

NII Table Linker at the NTCIR-15 Data Search Task

Re-ranking with Pre-trained Contextualized Embeddings,
Data Content, Entity-centric, and Cluster-based Approaches

Phuc Nguyen, Kazutoshi Shinoda, Taku Sakamoto, Diana Andreea Petrescu, Hung Nghiep Tran,
Atsuhiko Takasu, Akiko Aizawa, Hideaki Takeda

Outline (8 mins)

1. Introduction (1 mins)
2. Challenges (1 mins)
3. NII Table Linker (5 mins)
4. Conclusion (1 min)

1. Introduction (1 mins)

- NTCIR 15 Data Search Task

- **Task:** Ad-hoc retrieval for governmental statistical data
- **Input:** keywords
- **Output:** List of dataset ranked by a relevance score
- **Data stats:**

Statistic	Japanese	English
Number of datasets	1,338,402	46,615
Metadata	4.1 GB	0.2 GB
Data files	1,338,402 (434 GB)	90,388 (714 GB)
Total dataset	1,2 T	

2. Challenges (1 mins)

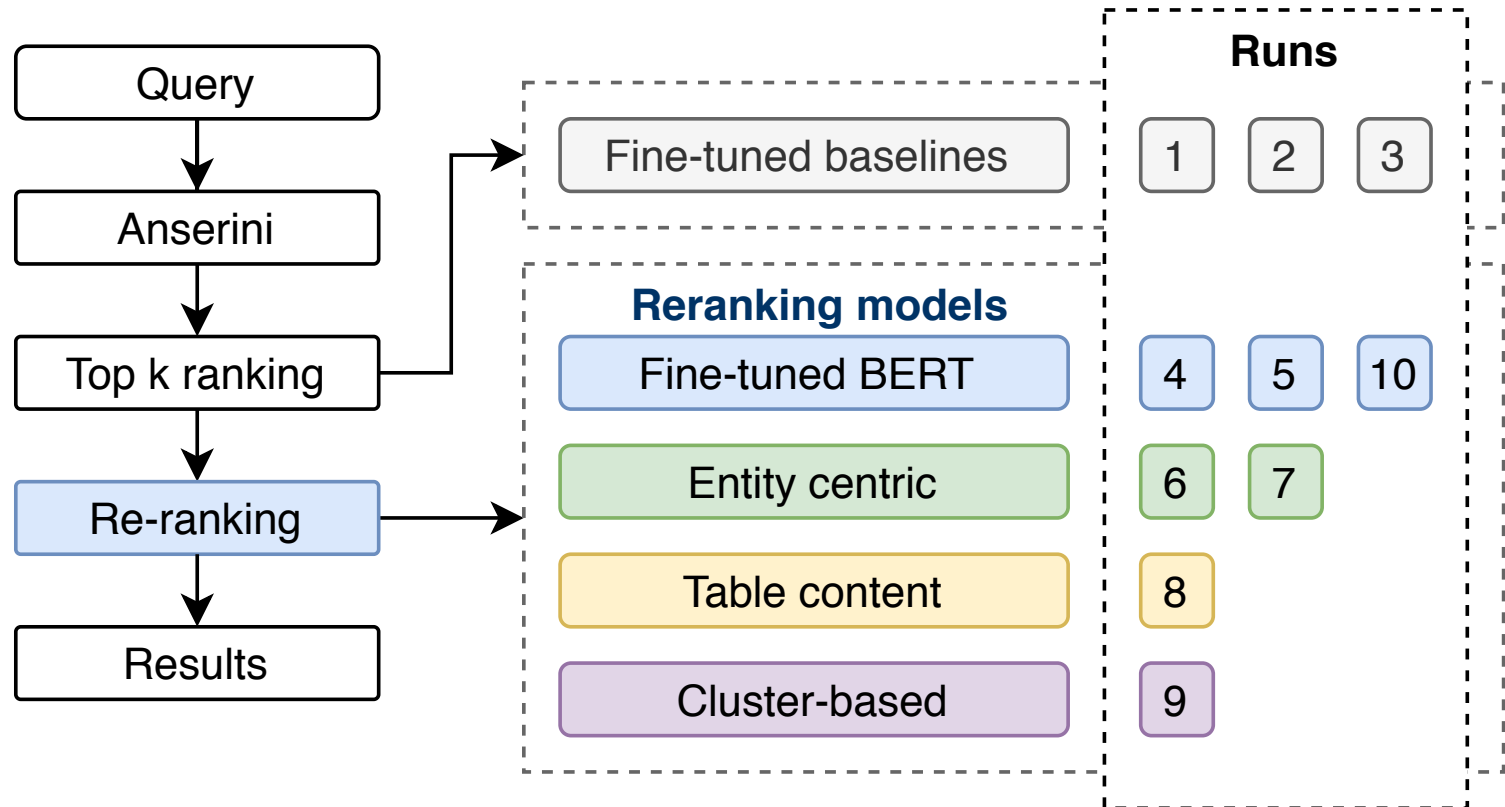
Data search challenges

- **Data:** Large number of datasets, and fast growth
- **Data context:** quality, trustability
- **Reusable:** machine non-readable
- **Data evolving:** data versioning, changing of publishers
- **Evaluation:** Bias relevant judgment

3. NII Table Linker (5 mins)

- Overall system
- Introduce each runs and results

NII Table Linker System



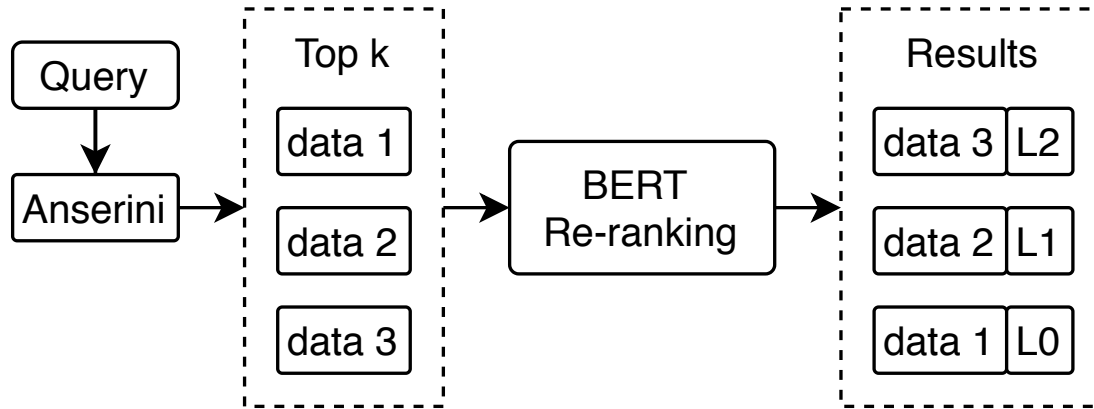
Fine-tuned baselines

- Run 1: Fine-tuned BM25
- Run 2: BM25+PRF
- Run 3: Fine-tuned BM25 + PRF

Avg metric	English		Japanese	
	Default	Fine-tuned	Default	Fine-tuned
BM25	0.238	E1: 0.231 -2.94%	0.415	J1: 0.415 +0.05%
BM25+PRF	0.219	E3: 0.203 -7.31%	0.405	J3: 0.412 +1.85%

- Fine-tuning on the standard methods:
- + English: bad
 - + Japanese: Good (J1: 1st in average score)

Sequential Re-ranking with BERT (1)

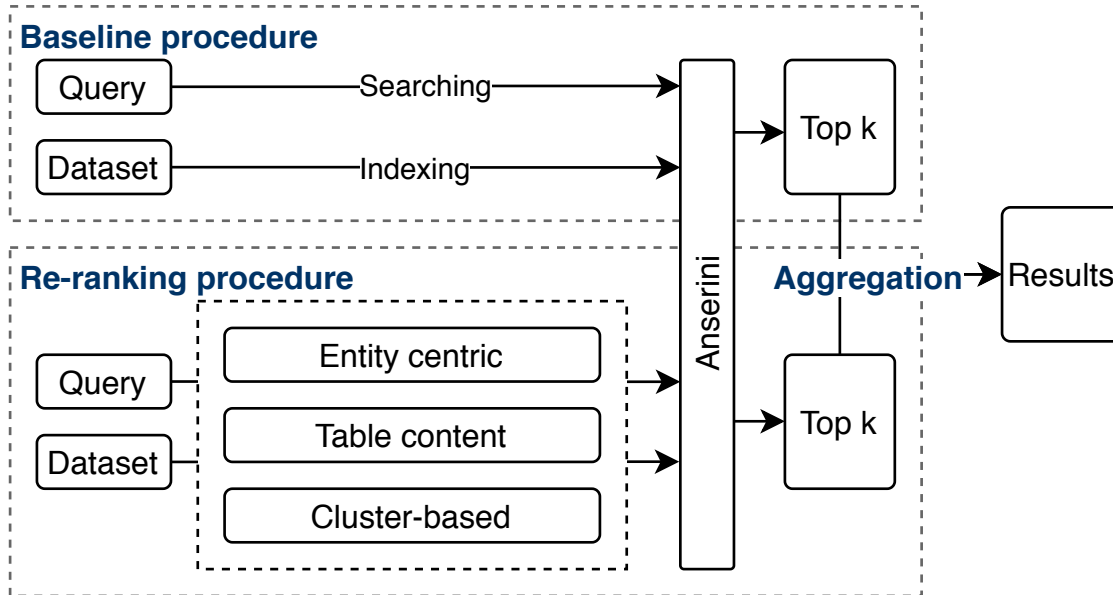


- Run 4: Top 20 of BM25+PMF (default)
- Run 5: Top 20 of BM25+PRF (fine-tuned)
- Run 10: Top 100 of BM25+PRF (fine-tuned)

Avg metric	English	Japanese
BM25+PRF (Top k)	E2: 0.222	J3: 0.435
Top k + BERT	E4: 0.252 +13.53%	J5: 0.430 -1.12%

→ English: Good (E-4 : 1st in average score)
Japanese: Bad

Parallel Re-ranking



- Run 6: Entity extraction
- Run 7: Date and Location
- Run 8: Table Content
- Run 9: Cluster based approach

Avg metrics	English	Japanese
BM25+PRF	E3: 0.209	J3: 0.435
Entity Centric	E6: 0.178 -14.6%	J6: 0.343 -21.13%
Date Location	E7: 0.192 -8.3%	-
Table Content	E8: 0.189 -9.8%	J8: 0.380 -12.6%
Cluster-based	E9: 0.221 +6%	J9: 0.421 -3.9%

→ The parallel re-ranking do not perform well

Extra-runs: Embedding Search

- Only perform on English dataset
- BM25+PMF (Top 10) with fasttext
- Two runs (2, 6) with different weighting

Avg metric	English
BM25+PRF	E2: 0.217
Embedding Search 1	E-Ex-2: 0.248 +14.3%
Embedding Search 2	E-Ex-6: 0.235 +8.3%

4. Conclusion (1 min)

- Contributions
- Future work

- **NII Table Linker Contributions:**

- Preliminary work on DST
- Study capability of standard methods
- The four re-ranking approaches
 - + pre-trained BERT
 - + Entity Centric
 - + Table Content
 - + Cluster-based

→English: Best performance using pre-trained BERT

→Japanese: Best performance using fine-tuned BM25

- **Future work**

- Data standardization
- Data semantic annotation
- Data content retrieval, fact checking

- **Discussion:**

Please go to our poster at 18:00-19:30 tonight.

Ranking (Top 10)

in the average of all metrics

OFFICIAL EVALUATION

English

NII-E-4	0.233	0.237	0.248	0.251	0.251	0.264	0.278
KSU-E-2	0.204	0.231	0.255	0.238	0.229	0.257	0.276
KSU-E-6	0.204	0.231	0.255	0.238	0.229	0.257	0.276
ORGE-E-8	0.224	0.230	0.233	0.238	0.244	0.255	0.264
STIS-E-2	0.230	0.228	0.237	0.217	0.248	0.255	0.264
ORGE-E-2	0.219	0.225	0.238	0.240	0.235	0.250	0.264
uhai-E-5	0.219	0.225	0.238	0.240	0.235	0.250	0.264
NII-E-10	0.221	0.226	0.237	0.238	0.235	0.248	0.264
NII-E-5	0.214	0.227	0.230	0.234	0.230	0.247	0.258
ORGE-E-7	0.216	0.220	0.236	0.237	0.228	0.242	0.256

Japanese

NII-J-10	0.395	0.409	0.416	0.445	0.471	0.482	0.464
NII-J-1	0.400	0.405	0.415	0.446	0.467	0.483	0.478
NII-J-3	0.396	0.407	0.412	0.445	0.468	0.480	0.463
NII-J-5	0.389	0.403	0.411	0.438	0.463	0.475	0.463
NII-J-2	0.382	0.396	0.405	0.426	0.452	0.464	0.445
NII-J-9	0.376	0.385	0.400	0.433	0.458	0.473	0.444
NII-J-4	0.378	0.393	0.404	0.424	0.450	0.463	0.445
ORGJ-J-3	0.407	0.413	0.421	0.325	0.450	0.470	0.484
ORGJ-J-2	0.402	0.405	0.415	0.328	0.447	0.467	0.483
uhai-J-10	0.403	0.406	0.415	0.312	0.447	0.466	0.484

EXTRA EVALUATION

English

NII-E-EX-2	0.227	0.236	0.230	0.255	0.271	0.277	0.239
NII-E-4	0.227	0.229	0.239	0.246	0.259	0.274	0.238
STIS-E-2	0.230	0.227	0.235	0.248	0.254	0.264	0.210
ORGE-E-8	0.219	0.224	0.226	0.242	0.252	0.261	0.227
NII-E-EX-6	0.220	0.216	0.222	0.243	0.250	0.263	0.232
uhai-E-5	0.215	0.220	0.232	0.233	0.247	0.261	0.230
ORGE-E-2	0.215	0.220	0.232	0.233	0.247	0.261	0.230
NII-E-10	0.219	0.222	0.228	0.234	0.246	0.260	0.226
KSU-E-2	0.198	0.223	0.240	0.226	0.253	0.270	0.223
KSU-E-6	0.198	0.223	0.240	0.226	0.253	0.270	0.223

Japanese

NII-J-EX-10	0.393	0.405	0.406	0.450	0.474	0.486	0.296
KSU-J-EX-1	0.389	0.396	0.426	0.443	0.466	0.480	0.298
KSU-J-EX-6	0.389	0.396	0.426	0.443	0.466	0.480	0.298
uhai-J-10	0.389	0.391	0.392	0.437	0.458	0.475	0.286
ORGJ-J-3	0.388	0.393	0.391	0.437	0.458	0.472	0.293
KSU-J-5	0.368	0.381	0.413	0.431	0.461	0.482	0.252
ORGJ-J-2	0.383	0.385	0.386	0.434	0.455	0.471	0.300
NII-J-EX-1	0.380	0.385	0.385	0.433	0.454	0.471	0.302
NII-J-EX-3	0.377	0.385	0.382	0.432	0.455	0.467	0.288
NII-J-EX-5	0.373	0.384	0.385	0.429	0.454	0.465	0.288