

NII Table Linker at the NTCIR-15 Data Search Task

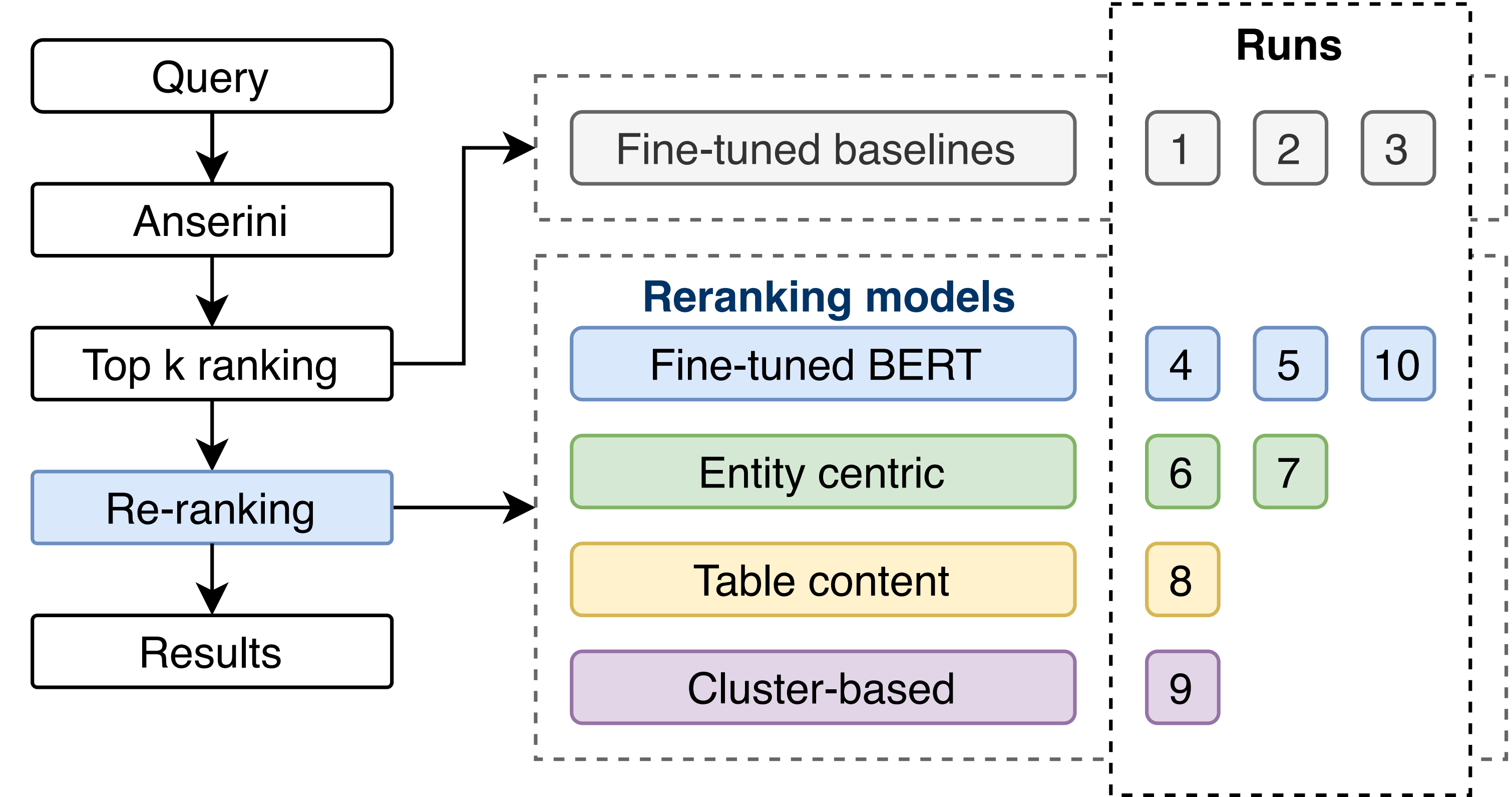


Re-ranking with Pre-trained Contextualized Embeddings, Data Content, Entity-centric, and Cluster-based Approaches

Phuc Nguyen¹, Kazutoshi Shinoda², Taku Sakamoto², Diana Andreea Petrescu³, Hung Nghiep Tran¹, Atsuhiko Takasu¹, Akiko Aizawa^{1,2}, Hideaki Takeda¹
¹ National Institute of Informatics (NII), Japan
² The University of Tokyo, Japan
³ The École Polytechnique Fédérale de Lausanne, Switzerland

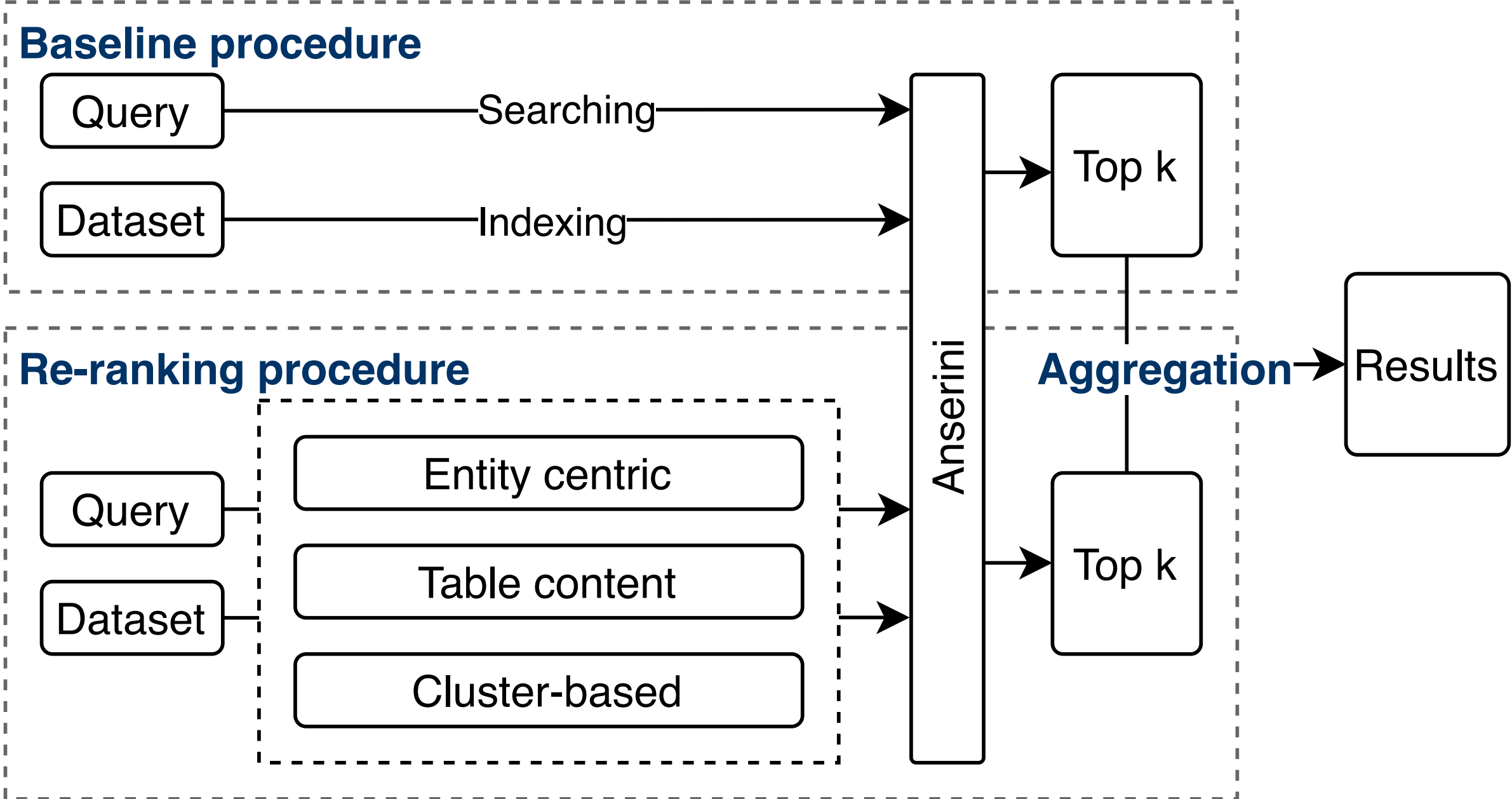
1. Introduction

- Tasks:
Data Search (English and Japanese sub-tasks)
- Overview NII Table Linker:



2. NII Table Linker

- Fine-tuned with Standard Methods:
 - Run 1: Fine-tuned BM25 English: NII-E-1, Japanese: NII-J-1
 - Run 2: BM25+PMF English: NII-E-2, Japanese: NII-J-2
 - Run 3: Fine-tuned BM25 + PMF English: NII-E-3, Japanese: NII-J-3
- Re-ranking with a Sequential Procedure:
 - Re-raking with BERT
 - Run 4:
Top 20 on Run 2
English: NII-E-4
Japanese: NII-J-4
 - Run 5:
Top 20 on Run 3
English: NII-E-5
Japanese: NII-J-5
 - Run 10:
Top 100 on Run 3
English: NII-E-10
Japanese: NII-J-10
- Re-ranking with a Parallel Procedure:
 - Run 6: Entity Centric English: NII-E-6, Japanese: NII-J-6
 - Run 7: Date and Location English: NII-E-7
 - Run 8: Table Content English: NII-E-8, Japanese: NII-J-7, NII-J-8
 - Run 9: Cluster based approach English: NII-E-9, Japanese: NII-J-9



- Embedding Search: (Extra runs)
 - Run-Ex-2: Run 2 + fasttext, English: NII-E-Ex-2
 - Run-Ex-6: Run 2 + fasttext, English: NII-E-Ex-6

4. Results

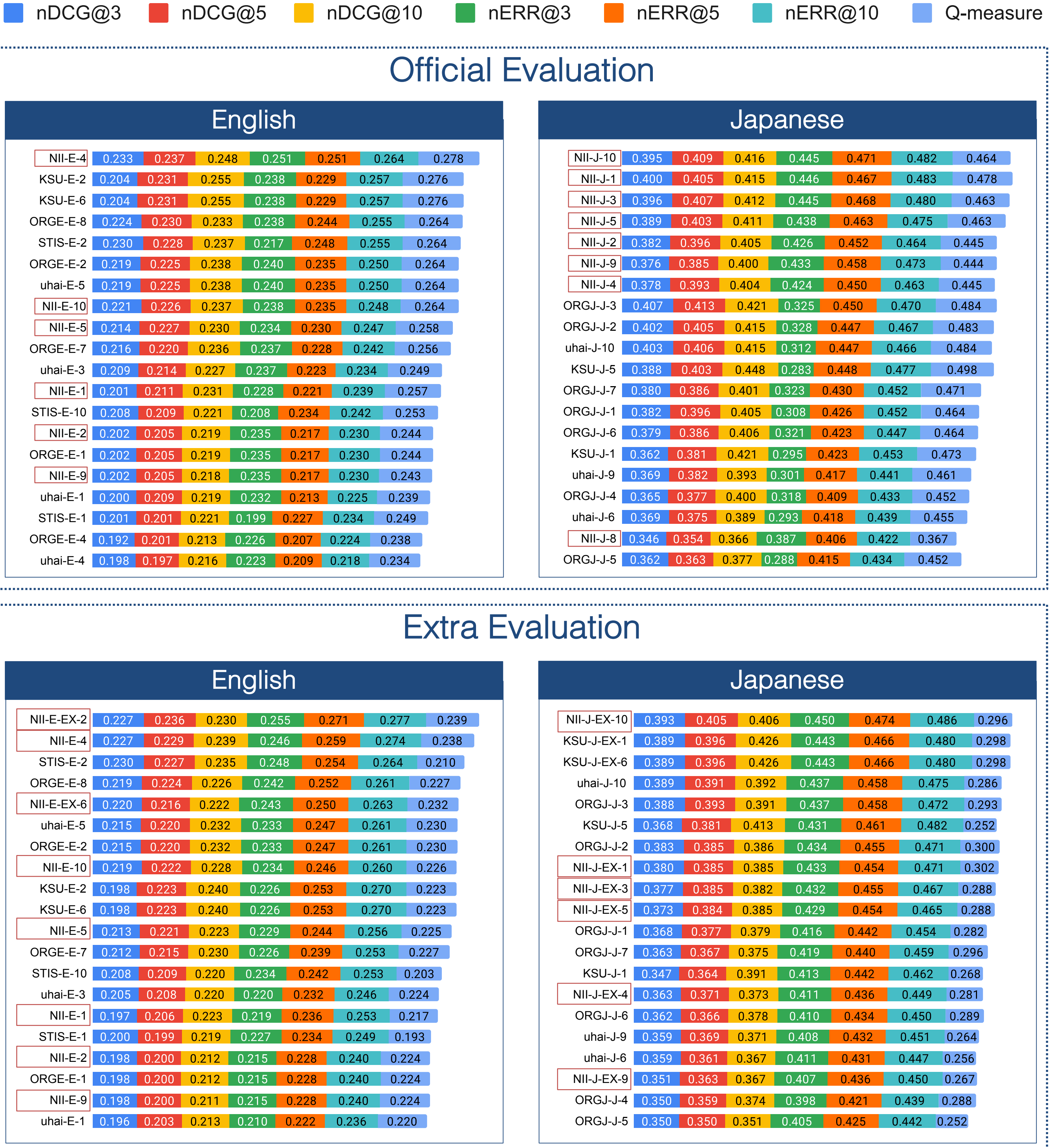
- Fine-tuning on the standard methods: bad (English), good (Japanese)

Avg score	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%
- Re-ranking (sequential) with BERT: good (English), bad (Japanese)

Avg score	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%
- Re-ranking (parallel) with other factors do not work well

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%
- Extra-runs: Re-ranking with embedding search is promising
Run Ex-E-2: +14.3% vs E-2 (BM25+PRF Top 10)

3. Results (Ranking of Top 20 runs)



5. Future work

- Dataset standardization, semantic annotation, knowledge graph matching
- Data content retrieval, fact checking

6. Acknowledgements

This work was supported by the Cross-ministerial Strategic Innovation Promotion Program (SIP) Second Phase, "Big-data and AI- enabled Cyberspace Technologies" by the New Energy and Industrial Technology Development Organization (NEDO).