KSU Systems at the NTCIR-16 Data Search 2 IR Subtask

Taku Okamoto, ○**Tomokazu Hayashi** and Hisashi Miyamori

Kyoto Sangyo University

NTCIR-16 Day 4: June17th 10:30~11:30（JST）

Outline

1. INTRODUCTION
2. METHOD
3. RESULTS & DISCUSSION
4. CONCLUSION

INTRODUCTION

- Ad-hoc search for statistical documents are important.
  - Previous studies have mainly used only metadata of statistical documents for ranking
  - However, it has not shown equal or better performance than traditional ad hoc search ranking for text documents.
- The ranking methods using only metadata might be reaching their limits.
  - The contents of the body of the statistical data tables are rarely used.
- Re-ranking method is proposed which employs the table body features of statistical data and neural network models.
  - Verify how much the ranking results improve

METHOD

- Baseline: Data augmentaion(*), BM25
- Category search(*)
- Re-ranking by features obtained from table body and by BERT and MLP

(*) Methods proposed in the previous Data Search Task

TABLE BODY FEATURES OF STATISTICAL DOCUMENTS [1]

Table 1. Features obtained from table body only ($\boldsymbol{F_B}$)

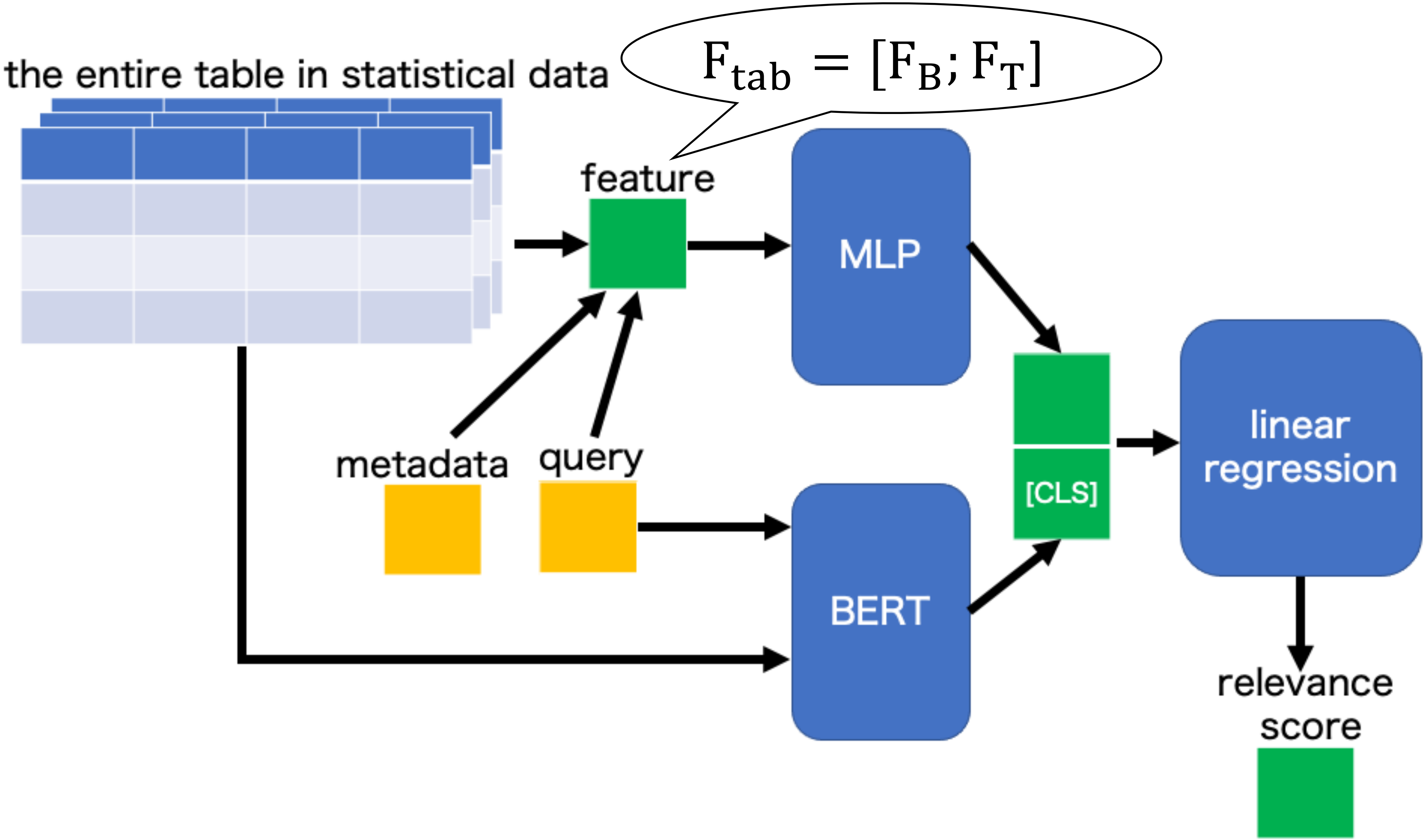
Table feature items	Table feature description
#Rows	Average number of rows in each table body in statistical data
#Cols	Average number of columns in each table body in statistical data
#EmptyCells	Average number of empty cells appearing in each table body in statistical data
#InLinks	The total number of predicates for the corresponding object in DBpedia when each word in the main body of the table is regarded as an RDF object

Table 2. Features obtained from entire table ($\boldsymbol{F_T}$)

Table feature items	Table feature description
hitsLC	Frequency of occurrence of query tokens in the leftmost column of the entire table
hitsSLC	Frequency of occurrence of query tokens in the leftmost two columns of the entire table of statistical data
hitsB	Frequency of occurrence of query tokens in the entire table of statistical data
qInPgTitle	Ratio of the number of query tokens in the title to the total number of tokens in the title of the metadata

[1] Shuo Zhang and Krisztian Balog. 2018. Ad Hoc Table Retrieval using Semantic Similarity. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 1553–1562. https://doi.org/10.1145/3178876.3186067

RE-RANKING USING BERT AND MLP (BERT + MLP)



RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

DISCUSSION

- The reranking method combining table body features and BERT-MLP model did not contribute to improve ranking results.
  - This may be due to the fact that the **information in the table body, which is mostly numerical, was converted to a feature vector using BERT.**
  - **The body of the table used in the previous work contains ordinary non-numeric word sequences** as values.
- Table clipping had no effect.
  - An examination of the feature vectors input to the MLP revealed that each feature had large outlier-like values, and scaling by MaxAbsScaler resulted in most feature values being close to zero.
  - As a result, the table's original features could not be fully utilized.
- Category search was effective in improving rankings.
  - The same results were obtained in the previous Data Search Task.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
$\mathrm{F}_{\mathrm{tab}}$+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+F_{tab}+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+F_{tab}+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+F_{tab}+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
F_{tab}+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

DISCUSSION

- The reranking method combining table body features and BERT-MLP model did not contribute to improve ranking results.
  - This may be due to the fact that the information in the table body, which is mostly numerical, was converted to a feature vector using BERT.
  - The body of the table used in the previous work contains ordinary non-numeric word sequences as values.
- Table clipping had no effect.
  - An examination of the feature vectors input to the MLP revealed that each feature had large outlier-like values, and scaling by MaxAbsScaler resulted in most feature values being close to zero.
  - As a result, the table's original features could not be fully utilized.
- Category search was effective in improving rankings.
  - The same results were obtained in the previous Data Search Task.

RESULTS

Table 3. Ranking evaluation result (nDCG@10)

Method	Japanese		English	
BM25 (optimized)	ORGJ-J-2	0.438	ORG-E-2	0.211
Cat+Clip+F_{tab}+BERT+MLP	RUN-J-2	0.218	RUN-E-1	0.037
Cat+F_{tab}+BERT+MLP	RUN-J-4	0.218	RUN-E-3	0.028
Clip+F_{tab}+BERT+MLP	RUN-J-6	0.151	RUN-E-5	0.044
F_{tab}+BERT+MLP	RUN-J-8	0.151	RUN-E-7	0.051
Cat+BM25	RUN-J-10	0.314	RUN-E-9	0.069

Clip: Indicates that the area of the table in which the features are calculated is cut out so that the proportion of the range of the headers becomes relatively high. The range is set to 15 columns and 5 rows based on the upper left corner.

DISCUSSION

- The reranking method combining table body features and BERT-MLP model did not contribute to improve ranking results.
  - This may be due to the fact that the information in the table body, which is mostly numerical, was converted to a feature vector using BERT.
  - The body of the table used in the previous work contains ordinary non-numeric word sequences as values.
- Table clipping had no effect.
  - An examination of the feature vectors input to the MLP revealed that each feature had large outlier-like values, and scaling by MaxAbsScaler resulted in most feature values being close to zero.
  - As a result, the table's original features could not be fully utilized.
- Category search was effective in improving rankings.
  - The same results were obtained in the previous Data Search Task.

CONCLUSION

- Method
    - Re-ranking method combining table body features and BERT-MLP
    - Together with Data augmentation and Category search
- Result
    - The proposed re-ranking with table body features and BERT-MLP showed no improvement.
    - Category search was effective in improving rankings.
- Future work
    - Consider features corresponding to the structure of the table as well as the information in the table itself.