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,
Atsuhiro Takasu, Akiko Aizawa, Hidetoshi Takeda

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:

  - Query
  - Anserini
  - Top k ranking
  - Re-ranking
  - Results

  **Fine-tuning with Standard Methods:**
  - Run 1: Fine-tuned BM25
  - Run 2: BM25 + PMF
  - Run 3: Fine-tuned BM25 + PMF

  **Re-ranking with a Sequential Procedure:**
  - 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
  - Run 7: Date and Location
  - Run 8: Table Content
  - Run 9: Cluster based approach

- 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

**2. NII Table Linker**

- Fine-tuning with Standard Methods:
  - English: NII-E-1, Japanese: NII-J-1

- Re-ranking 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 Sequential Procedure:
  - Top k ranking with a Parallel Procedure:
    - English: NII-E-6, Japanese: NII-J-6
    - English: NII-E-7
    - English: NII-E-9, Japanese: NII-J-9

**3. Results (Ranking of Top 20 runs)**

- Official Evaluation
  - English
  - Japanese
  - Extra Evaluation

**4. Results**

- Fine-tuning on the standard methods: bad (English), good (Japanese)
  - BM25:
    - Default: 0.238
    - Fine-tuned: 0.219
  - BM25 + PRF:
    - Default: 0.203 - 0.94%
    - Fine-tuned: 0.415
  - J6: 0.415 + 0.05%

- Re-ranking (sequential) with BERT: good (English), bad (Japanese)
  - BM25+PRF (Top k):
    - Default: 0.222
    - Fine-tuned: 0.262 + 13.53%
  - J6: 0.435
  - J6: 0.435 - 1.12%

- Re-ranking (parallel) with other factors do not work well
  - Avg metrics:
    - English
    - Japanese
    - BM25 + PRF:
      - English: E3: 0.209
      - Japanese: J3: 0.435
  - Re-ranking with embedding search is promising
  - Run Ex-2: +14.3% vs E-2 (BM25+PRF Top 10)

**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).