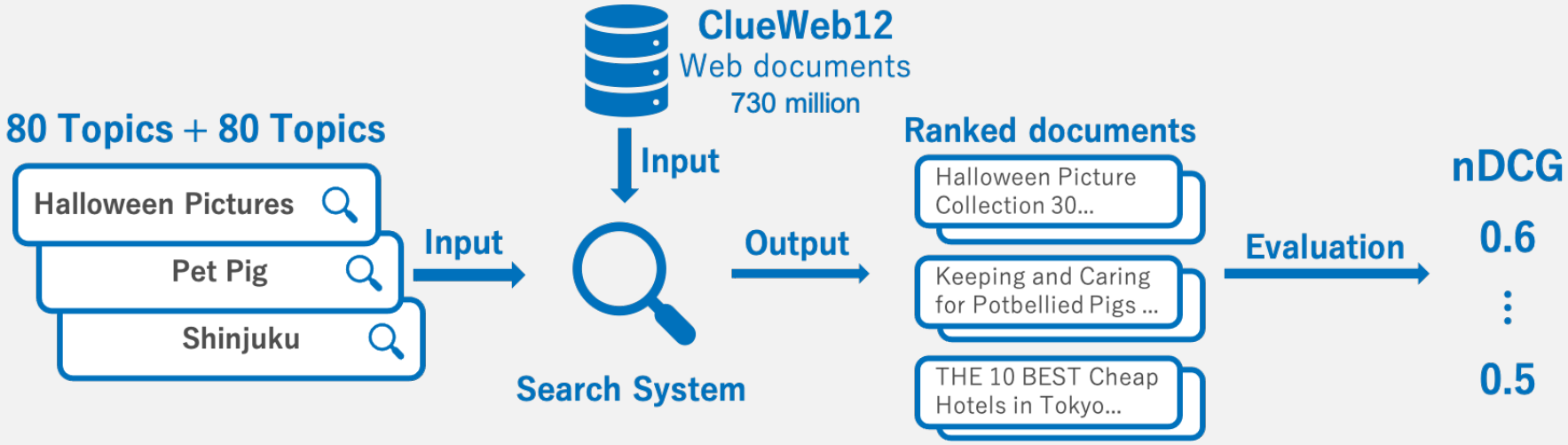
KASYS
筑波大学知識獲得システム研究室
UNIVERSITY OF TSUKUBA
KNOWLEDGE ACQUISITION SYSTEM LAB.

KASYS at the NTCIR-15 WWW-3 Task

Kohei Shinden, Atsuki Maruta, Makoto P. Kato

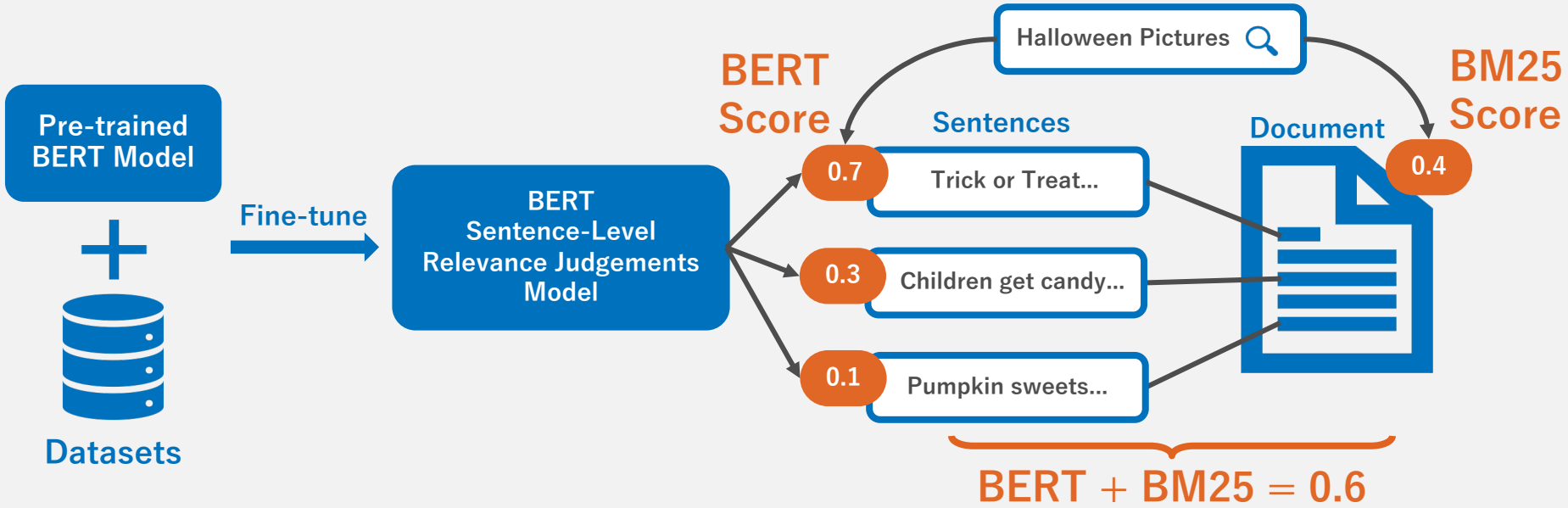
University of Tsukuba

- **NTCIR-15 WWW-3 Task**
  - Ad-hoc document retrieval tasks for web documents



- **Proposed search model using BERT (Birch)**
  - Yilmaz et al: Cross-Domain Modeling of Sentence-level Evidence for Document Retrieval, EMNLP 2019
  - BERT has been successfully applied to a broad range of NLP tasks including document ranking tasks.

- **Applying a sentence-level relevance estimator learned by QA and microblog search datasets to ad-hoc document retrieval**



1. The sentence-level relevance estimator is obtained by fine-tuning the pre-trained BERT model with QA and microblog search data.
2. Calculate BM25 scores and BERT scores for query and document sentences.
3. Weighted sum of the BM25 and the score of the highest BERT-score sentence in the document.

- **Weighted 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 [1]
    - $f_{\mathrm{BM25}}(d)$: The BM25 score of document d
    - $f_{\mathrm{BERT}}(p_i)$: The sentence relevance of the top i-th sentence obtained by BERT
    - w_i : The hyper-parameter w_i is to be tuned with a validation set

$$f(d) = f_{BM25}(d) + \sum_{i=1}^{k} w_i \cdot f_{BERT}(p_i)$$

[1] Yilmaz et al: Cross-Domain Modeling of Sentence-level Evidence for Document Retrieval, EMNLP 2019

- **Preliminary experiments to select datasets and hyper-parameters suitable for ranking web documents**

Train

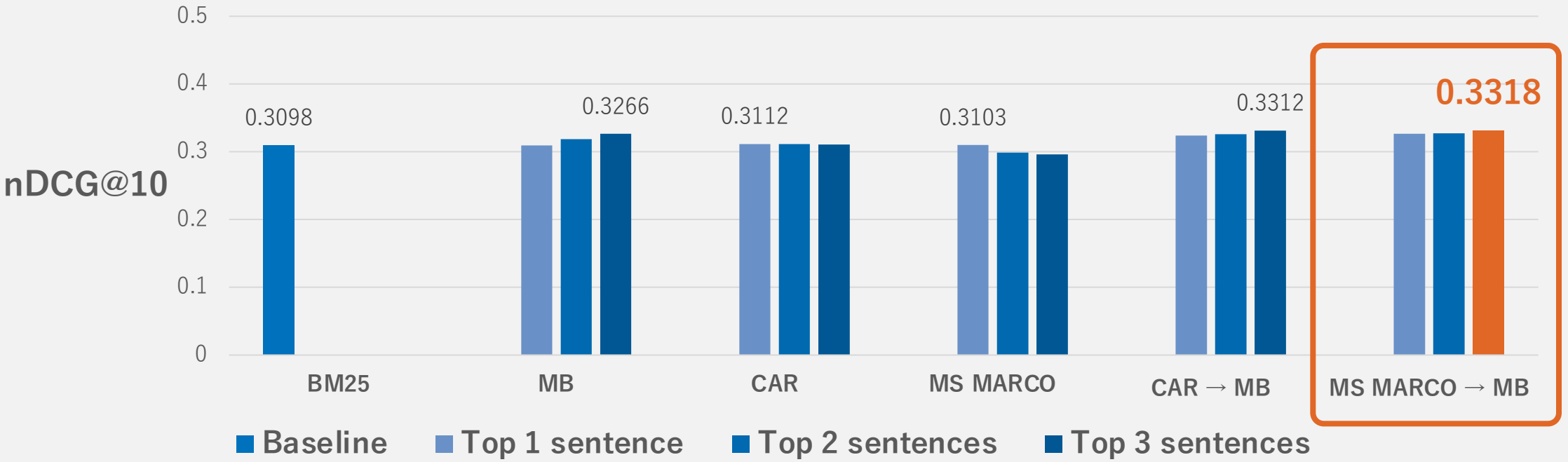
	Robust04	MS MARCO	TREC CAR	TREC MB
Model MB	✓			✓
Model CAR	✓		✓	
Model MS MARCO	✓	✓		
Model CAR → MB	✓		✓	✓
Model MS MARCO → MB	✓	✓		✓

The checkmarks represent the data set used for training.

Validation

NTCIR-14 WWW-2
Test Collection
(with its original qrels)

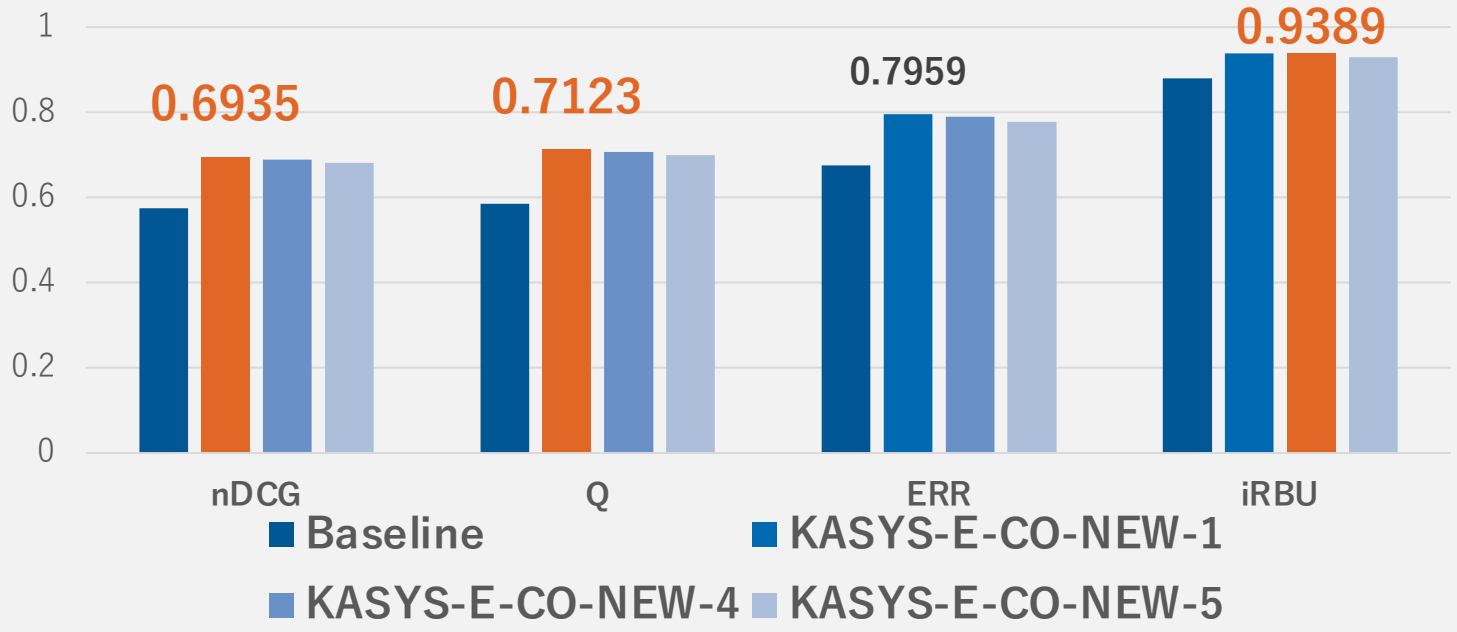
- **Evaluated the prediction results of Birch models**
  - Top k sentences: Uses the k-sentence with the highest BERT score for ranking



MSMARCO → MB is the best.

Thus, we submitted runs based on MS MARCO → MB and CAR → MB.

- **Achieved the best performances in terms of nDCG, Q and iRBU among all the participants.**



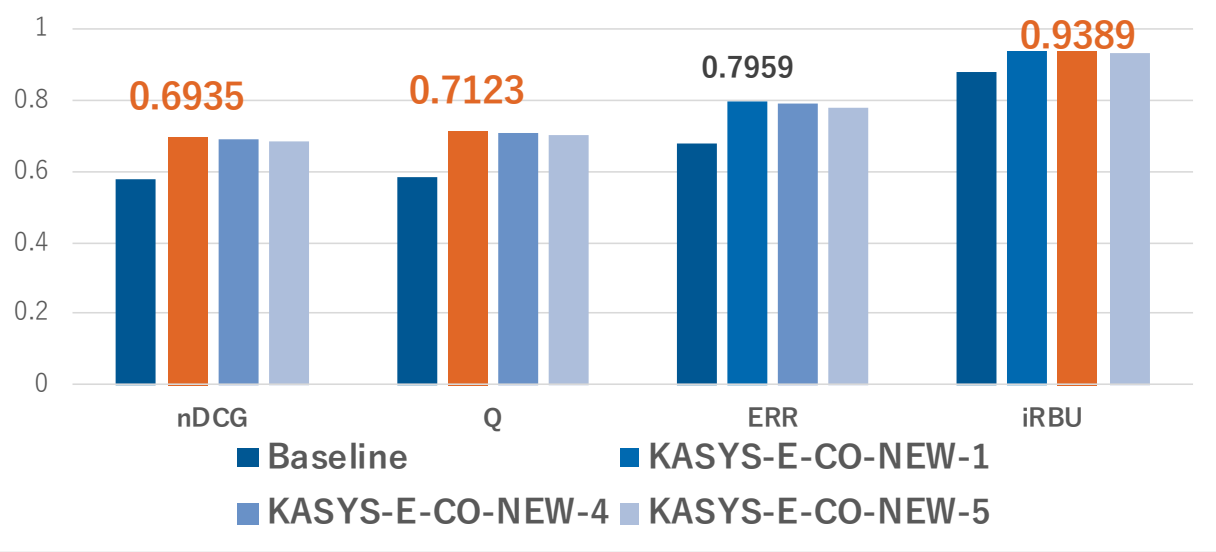
KASYS-E-CO-NEW-1:
- MS MARCO→MB
- Top 3 sentences

KASYS-E-CO-NEW-4:
- MS MARCO→MB
- Top 2 sentences

KASYS-E-CO-NEW-5:
- CAR→MB
- Top 3 sentences

- **MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.**
  - The reason why MS MARCO and TREC CAR's results are better probably because they are web documents retrieval and have a large amount of data.
- **BERT is also valid for web document retrieval.**

- **Achieved the best performances in terms of nDCG, Q and iRBU among all the participants.**
- **The effectiveness of BERT in ad hoc web document retrieval tasks was verified.**
  - MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.
  - BERT is also valid for web document retrieval.



KASYS-E-CO-NEW-1:
- MS MARCO→MB
- Top 3 sentences

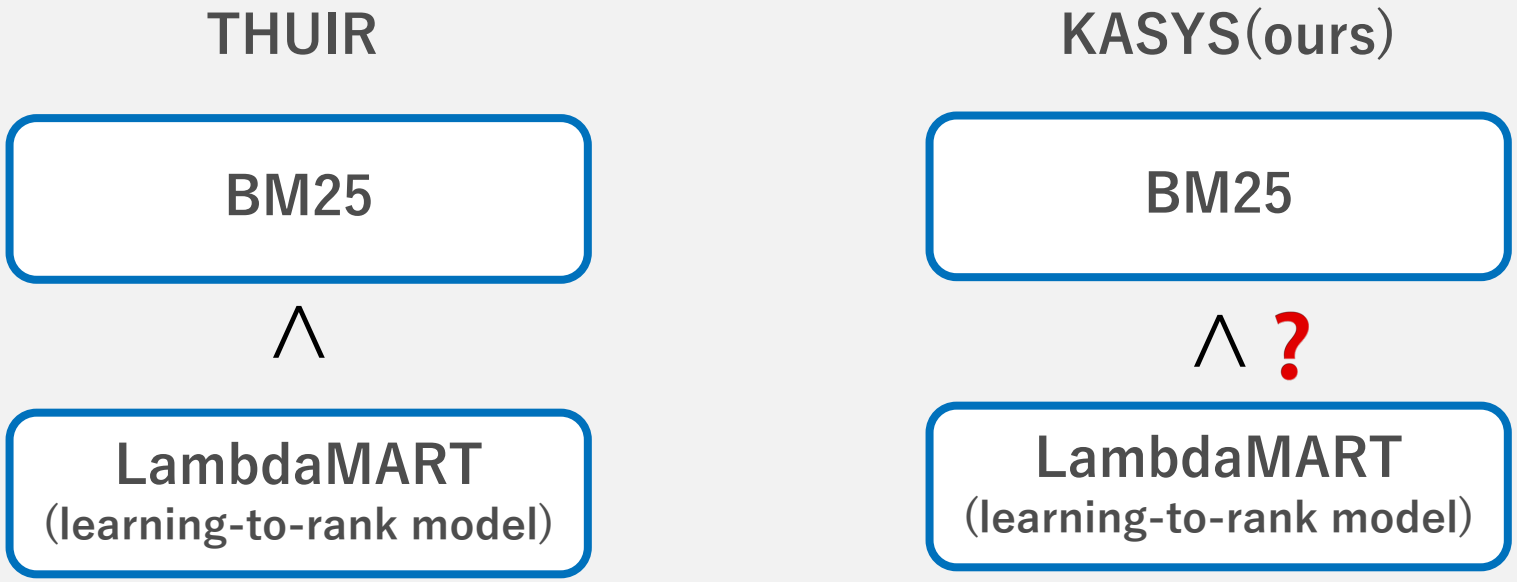
KASYS-E-CO-NEW-4:
- MS MARCO→MB
- Top 2 sentences

KASYS-E-CO-NEW-5:
- CAR→MB
- Top 3 sentences

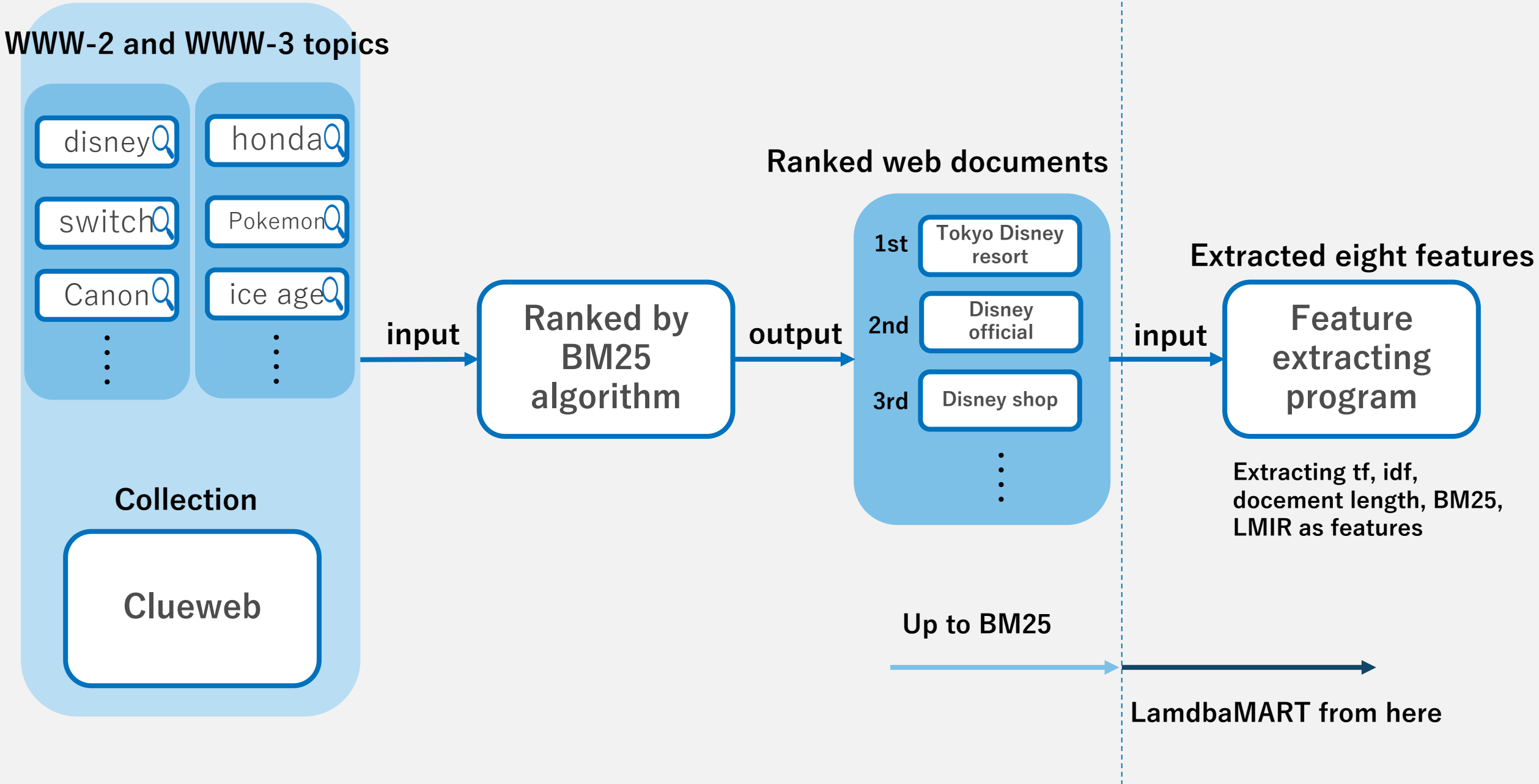
REP Runs

Replicating and reproducing the THUIR runs at the NTCIR 14 WWW-2 Task

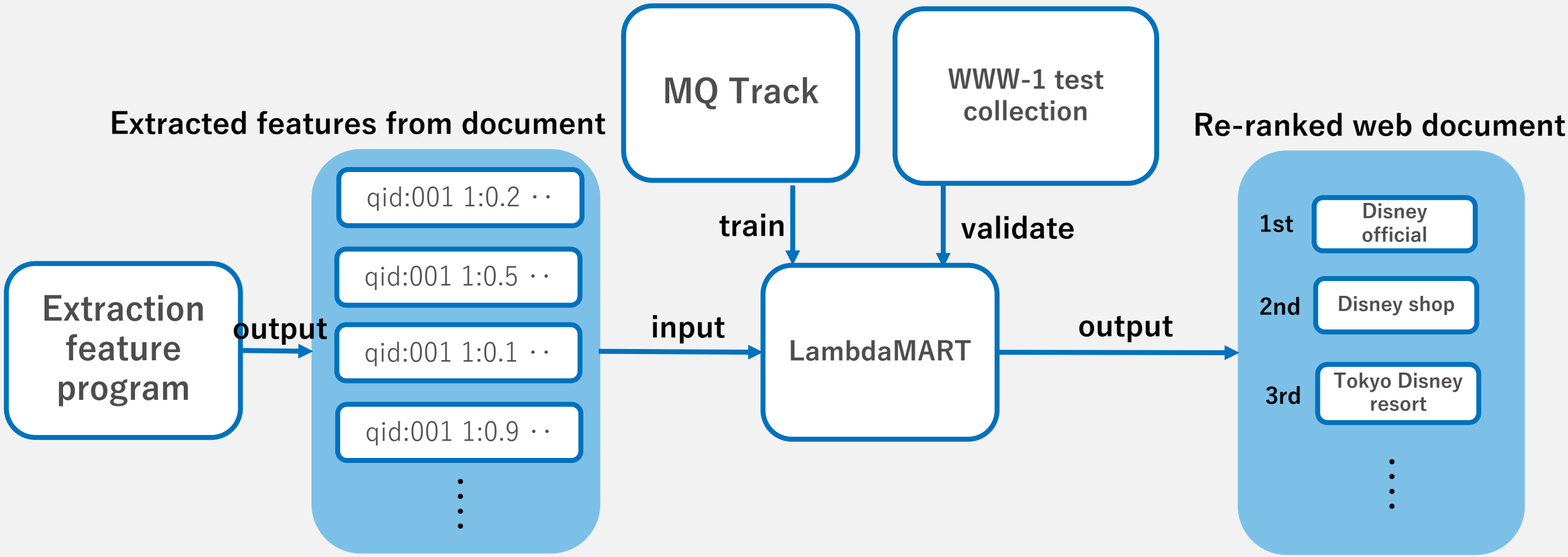
Whether the results between models are consistent with each result.



Replication Procedure 1



Replication Procedure 2

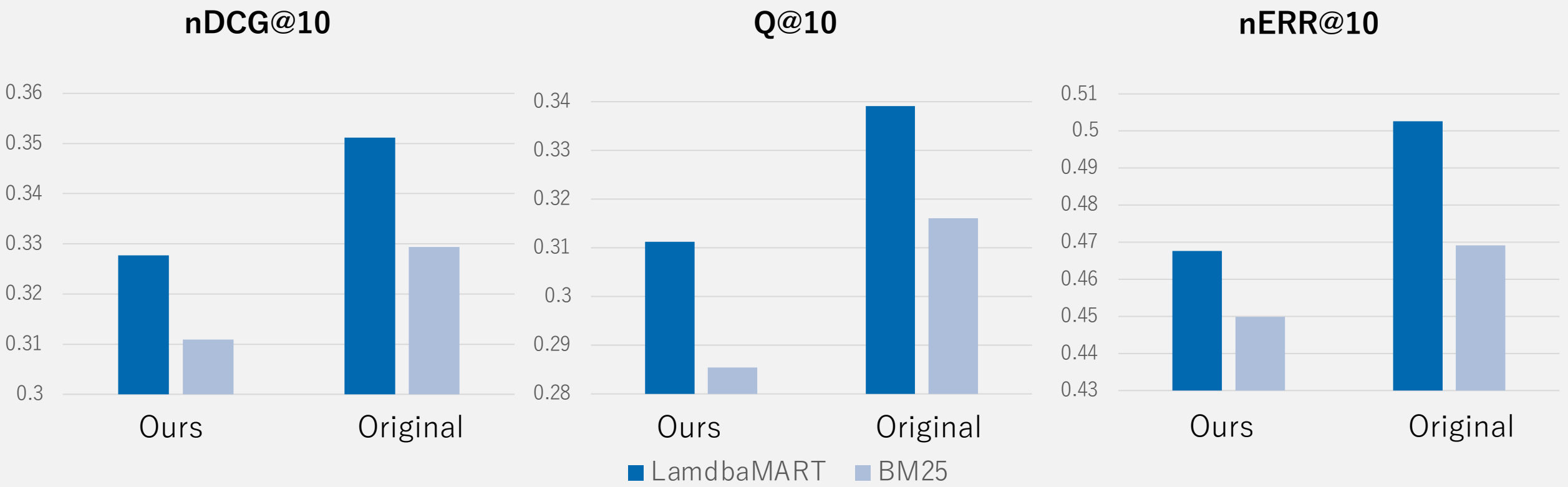


・MQ Track : A dataset of the relevance of a topic and a document.

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

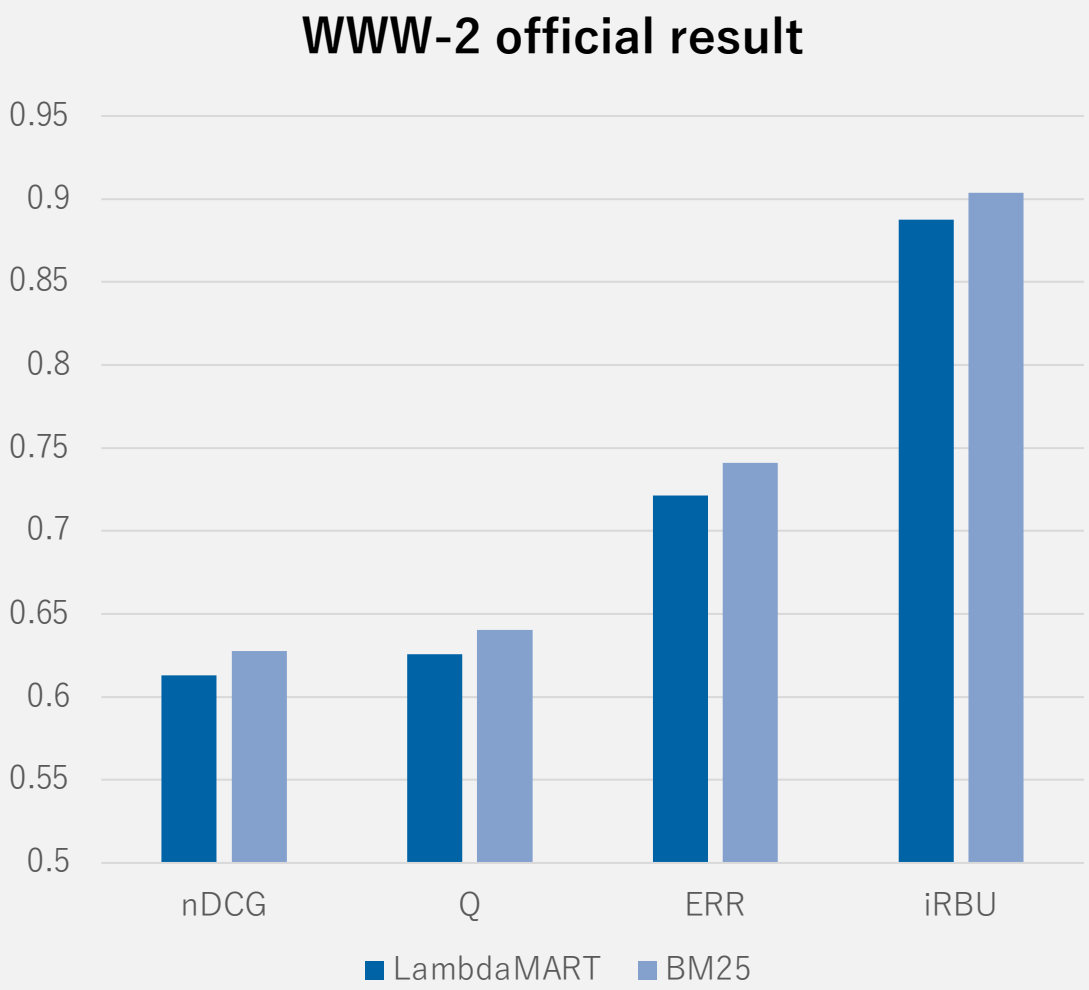
- **The differences from original paper**
  - Although THUIR extracted the features from four fields (whole document, anchor text, title, and URL), we extracted the features from only the whole document
  - Normalization is used by maximum and minimum values because the normalization of features was not described in the original paper

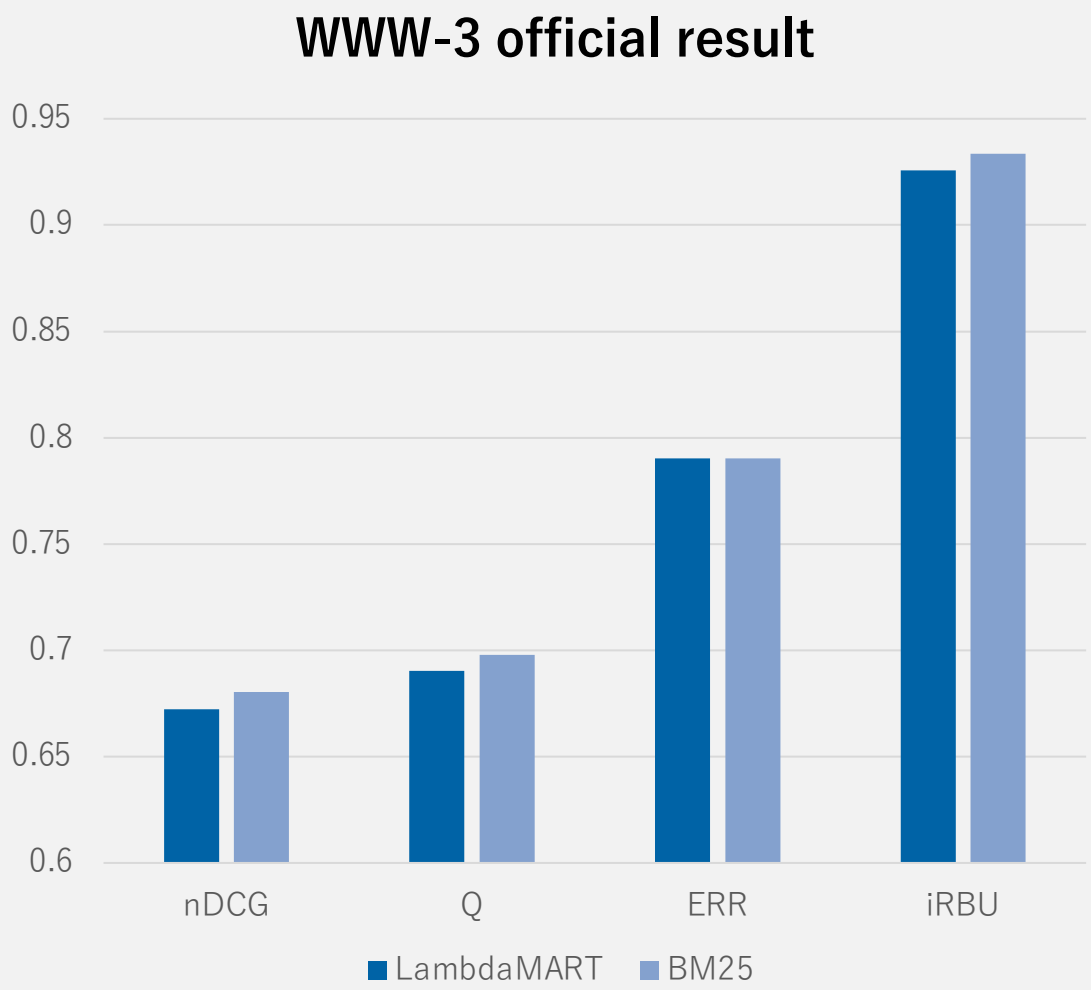
Preliminary Evaluation Results with Original WWW-2 qrels



- **Our results is lower than original results**
- **LambdaMART results were above BM25 for all evaluation metrics**
  - **Succeeded in reproducing the run**

Official Evaluation Results





- **BM25 results were above LambdaMART for all evaluation metrics**
  - **Failed to reproduce the run**

- In the original paper, LambdaMART gave better results than BM25, but on the contrary, our BM25 result was better than LambdaMART
  - We failed to replicate and reproduce the original paper

Suggestions

- In web search tasks, more effective to extract features from all fields
- Better to clarify the method of normalization in a paper

NEW runs

- **Achieved the best performances in terms of nDCG, Q and iRBU among all the participants**
- **The effectiveness of BERT in ad hoc web document retrieval tasks was verified.**
  - MSMARCO→MB is the best. The CAR→MB model also achieved similar scores.
  - BERT is also valid for web document retrieval.

REP runs

- **In the original paper, LambdaMART gave better results than BM25, but on the contrary, our BM25 result was better than LambdaMART**
  - **We failed to replicate and reproduce the original paper**