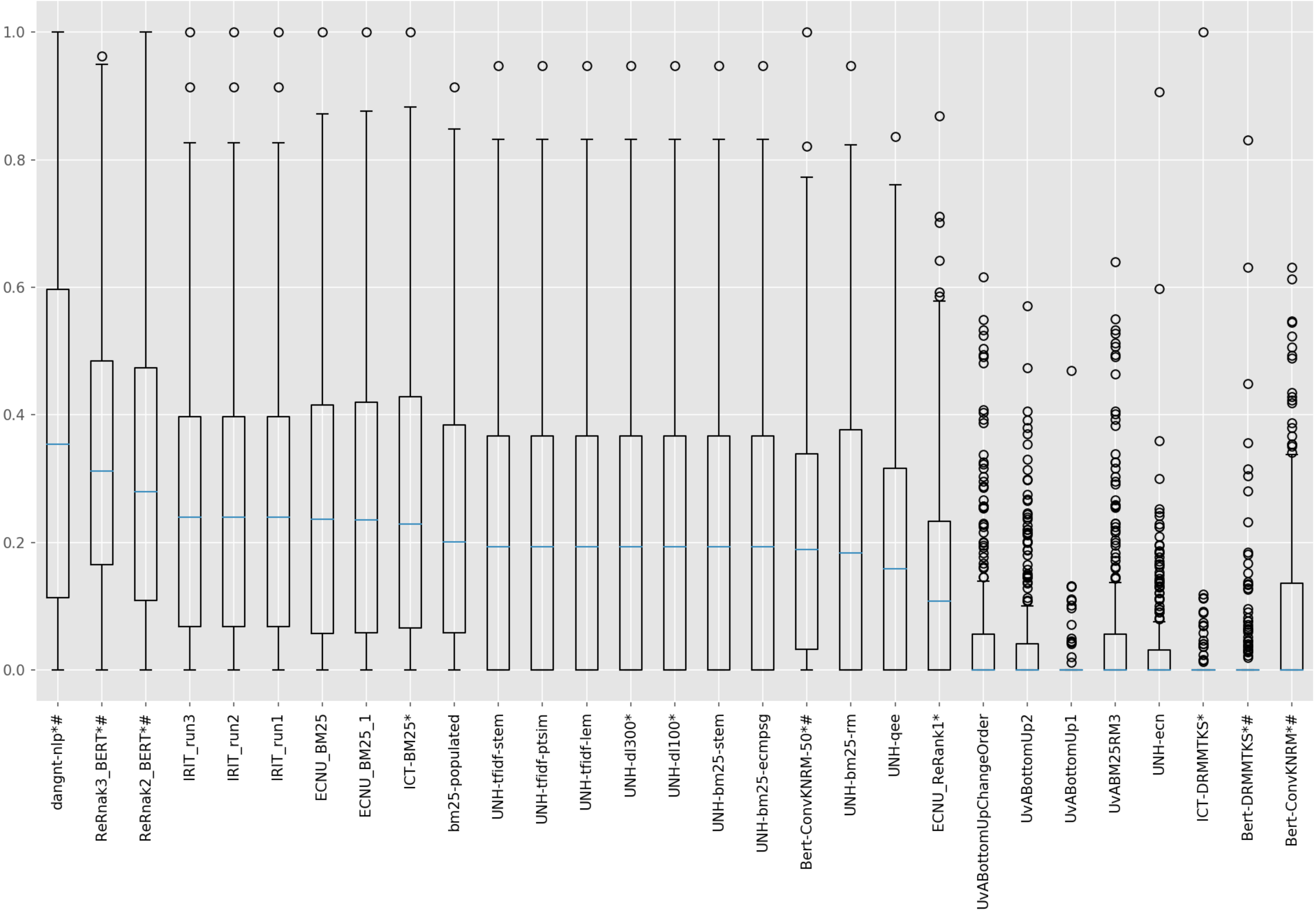
Per-Heading-NDCG



Per-Heading-NDCG

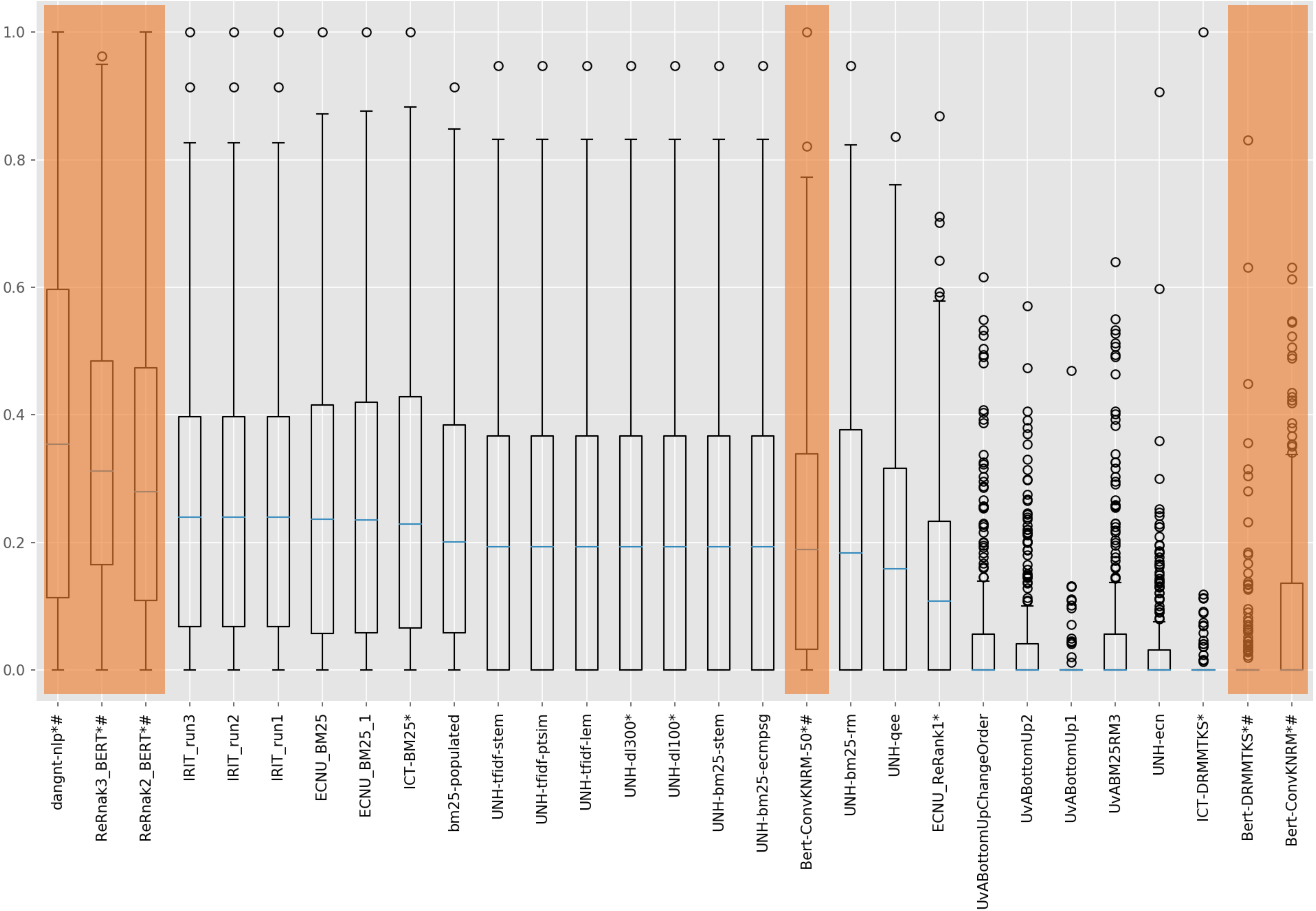


Per-Heading-NDCG



BERT

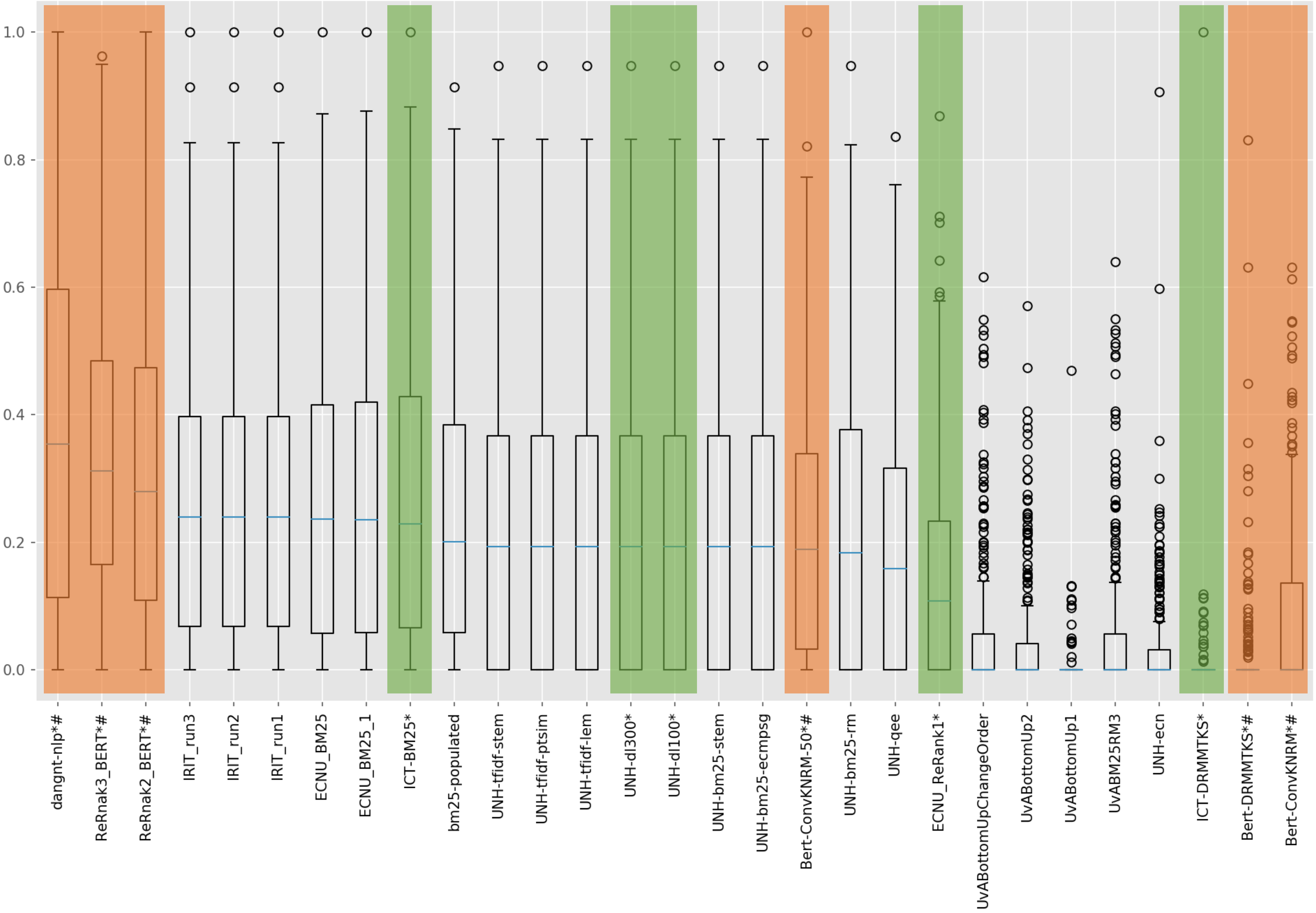
Per-Heading-NDCG



BERT

Neural

Per-Heading-NDCG



Ranking-Task Conclusions

- BERT takes the lead, but not without careful tuning.
  - Top-3 runs include BERT … all different ($p < 0.01$)
  - Bottom-2 runs also include BERT
- Open Research Questions:
  - Has BERT seen Wikipedia before? Is this a reasonable method?
  - How to generate good/coherent articles?
    - Most if not all teams used population script.
- Other Neural approaches fairly indistinguishable from BM25
  - Some BM25 runs are much stronger than others

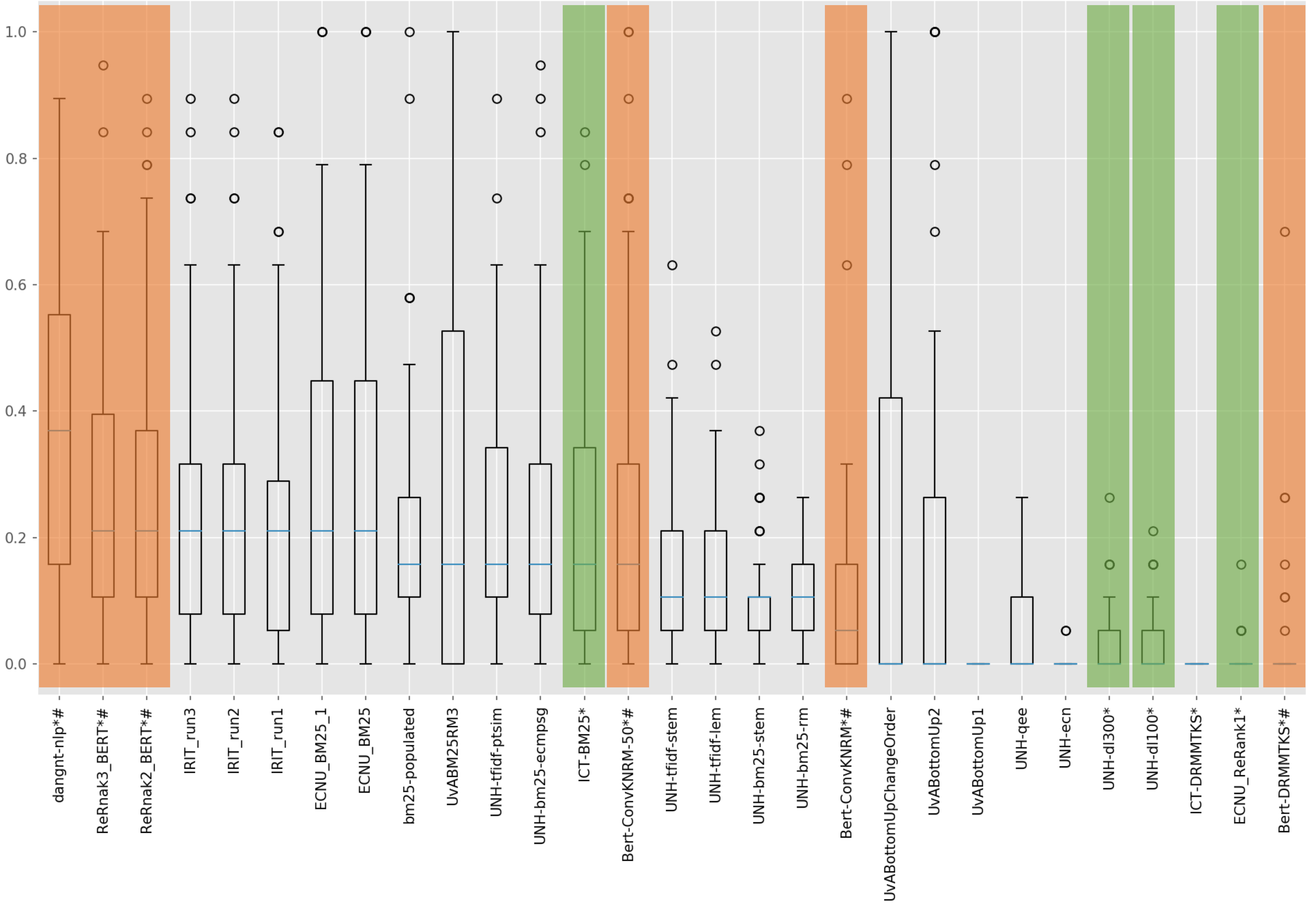
Coherence / Transition Quality

- No teams spent much time on this; preferring to treat CAR as a standard ranking task.
- Assessors marked the relationship between two paragraphs as:
  - Same-Topic ☺ => 1
  - Appropriate-Transition ☺ => 1
  - Switch-Transition ☹ => 0
- The coherence measure is the mean of the provided transitions.
  - Somewhere between 0 (all shuffled paragraphs) and 1 (all perfect transitions)

Per-Query Transition Quality

🟧 BERT

🟩 Neural



Coherence Conclusions

- Definitely correlated with retrieval quality, but not exactly.
  - System ordering roughly similar.
- Neural (but-not-BERT) models are much lower in the ranking.
  - Unclear if this is meaningful or spurious.
- Again, most did not tackle this challenge.
  - Could study this independently atop existing TREC runs…

CAR is over!

- Thanks to everyone who participated.
- Analysis/plotting code is publicly-available
  - VSCode / Jupyter Notebook style: #%%
  - https://github.com/jjfiv/car2019eval