MSRA at NTCIR-10 1CLICK-2

Kazuya Narita (Tohoku University, Japan), Tetsuya Sakai, Zhicheng Dou (MSRA, China), Young-In Song (NHN Corporation, Korea)
Overview

Input: Query
- Web search results

Sentence Ranker
- Ranked sentences

Post-Processor
- Output: X-string

Classify query type by some heuristic rules

Query Type Classifier
- Query Type

Rank sentences based on heuristic rules and attributes

Sentence Ranker
- Attributes

Find query attributes

Attribute Extractor
- Diversify ranked sentences
Overview

Input: Query

Web search results

Sentences

Sentence Ranker

Ranked sentences

Post-Processor

Output: X-string

Classify query type by some heuristic rules

Query Type Classifier

Query Type

Rank sentences based on heuristic rules and attributes

Sentence Ranker

Find query attributes

Attribute Extractor

Attributes

Diversify ranked sentences
Classification Rules

• Classification is based on some heuristic rules:
  – E.g. Query length → QA

“世界で初めてノーベル賞をとったのは誰か”
(Who took the Nobel Prize for the first time in the world?)

Long query may be QA
Classification Rules

- Classification is based on some heuristic rules:
  - E.g. Query length → QA
  - GEO clue suffixes at middle → GEO

  “博多駅 ホテル”
  (hotel near Hakata station)

  Other suffixes: “市” (city), ”区” (ward), “町” (town) ...
Classification Rules

• Classification is based on some heuristic rules:
  - E.g. Query length
  - GEO clue suffixes at middle
  - FACILITY clue suffixes

“須磨海浜水族園”
(Suma Aqualife Park)

Other suffixes: “学校” (school), ”病院” (hospital) ...
Classification Rules

- Classification is based on some heuristic rules:

  - E.g. Query length → QA
  - GEO clue suffixes at middle → GEO
  - FACILITY clue suffixes → FACILITY
  - Clue words in wiki pages → ARTIST, ACTOR, POLITICIAN, ATHLETE

Find a clue word "選手" (player) = ATHLETE

Other clue words: “小説家” (novelist), ”女優” (actress) ...
Classification Rules

- Classification is based on some heuristic rules:
  - E.g.
    - Query length
    - GEO clue suffixes at middle
    - FACILITY clue suffixes
    - Clue words in wiki pages
      - Artist
      - Actor
      - Politician
      - Athlete
    - Else

  → QA
  → GEO
  → FACILITY
  → ARTIST
  → ACTOR
  → POLITICIAN
  → ATHLETE
  → DEFINITION
Results of Query Type Classifier

- Accuracy: 83% (83/100)

<table>
<thead>
<tr>
<th>sys \ gold</th>
<th>ART</th>
<th>ACT</th>
<th>POL</th>
<th>ATH</th>
<th>FAC</th>
<th>GEO</th>
<th>DEF</th>
<th>QA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ARTIST</td>
<td>9</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACTOR</td>
<td>1</td>
<td>9</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>POLITICIAN</td>
<td></td>
<td></td>
<td>8</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ATHLETE</td>
<td></td>
<td></td>
<td></td>
<td>10</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>FACILITY</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>7</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>GEO</td>
<td></td>
<td></td>
<td>1</td>
<td></td>
<td>14</td>
<td></td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>DEFINITION</td>
<td>1</td>
<td>5</td>
<td></td>
<td></td>
<td></td>
<td>12</td>
<td>1</td>
<td></td>
</tr>
<tr>
<td>QA</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>14</td>
</tr>
</tbody>
</table>
## Error Analysis for FACILITY

### Correct Example

- Correct only by suffix rules
  - 須磨海浜水族園
    
    *(Suma Aqualife Park)*
  - 横浜市役所
    
    *(Yokohama City Hall)*
  - 小金井図書館
    
    *(Koganei library)*
  - ハワイパシフィック大学
    
    *(Hawaii Pacific University)*
  - あおやま矯正歯科医院
    
    *(Aoyama Orthodontic Office)*

### Wrong Example

- **Insufficient suffixes**
  - 京都真如堂
    
    *(Kyoto Shinnyo-do temple)*
  - 南ヶ丘牧場
    
    *(Minamigaoka dairy)*

- **Difficult only by suffix rules**
  - カーサ・ディ・ナポリ
    
    *(Casa di Napoli)*
  - ザ・ペニンシュラ東京
    
    *(The Peninsula Tokyo)*
  - らーめんてつや
    
    *(Ramen Tetsuya)*
Overview

Input: Query

Web search results

Sentences

Sentence Ranker

Ranked sentences

Post-Processor

Output: X-string

Classify query type by some heuristic rules

Query Type Classifier

Find query attributes

Attribute Extractor

Rank sentences based on heuristic rules and attributes
Query Attributes

Ichiro

Future Science Museum

Born

Height

Position

Address

Tel

Opening Hours
Query Attributes

**Ichiro**

**Future Science Museum**

These attributes may be useful when users are seeking for basic information of entities.

- Born
- Height
- Position

- Address
- Tel
- Opening Hours
Attribute Extraction

Web search results
Attribute Extraction

Born: October 22, 1973 (age 38)
Height: 5 ft 11 in (1.80m)
Position: Outfielder

Web search results

Extract tables or text containing “:”
Attribute Extraction

Web search results

Extract tables or
text containing “:”

Extract as query attributes

<table>
<thead>
<tr>
<th>Born</th>
<th>October 22, 1973 (age 38)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Height</td>
<td>5 ft 11 in (1.80m)</td>
</tr>
<tr>
<td>Position</td>
<td>Outfielder</td>
</tr>
</tbody>
</table>

**Born**: October 22, 1973 (age 38)

**Height**: 5 ft 11 in (1.80m)

**Position**: Outfielder
Overview

Input: Query

Web search results

Sentences

Sentence Ranker

Ranked sentences

Post-Processor

Output: X-string

Classify query type by some heuristic rules

Find query attributes

Diversify ranked sentences

Query Type Classifier

Query Type

Sentence Ranker

Rank sentences based on heuristic rules and attributes

Attribute Extractor

Attributes
Sentence Ranking

• Scoring based on following components:
  1. Content similarity
  2. Query attributes
  3. Manual clue words and URLs
  4. Search Rank
Scoring

1. Content similarity (**LexRank** [Erkan and Radev 2004])
   – Similar sentences are more relevant

   Importance of each sentences
   (similar to many other sentences -> High)

   \[
   p = \left[ dU + (1 - d)B \right]^T p
   \]

   All elements are 1/n
   (n: the number of sentences)

   Damping factor

   Adjacency matrix of the cosine similarity
Scoring

2. Query attributes

– Attributes extracted automatically are noisy

Ichiro

Born

Height

Address...

attribute list

Noisy attribute!
Scoring

2. Query attributes

– Attributes extracted automatically are noisy

Ichiro

Born

Height

Address

…

attribute list

Kohei Uchimura

Born

Height

…

attribute list

Same type (ATHLETE)

important attributes for ATHLETE

Noisy attribute for ATHLETE
Scoring

2. Query attributes

– Attributes extracted automatically are noisy

<table>
<thead>
<tr>
<th>Born</th>
<th>Height</th>
<th>Address</th>
<th>...</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ichiro</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Tel | Near Station | Address | ... |
| Facility | | | |

attribute list | Noisy attribute | attribute list
Scoring

2. Query attributes
   – Weight each attribute as importance like TF-IDF

We can give high priority to characteristic attributes

Weight of A > Weight of B
Scoring

3. Manual clue words and URLs

- “birthday” or “birth town”: more relevant for celebrity
- “talent.yahoo.co.jp”: more relevant for celebrity
- “address” or “phone number”: more relevant for facility

- “posted by” or “answerer”: less relevant for all
- “amazon.co.jp”: less important for all
4. Search rank

– Sentences of high rank webpage is more important

\[ S_{SRank}(s) = 1 - \frac{\text{rank}(s)}{N_{search}} \]
Sentence Ranking

• Scoring based on following components:
  1. Content similarity
     • Similar sentences are more relevant
  2. Query attributes
     • Weight for each attribute as importance like TF-IDF
  3. Manual clue words and URLs
  4. Search Rank
     • Sentences of high rank webpage is more important
Overview

Input: Query

Web search results

Sentences

Sentence Ranker

Ranged sentences

Post-Processor

Output: X-string

Classify query type by some heuristic rules

Query Type Classifier

Query Type

Rank sentences based on heuristic rules and attributes

Sentence Ranker

Find query attributes

Attribute Extractor

Attributes

Diversify ranked sentences
Sentence diversity

• Output top $N$ sentences as X-string
  – But similar sentences tend to be output
• Redundancy is undesirable for 1CLICK

We should diversify sentences
To consider sentence diversity

Similar sentences are eliminated

\[ \text{compare with already selected sentences (MMR algorithm and comparing set of sentences)} \]

\[ MMR = \arg \max_{s_i \in R \setminus S} \left[ \lambda \text{Score}(s_i) - (1 - \lambda) \max_{s_j \in S} \text{Sim}(s_i, s_j) \right] \]

Give a penalty depending on similarity to already selected sentences
To consider sentence diversity

Similar sentences are eliminated

Compare a sentence with bag of words of already selected sentences

```
sentence_1
sentence_2
sentence_3
...
sentence_n
```

```
sentence_1
sentence_3
...
```

Bag of words

**compare with already selected sentences**

*(MMR algorithm and comparing set of sentences)*

X-string
EVALUATION RESULTS
Results (mean S#-measures)

• Query attributes are effective for celebrity
  – Though the difference is not significant

Without query type
Without attribute
With attribute (binary)
With attribute (tf-idf)

Overall  ARTIST  ACTOR  POLITICIAN  ATHLETE
Results (mean S#-measures)

• Query attributes do **not work well for** DEFINITION and QA queries
Comparison with other participants

- Top performance among MANDATORY runs
  - Though the difference is not significant
Conclusions

- Automatic query attribute extraction is effective especially for celebrity queries
  - tf-idf weight is effective

- Future work:
  - Automatic extraction of clue words for query type classification
  - New framework instead of query attributes for DEFINITION and QA queries