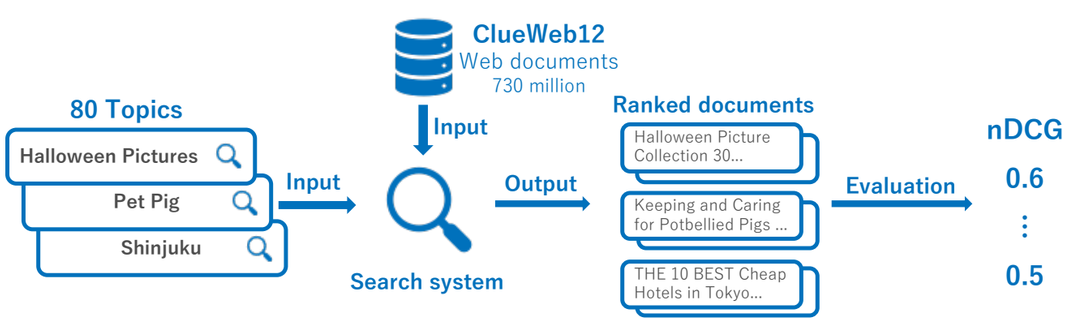
KASYS at the NTCIR-15 WWW-3 Task

Kohei Shinden, Atsuki Maruta and Makoto P. Kato (University of Tsukuba)

KASYS for NEW Runs

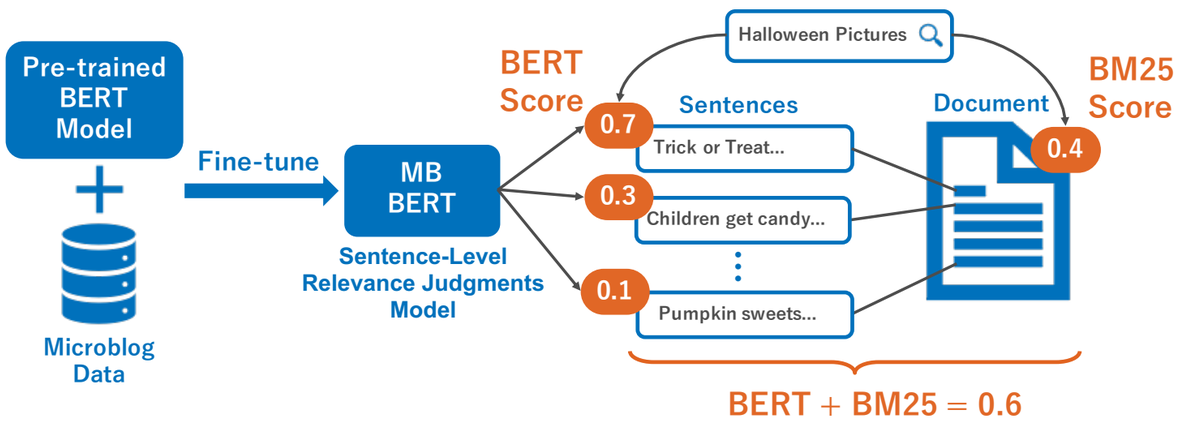
- **NTCIR WWW-3 Task**
  - Ad-hoc search tasks for Web documents



- **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

Birch (Yilmaz et al, 2019)

- **Applying the relevance between sentences learned in QA and Microblog search datasets to Ad-hoc document retrieval**
  - Estimate the relevance of the sentences in the query and document



Key Ideas in Birch

Common Web Documents Do Not Fit BERT

↓

Using Query and Sentence Relevance

- A typical web document exceeds BERT's maximum input sequence length of 512 tokens.

Less Query and Sentence Relevance Data in Ad-hoc Document Retrieval Tasks

↓

Using Models Learned in Other Tasks

Use Test Collection
- MS MARCO: QA Task
- TREC CAR : A is a complex QA task
- TREC MB : Tweet search task

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
    - f BM25 (d): The BM25 score of document d
    - f BERT(pi): The sentence relevance of the top i-th sentence obtained by BERT
    - w_i: The hyper-parameter wi 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

Experimental Results

Run name	Mean nDCG
KASYS-E-CO-NEW-1	0.6935
mpii-E-CO-NEW-1	0.6897
KASYS-E-CO-NEW-4	0.6893
KASYS-E-CO-NEW-5	0.6812
mpii-E-CO-NEW-2	0.6743
Technion-E-CO-NEW-1	0.6581
Technion-E-CO-NEW-2	0.6560
ESTUCeng-E-CO-NEW-3	0.6537
ESTUCeng-E-CO-NEW-1	0.6508
Technion-E-CO-NEW-4	0.6505

Run name	Mean Q
KASYS-E-CO-NEW-1	0.7123
KASYS-E-CO-NEW-4	0.7073
mpii-E-CO-NEW-1	0.7016
KASYS-E-CO-NEW-5	0.6990
mpii-E-CO-NEW-2	0.6905
Technion-E-CO-NEW-1	0.6815
Technion-E-CO-NEW-2	0.6739
Technion-E-CO-NEW-4	0.6717
ESTUCeng-E-CO-NEW-3	0.6644
ESTUCeng-E-CO-NEW-1	0.6638

Run name	Mean ERR
mpii-E-CO-NEW-1	0.8090
KASYS-E-CO-NEW-1	0.7959
KASYS-E-CO-NEW-4	0.7893
Technion-E-CO-NEW-1	0.7791
mpii-E-CO-NEW-2	0.7787
KASYS-E-CO-NEW-5	0.7768
ESTUCeng-E-CO-NEW-1	0.7597
ESTUCeng-E-CO-NEW-3	0.7561
Technion-E-CO-NEW-2	0.7502
Technion-E-CO-NEW-4	0.7471

Run name	Mean iRBU
KASYS-E-CO-NEW-4	0.9389
KASYS-E-CO-NEW-1	0.9382
KASYS-E-CO-NEW-5	0.9298
Technion-E-CO-NEW-1	0.9217
Technion-E-CO-NEW-2	0.9187
mpii-E-CO-NEW-2	0.9175
mpii-E-CO-NEW-1	0.9165
ESTUCeng-E-CO-NEW-1	0.9163
ESTUCeng-E-CO-NEW-3	0.9161
THUIR-E-CO-REV-1	0.9112

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)

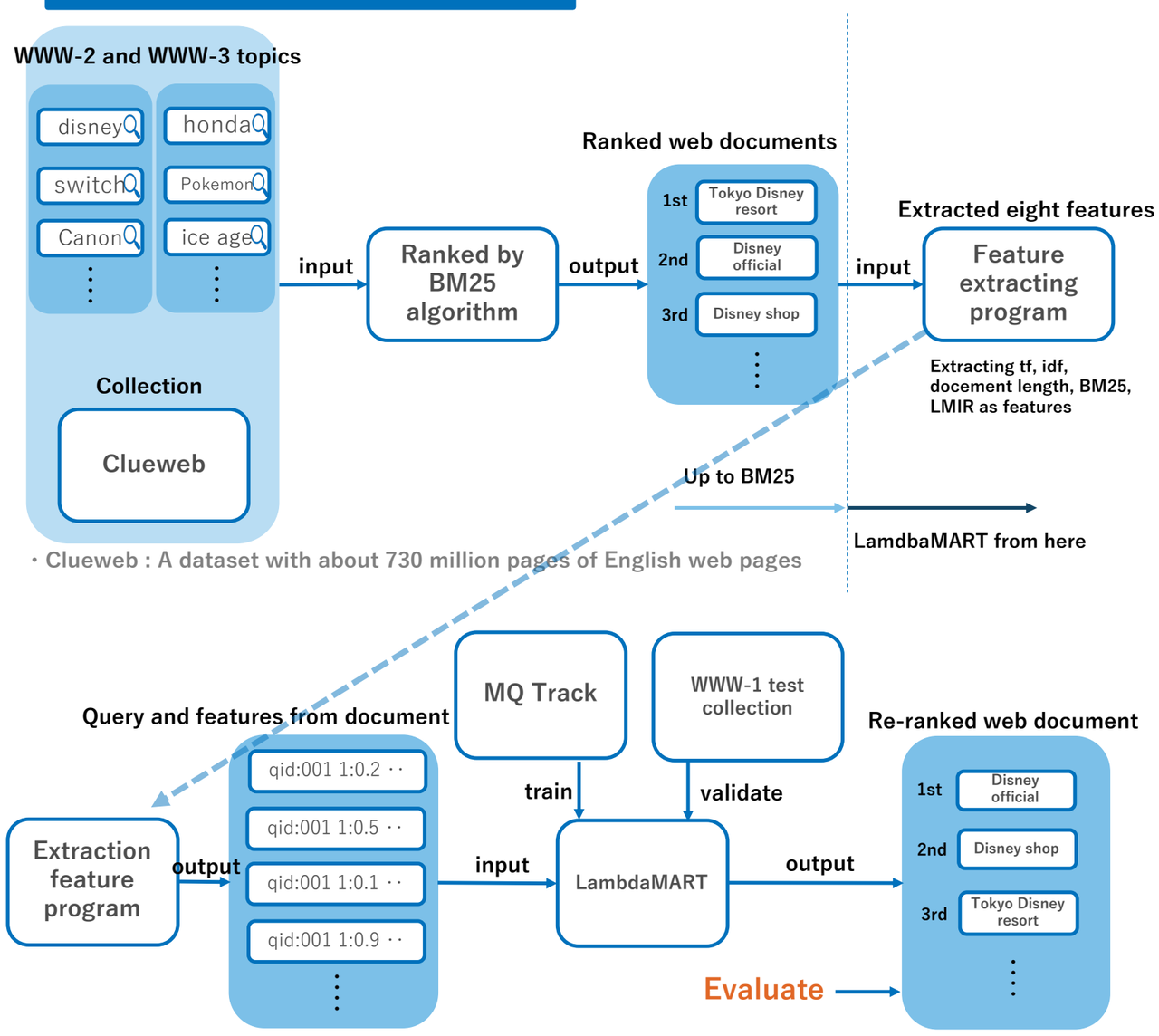
KASYS for REP Runs

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

Whether the results across models are consistent in each experiment

THUIR	KASYS(ours)
BM25	BM25
∧	∧ ?
LambdaMART (learning-to-rank model)	LambdaMART (learning-to-rank-model)

Replication Procedure



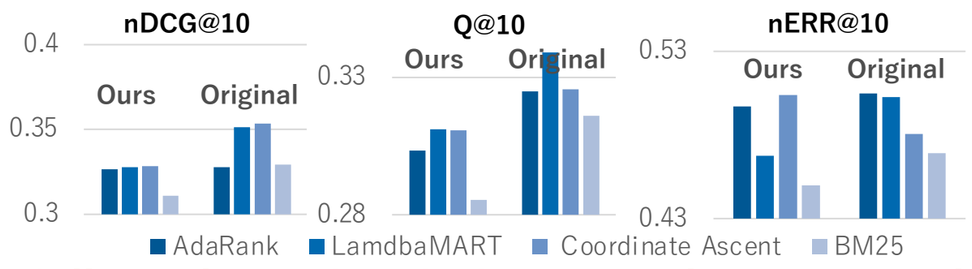
- Clueweb : A dataset with about 730 million pages of English web pages

Implementation Details

- **Features for learning to rank**
  - TF, IDF, TF-IDF, document length, BM25 score, and language-model-based IR scores
- **The differences from the original run (THUIR)**
  - Although THUIR extracted the features from four fields (whole document, anchor text, title, URL), we extracted from only the whole document
  - Normalized by maximum and minimum values because the normalization of features was not described in THUIR paper

Experimental Results

Original WWW-2 qrels



Successfully replicated with the original WWW-2 qrels

: Evaluation results of KASYS's REP runs for WWW-2 topics (Replicability evaluation).

Run	Model	nDCG	Q	ERR	iRBU
KASYS-E-CO-REP-2	Rep A-run (LambdaMART)	0.6742	0.6904	0.7904	0.9256
KASYS-E-CO-REP-3	Rep B-run (BM25)	0.6803	0.6979	0.7904	0.9334

Evaluation results of KASYS's REP runs for WWW-3 topics (Reproducibility evaluation).

Run	Model	nDCG	Q	ERR	iRBU
KASYS-E-CO-REP-2	Rep A-run (LambdaMART)	0.6131	0.6256	0.7213	0.8876
KASYS-E-CO-REP-3	Rep B-run (BM25)	0.6275	0.6402	0.7410	0.9037

WWW-3 official results

Seems that KASYS failed to replicate and reproduce the THUIR run