Information Extraction based Approach for the NTCIR-9 1CLICK Task

Meng Zhao, Kosetsu Tsukuda, Yoshiyuki Shoji, Makoto P. Kato, Takehiro Yamamoto, Hiroaki Ohshima, Katsumi Tanaka

Graduate School of Informatics, Kyoto University

NTCIR-9 Workshop Meeting, 2011/12/9, NII Tokyo, Japan
Our Framework

query

Query Classifier

Information Extractor

IE for CELEBRITY
Wikipedia

IE for LOCATION
Official Web pages

IE for DEFINITION
Summarizing Web search results

IE for QA
Yahoo! Chiebukuro

Information Summarizer

Summarized Text
Query Classifier

Classify four types of query by multi class SVM

185 Features

<table>
<thead>
<tr>
<th>Feature</th>
<th># of features</th>
</tr>
</thead>
<tbody>
<tr>
<td>Has Wikipedia article</td>
<td>1</td>
</tr>
<tr>
<td>Frequency of parts-of-speech</td>
<td>44</td>
</tr>
<tr>
<td>Query unigram</td>
<td>85</td>
</tr>
<tr>
<td>Sentence pattern</td>
<td>2</td>
</tr>
<tr>
<td>Number of documents containing expanded query</td>
<td>15</td>
</tr>
<tr>
<td>Has travel services</td>
<td>1</td>
</tr>
<tr>
<td>Number of search results</td>
<td>1</td>
</tr>
<tr>
<td>Terms in search results</td>
<td>39</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>185</strong></td>
</tr>
</tbody>
</table>

Search with 15 expanded query such as “query-san(さん)” and “query-senshu(選手)” to distinguish name of CELEBRITY.

Count selected 39 terms such as “profile,” “born at” and “Chome (丁目)” in snippets of Yahoo Japan search result
Information Summarizer

- Eliminates overlap from the extracted sentences.

**MMR-based Information Summarization Method**

\[ MMR = \arg \max_{d_i \in D/S} \left[ \lambda (Score(d_i)) - (1 - \lambda) \max_{d_j \in S} \text{sim}(d_i, d_j) \right] \]

- Score of the sentence
- Similarity to the output sentences

Because it is too similar to the already selected sentences

Because of the output length limitation

Output

Shorter than the limitation
IE for CELEBLITY (Attribute-based)

- Extract pairs of an attribute name and its value from Infobox in Wikipedia

http://en.wikipedia.org/wiki/Andre_Agassi

“Andre Agassi”

“Country is USA, residence is Las Vegas, Nevada, U.S., height is ...., and highest ranking is No.1.”
IE for CELEBLITY (Sentence-based)

- Extract important sentences from Wikipedia
- 12 features employed

- Train a regressor to predict the importance score of each sentence (using pre-distributed queries)

---

**Hayao Miyazaki**

ContainsQuery

```
Hayao Miyazaki (Japanese: 宮崎駿, born January 5, 1941) is a Japanese manga artist and prominent film director and animator of many popular anime series. He is known for his films such as Spirited Away, My Neighbor Totoro, and Princess Mononoke. Miyazaki's films often incorporate recurring themes like humanity's relationship to nature and technology, and the difficulty of maintaining a pacifist ethic. Reflecting Miyazaki's feminism, the protagonists of his films are often strong, independent girls or young women. Miyazaki is a vocal critic of capitalism and globalization.
```

- The position of a sentence in a section
- RelativePosition

---

query: Hayao Miyazaki

(Yahoo! Japan Web search API)

Summaries(Top n)

MaxCos, MinCos, MaxInnerProd, AvgCos, AvgInnerProd

Calculate cosine, inner-product similarity with each summary

---

http://en.wikipedia.org/wiki/Hayao_Miyazaki
IE for LOCATION

Collect official Web pages

**query**: Kyoto Royal Hotel

www.royalhotel.com AND ( address OR access OR ...)

(Yahoo! Japan Web search API)

Extract attributes and their values

**Regular Expression based extraction**

〒 [0-9]{3} -[0-9]{4}
→ Address: 〒606-8501
[0-9]{3} -[0-9]{3} -[0-9]{4}
→ Tel: 075-753-5385
[a-Z]*@[a-Z]*
→ Mail: hotel@kyoto.com

**Sentence based extraction**

10 minutes walk from Kyoto station

access information? (SVM classifier)

**Table based extraction**

<table>
<thead>
<tr>
<th>check in</th>
<th>15:00</th>
</tr>
</thead>
<tbody>
<tr>
<td>check out</td>
<td>10:00</td>
</tr>
</tbody>
</table>

Check in : 15:00
Check out: 10:00
• Query: $x$
• Search with a new query “$x$ towa (とは)”
• Extract a sentence including the phrase of “$x$ towa” from each retrieved Web page
• Apply the *LexRank*[^1] algorithm to the set of sentences for estimating the sentence importance

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

- $p$: a vector representing the importance degree of each sentence
- $U$: $n \times n$ square matrix whose elements are $1/n$
- $B$: adjacency matrix of the cosine similarity between two sentences
- $d$: damping factor

IE for QA

- **Query:** Which is taller, Tsutenkaku or Ustunomiya tower?

  Calculate $n$-gram similarity

- **Extract answers for most similar questions to the given query from Yahoo! Chiebukuro**

- **Return the best answer**
  - if no best answer is given, return the most similar answer to the query
Results

• The accuracy of our query classifier: \textbf{0.93}

• Highest S-measures for CELEBLITY and QA queries among participants
  – though the difference is not significant
• For the 1CLICK task, the object identification problem should be tackled (see Kanazawa Univ.)
• QA queries obtained high S-measures when exactly the same questions are available (see the two questions)
Conclusions

• Information extraction framework for 1CLICK
  – Query classifier (**multiclass SVM**)  
  – Four types of information extractor
    • CELEBRITY (**attribute and sentence extraction from Wikipedia**)  
    • LOCATION (**postal and access information extraction**)  
    • DEFINITION (**summarization of Web search results**)  
    • QA (**QA pair extraction from Yahoo! Chiebukuro**)  

– Information summarizer (**MMR**)  

Thanks! ntcir-9@dl.kuis.kyoto-u.ac.jp