# KSU Systems at the NTCIR-15 Data Search Task


Taku Okamoto and Hisashi Miyamori

Affiliation 京都産業大学 KYOTO SANGYO UNIVERSITY

Conference Workshop NTCIR Task



## ABSTRACT

### Goal

- Data Search task
  - Ad-hoc retrieval of governmental statistical data
  - The main body basically composed of numbers, and the length of metadata is short

### Our methods

1. Category search
   - Narrows down the documents by category
2. Metadata augmentation by table headers
3. Reranking using BERT

### Result

The combination of category search and BM25 resulted in 0.426 for Japanese subtask and 0.240 for English subtask, both at nDCG@10, where each showed the highest score among all the official runs.


## Category Search

### Aim

- Narrows down the documents by category, to properly capture the scope of the query.

### When indexing

- Adopts the categories used in Yahoo! Chiebukuro or Yahoo! Answers
- Assigns each document 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

### Building a classifier

#### 1. Collection of datasets

Scrape all QA data and the categories set on each page from Yahoo! Chiebukuro and Yahoo! Answers. Table 1 shows the average and standard deviation for each category and total QA number.

**Tab. 1.** Distribution of collected data

| | Average | Stddev | QA datas | Category |
|---|---|---|---|---|
| English | 158 | 99.26 | 3,541 | 23 |
| Japanese | 149.4 | 6.28 | 1,494 | 10 |

#### 2. Training the classifier

Select the best combination of part-of-speech, vectors, and training methods via 10-fold cross-validation.

**Part-of-speech：** Noun, verb, all part-of-speech

**Vectors：** fastText, GloVe, TF

**Training：** MLP, SVM, Logistic Regression

**Accuracy：**

- Japanese → 69% （N+V, fastText, SVM)
- English → 58% （all POS, fastText, SVM)

## Augmentation by Table Header

### Aim

- Compensates for the short document length of the metadata

### Preliminary Analysis

After examining the metadata, we found that their document length was short. A typical metadata has an average length of 300-400 words.

**Tab. 2.** Statistics on metadata in the data collection

| Sub Task | document length | | number of documents |
|---|---|---|---|
| | average | stddev | |
| English | 101.93 | 81.19 | 46,615 |
| Japanese | 11.83 | 3.02 | 1,338,402 |

### Two approaches for header extraction:

#### 1. Extraction through images with OCR

In order to deal with various formats of statistical data, the table data is first converted into images, and contour extraction is performed to recognize the cell regions. Then, the header is recognized for each cell using the classifier. Finally, the text is extracted from the cells recognized as header by OCR.

#### 2. Simple heuristics to avoid misidentification of OCR

Gets the text from the rough area where the header is likely to present.

#### 2-1. English subtask

Limits to PDF files and obtains the entire string from each file.

#### 2-2. Japanese subtask

Extracts headers based on changes in the number of non-empty cells as shown in Fig.2.

```
Input: statistical data sd
Output: column headers hdr_col
  prev = 0
  hdr_col = []
  max_col = sd.column.length
  for i = 1,...,max_col do
    curr = sd.column[i].unempty_cells.length
    if curr > prev then

  hdr_col.append(sd.column[i].unempty_cells)
    end if
    prev = curr
  end for
  return hdr_col
```

**prev**：number of non-empty cells in the previous column
**hdr_col**：column header
**max_col**：number of columns of the statistical data
**curr**：number of non-empty cells

**Fig. 2.** Extraction focused on changes in the number of non-empty cells

## Reranking by BERT

### Aim

- Understands how much contribution can be expected from the pretrained language model.

### BERT and reranking

- BERT is a pre-trained language model that has been reported to have high performance in various fields.
- Applying reranking using BERT to the top set of documents obtained by normal search with BM25 could be more accurate than normal search.

### Score calculation

Inference by BERT is performed for each sentence of the candidate document, and the sentence level score 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 and Discussion

### Result for Japanese subtask

**Tab. 3.** 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 for English subtask

**Tab. 4.** 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 for Japanese subtask

- There were **some tables** where header extraction **did not work properly**. The **semantic content of the header may need to be considered** in extracting headers.

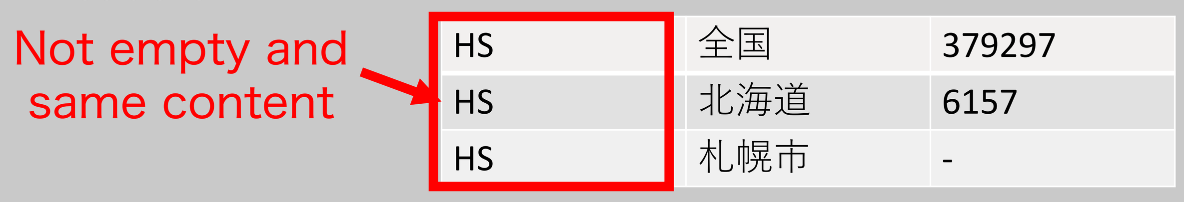


**Fig. 3.** Example of a table where header extraction failed

### Discussion for English subtask

- The maximum number of documents per category in the collected dataset was 260 and the minimum was 20, **indicating a large variation in the dataset**. Therefore, **the classification accuracy varies greatly depending on the category.**
- **All strings were extracted as table headers**. The extracted data contained a lot of numbers and some text consisting of ordinary words. **Excluding the numbers might have led to better results.**