SMM at the NTCIR-18 U4 Task

Yukihiro Seito
Keio University

Team: SMM **Tasks:** Table Retrieval (Japanese), Table Question Answering (Japanese)

1. Abstract

- **Hybrid Table Retrieval (TR):**
  - Single-Vector (Global Semantics)
  - Cell-Level Multi-Vector (Local Details)
- **Effective Table Question Answering (TQA) with Cell ID Estimator:**
  - T5 for candidate generation
  - Cell ID Estimator for evidence-based selection
- **High Performance:**
  - TR Acc: ∼97.6%
  - TQA Value Acc: ∼86% (Pub) / ∼82% (Priv)

Keywords: Information Retrieval, Question Answering, Tabular Data, Financial Reports

2. Task Overview: NTCIR-18 U4

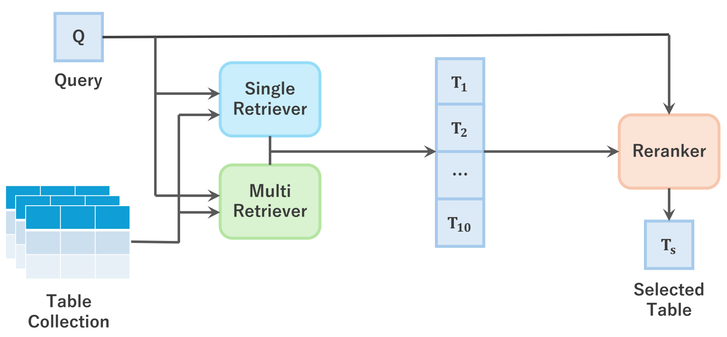
Challenge: Automated information extraction from Japanese annual securities reports.

- **Table Retrieval (TR):** Query → Most relevant table.
- **Table Question Answering (TQA):** Query + Table → Precise answer (value & cell ID).

Our Goal: High-accuracy system considering table structure & cell-level info.

3. Table Retrieval (TR) Method

Pipeline: Query → Hybrid Retrievers → Top 10 → Reranker → Final Table



TR Pipeline Diagram

Hybrid Retrieval Strategy:

(1) Single-Vector (Global Semantics)

- **Model:** 'pkshatech/GLuCoSE-base-ja'
- **Representation:** Avg. BERT hidden states of concatenated cell text.

$$\mathbf{v}_{\text{table}} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i$$

- **Similarity:** Cosine Similarity.

$$S_{\text{single}} = \text{sim}(\mathbf{q}, \mathbf{v}_{\text{table}})$$

- **Training: Iterative Negative Re-mining [3]**
  - Initial: BM25 Hard Negatives.
  - Iterative: Add top-k False Positives (from self & Multi-Vector) as new Hard Negatives.

(2) Cell-Multi-Vector (Local Details)

- **Model:** 'pkshatech/GLuCoSE-base-ja'
- **Representation:** Each cell as independent vector.
- **Similarity:** Query-cell similarities + LogSumExp Aggregation (τ=20).

$$S_{\text{multi}} = \frac{1}{\tau}\log\sum_{j}\exp(\tau \cdot s_j)$$

- **Training: In-passage Negatives [2]**
  - Negatives: Low BM25 score cells from the same table.
  - Enhances distinguishing contextually similar but irrelevant cells.

Score Integration

- **Combined Score:** Weighted sum of S_{single} and S_{multi}.
- Weight $\alpha = 0.6$ (optimized on validation set).

$$S_{\text{final}} = \alpha S_{\text{single}} + (1-\alpha)S_{\text{multi}}$$

Key Preprocessing Steps

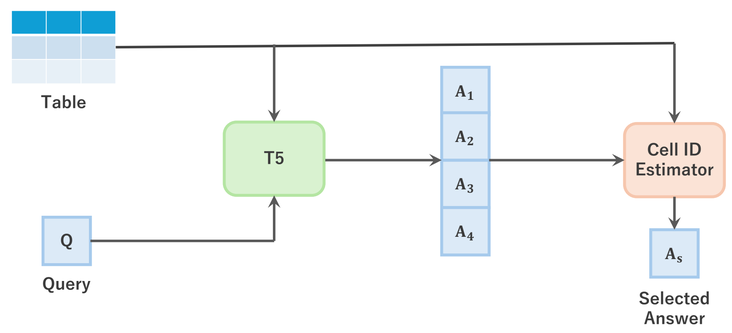
- HTML Table Extraction & Cleaning
- Text Normalization (Unicode, Romaji/Katakana)

TR: Cross-Encoder Reranker

- **Model:** 'hotchpotch/japanese-reranker-cross-encoder-base-v1'
- **Objective:** Refine Top-10 candidates from TR hybrid retrieval.
- **Input:** (Query, Table Content) pairs.
- **Output:** 3-class (Pos/Neg/HardNeg) Softmax score.
- **Training:** Hard Negatives from high-scoring incorrect examples from hybrid retrieval. Cross-Entropy Loss.

4. Table Question Answering (TQA) Method

Pipeline: Query + Table → T5 Candidates → Cell ID Estimator → Final Answer



TQA Pipeline Diagram

Two-Stage Approach:

(1) T5 for Candidate Generation

- **Model:** 'retrieva-jp/t5-base-long'
- **Input Format for T5:** Linearized table with structural markers.
  `|Query|Table Title|cell1_row1,cell2_row1,...|cell1_row2,cell2_row2,...|...`
  `(and so on for all relevant rows)|`
- **Output:** Multiple diverse answer candidates (Beam Search).

(2) Cell ID Estimator

- **Goal:** Select most plausible T5 answer by identifying its best supporting evidence cell.
- **Core Context Scoring (for text T and cell C_{target}'s context type $X \in \{\text{Row, Col}\}$):** Let $\mathcal{C}_X$ be the set of cells in context X (row or column) of C_{target}. Weight $\alpha = 0.6$.
  - $S_{\text{single}}(T, \mathcal{C}_X)$: Sim. between T and concatenated text of cells in $\mathcal{C}_X$. (Cosine Sim.)
  - $S_{\text{multi}}(T, \mathcal{C}_X)$: Aggregated sim. between T and each cell in $\mathcal{C}_X$. (LSE Agg.)
  - $S_{\text{context}}(T, C_{\text{target}}, X) = \alpha \cdot S_{\text{single}}(T, \mathcal{C}_X) + (1-\alpha) \cdot S_{\text{multi}}(T, \mathcal{C}_X)$
- **Two-Stage Selection Process:** Let $S_{\text{cell-score}}(T, C_j) = S_{\text{context}}(T, C_j, \text{Row}) + S_{\text{context}}(T, C_j, \text{Col})$.
  1. **Candidate-Cell Matching (for each T5 answer candidate A_i):**
     - For every cell C_j: Calculate $S_{\text{cell-score}}(A_i, C_j)$.
     - Identify best matching cell C_i^* for A_i (cell with max $S_{\text{cell-score}}(A_i, C_j)$).
  2. **Query-Evidence Scoring (for each pair (A_i, C_i^*) from Stage 1):**
     - Calculate final evidence score $S_{\text{evidence}}(A_i, C_i^*) = S_{\text{cell-score}}(\text{Query}, C_i^*)$.
- **Final Selection:** Choose A_i (and its cell C_i^*) with max S_{evidence}.

5. Key Results (NTCIR-18 U4)

Task	Metric	Public	Private	Key Findings
TR	Accuracy	97.70%	97.55%	Very high accuracy achieved.
	Accuracy@K	@1: ∼95%, @3: ∼99%		Strong performance (**Retriever stage**).
TQA	Value Acc.	86.57%	81.94%	Highly competitive.
	Cell ID Acc.	86.34%	82.76%	Accurate evidence identification.

Ablation studies confirmed component effectiveness: TR Reranker (+2.8pt), Cell-Multi-Vector (+7pt); TQA Cell ID Estimator (+4pt Value Acc).

6. Conclusion

Achieved:

- High-performance TR & TQA for financial tables.
- Hybrid TR and Cell ID Estimator for TQA proved effective.

References

[1] Y. Kimura, et al. 2025. Overview of NTCIR-18 U4. *Proc. NTCIR-18.*
[2] J. Lee, et al. 2021. Phrase Retrieval Learns Passage Retrieval. *EMNLP.*
[3] T. Shen, et al. 2023. UnifieR: Unified Retriever. *KDD.*