• Achieved the best performances in terms of nDCG, Q, iRBU

- KASYS-E-CO-NEW-1 (Fine-tuning with MS MARCO & TREC MB)
- KASYS-E-CO-NEW-4 (Fine-tuning with MS MARCO & TREC MB)
- KASYS-E-CO-NEW-5 (Fine-tuning with TREC CAR & TREC MB)

Key Ideas in Birch

Common Web Documents Do Not Fit BERT

• Using Query and Sentence Relevance
• Using Models Learned in Other Tasks

Details of Birch

• Linear sum of the BM25 and the score of the highest BERT-scoring sentence in the document
  - Assuming that the most relevant sentences in a document are good indicators of the document-level relevance
  - fBM25(d): The BM25 score of document d
  - fBERT(p): The sentence relevance of the top-pth sentence obtained by BERT
  - \( f(d) = f_{BM25}(d) + \sum_{p=1}^{k} w_p \cdot f_{BERT}(p) \)

Implementation Details

• Features for learning to rank
  - TF, IDF, TF-IDF, document length, BM25 score, and language-model-based IR scores

Experimental Results

Successfully replicated with the original WWW-2 qrels

WWW-3 official results

Seems that KASYS failed to replicate and reproduce the THUIR run