KSU Systems at the NTCIR-15 Data Search Task

○**Taku Okamoto** and Hisashi Miyamori
Kyoto Sangyo University
NTCIR-15 DAY 3: December 10 Thu. 15:10 - 15:20

Methods

1. **Category search**
   Narrows down the set of documents to be retrieved by category, to properly capture the scope of the query.
2. **Metadata augmentation by table headers**
   Augments metadata by table header information, to compensate the short document length of metadata.
3. **Reranking by BERT**
   Applied the BERT-based reranking, to see how much contribution it achieves for this task.

Category search

- Narrow down the documents by the categories used in Yahoo! Chiebukuro or Yahoo Answers.
  - Japanese: 10 categories, English: 23 categories
- When indexing, each document is assigned a category by a pre-built text classifier.
- When searching, the result is ranked only on the set of documents belonging to the category estimated from the given query

Category search

- Collected documents from Yahoo! Chiebukuro and Yahoo Answers.
- Trained a text classifier with the collected documents.

Table 1. Statistics for the set of documents collected to build the classifier

Subtask	# of categories	Avg # of doc/cat	Stddev of # of doc/cat	Total # of QA pairs
Japanese	10	149.4	6.28	1494
English	23	158	99.26	3541

Augmentation by table headers

- Compensates for the short document length of the metadata.
- Procedure:
  1. Tables in various formats are **converted into images**.
  2. **Cell type, i.e. header or not, is recognized** for each cell using a classifier.
  3. **Texts are extracted by OCR** from cells recognized as header.

Augmentation by table headers

- Compensates for the short document length of the metadata.
- Procedure:
  1. Tables in various formats are **converted into images**.
  2. **Cell type, i.e. header or not, is recognized** for each cell using a classifier.
  3. **Texts are extracted by OCR** from cells recognized as header.

→ **Low accuracy of OCR**

Augmentation by table headers

- Revised for extra run of Japanese subtask:
  - **Identify the header with the rule:**
    - If the **number of non-empty cells in the current line increased** from that in the previous line, the current line is **considered as a header.**
    - Same for columns

Non-empty cell count					
0					
2	**Origin**		**Male**		identified headers
4	Prefectures	Area code	Total	0~9 years old	
4	Hokkaido	0000	8,105	925	

Figure 1. Revised extraction of table headers for Japanese subtask

Augmentation by table headers

- Revised for extra run of English subtask:
  - Limit ourselves to PDF files
  - Obtained the entire strings from each file

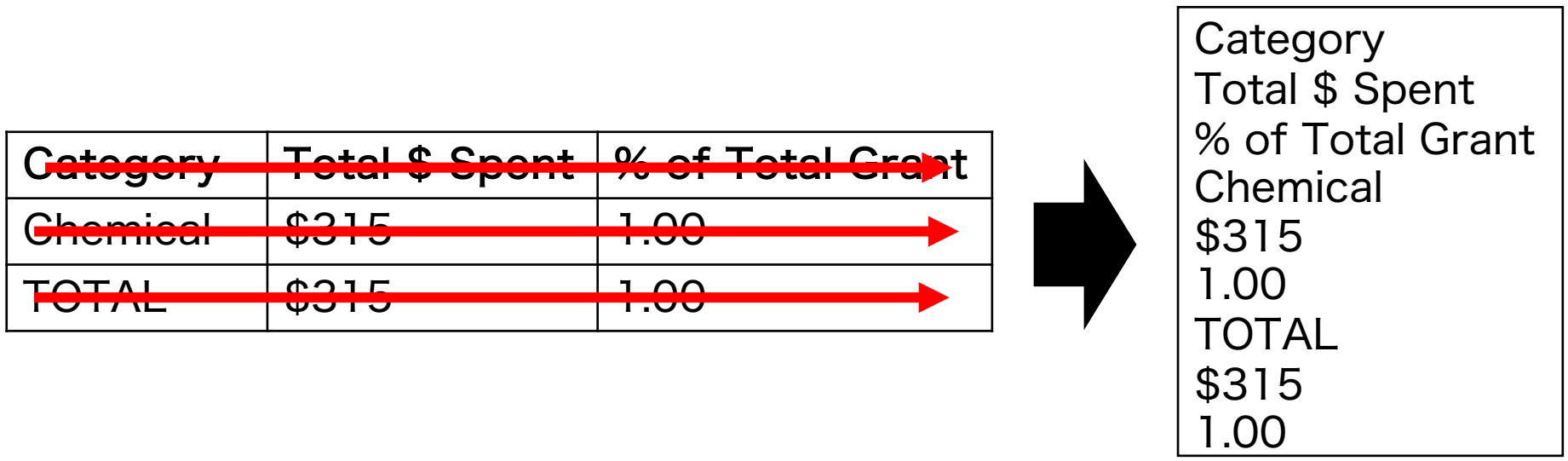

Figure 2. Revised extraction of table headers for English subtask

Reranking by BERT

- Apply **reranking by BERT to the top set of documents obtained by normal search** with BM25.
- Specifically, the sentence level score inferenced by BERT is combined with the normal document score according to the following equation:

$$S_f = a \cdot S_{doc} + (1 - a) \cdot \sum_{i=1}^{n} w_i \cdot S_i$$

S_f : final doc score　S_i : top i-th sentence score by BERT

S_{doc} : doc score before reranking　a, w_i : parameters

Result：Japanese

Table 2. Evaluation result for Japanese subtask

RUN	Category search	Table header	ranking	Text classifier for category search			Table header extraction	nDCG@10
				POS	Vector	training		
KSU-J-1	✓	✓	BM25	ALL	TF	MLP	OCR+CRF	0.391
KSU-J-3		✓	Bert reranking				OCR+CRF	0.110
KSU-J-5	✓		BM25	ALL	TF	MLP		**0.413**
KSU-J-7			Bert reranking					0.110
KSU-J-EX-1	✓	✓	BM25	N+V	Fasttest	SVM	ROW+COL	**0.426**
KSU-J-EX-2	✓	✓	BM25	N+V	Fasttest	SVM	ROW	0.276
KSU-J-EX-3	✓		BM25	N+V	Fasttest	SVM		0.353
KSU-J-EX-6	✓	✓	BM25	N+V	Fasttest	LR	ROW+COL	0.426
KSU-J-EX-7	✓	✓	BM25	N+V	Fasttest	LR	ROW	0.276
KSU-J-EX-8	✓	✓	BM25	N+V	Fasttest	LR		0.342

Result：English

Table 3. Evaluation result for English subtask

RUN	Category search	Table header	ranking	Text classifier for category search			Table header extraction	nDCG@10
				POS	Vector	training		
KSU-E-2	✓	✓	BM25	ALL	TF	MLP	OCR+CRF	**0.240**
KSU-E-4		✓	Bert reranking				OCR+CRF	0. 051
KSU-E-6	✓		BM25	ALL	TF	MLP		0.240
KSU-E-8			Bert reranking					0.038
KSU-E-EX-4	✓	✓	BM25	ALL	Fasttest	SVM	ALL	0.042
KSU-E-EX-5	✓		BM25	ALL	Fasttest	SVM		0.181
KSU-E-EX-9	✓	✓	BM25	ALL	Fasttest	LR	ALL	0.043
KSU-E-EX-10	✓		BM25	ALL	Fasttest	LR		**0.216**

Discussion: Japanese

- Modification on category classifier and header extraction method successfully improved the result.
- Confirmed that headers failed to be extracted properly for some tables.
- Semantic content of the header may need to be considered.

HS	00　全　国	379297	300059	7581
HS	01　北海道	6157	4919	131
HS	01100札幌市	-	-	-
HS	01202函館市	-	-	-
HS	01203小樽市	-	-	-
HS	01204旭川市	-	-	-
HS	01205室蘭市	-	-	-
HS	01206釧路市	-	-	-
HS	01207帯広市	-	-	-
HS	01208北見市	-	-	-
HS	01209夕張市	-	-	-
HS	01210岩見沢	-	-	-
HS	01211網走市	-	-	-
HS	01212留萌市	-	-	-
HS	01213苫小牧	130	112	-
HS	01214稚内市	-	-	-
HS	01215美唄市	157	123	1

Figure 2. Example of a table where header extraction failed

Discussion: Japanese

- Modification on category classifier and header extraction method successfully improved the result.
- Confirmed that headers failed to be extracted properly for some tables.
- Semantic content of the header may need to be considered.

Figure 2. Example of a table where header extraction failed

Discussion: English

- No improvement could be achieved in the extra run.
- Two reasons:
  1. Large variability of the training data
     - Ave：158, STD：99.26, Max：260, Min：20
  2. Full text were inappropriate as table headers
     - Excluding the numbers might have led to better results.

Conclusion

- We Introduced three methods:
  1. **Category search**
  2. **Metadata augmentation by table headers**
  3. **Reranking by BERT**
- Combined method of **category search and BM25** showed the highest score on NDCG@10 among all the official runs.