ABSTRACT

Our group NEXTI participated in the Query Understanding subtask for Japanese. We extracted subtopic candidates from retrieved web documents. Then, we merged them with query suggestion and query log data. We also identified the vertical intent of each subtopic using a method that combines machine learning- and k-NN-based methods. We conducted experiments and confirmed the effectiveness of our method.

Keywords
k-NN, SVM, clustering

1. INTRODUCTION

Our group started Project Next IR1 in July 2014. The purpose of this project is to investigate the kinds of technologies that are required to enhance the current IR systems and the related problems that need solving. We conducted failure analysis using various test collections [Nanba 2016] and found the limitation of the bag-of-words approach in IR. We believe that the relations between words in a document should be taken into account in some way, and that the IMine-2 task is a good starting point to address this challenge.

The remainder of this paper is organized as follows. In Section 2, we propose the method for identifying subtopics with their vertical intent. Section 3 reports the experimental results and Section 4 concludes.

2. IDENTIFICATION OF SUBTOPICS

Our method of identifying subtopics consists of the following three steps.

1. Extraction of subtopic candidates from web documents
2. Merging of the candidates with query suggestion and query log data
3. Identification of the vertical intent for each subtopic

We describe the details of each step as follows.

2.1 Extraction of Subtopic Candidates from Web Documents

In this step, we extract subtopic candidates from web documents, which were provided by the organizers. For the extraction of candidates, we focused on some linguistic patterns. We assumed that a subtopic appears with its (main) topic in the same sentence. Let us consider that “iPhone 6” is a topic and “レビュー (review)” is a subtopic, and there is an expression “iPhone 6 のレビュー (review of iPhone 6)” in documents. We call character strings between a topic and a subtopic, such as “(of)” linguistic patterns.

If we apply a pattern “iPhone 6 の [noun phrase]” to documents, and collect noun phrases, we may find new subtopics in the noun phrases.

To collect these patterns, we used NTCIR-11 IMine data [Liu 2014]. First, we retrieved 500 web documents for each topic using Google. Second, we collected character strings between a topic and a subtopic. Finally, we collected 19 linguistic patterns in Table 1. S and T indicate a topic and a subtopic, respectively. <NP> and </NP> indicate start and end points of noun phrases. We identified nouns in a sentence using a Japanese morphological analyzer, JUMAN2, and regarded continuous nouns as a noun phrase. In Table 1, “TS” indicates that both topic and subtopic appear in the same noun phrase, while “TS<NP>” indicates that the topic and the subtopic appear in different noun phrases, and “(of)” appears between the topic and the subtopic. Using these patterns, we extracted subtopic candidates from web documents.

2.2 Mergence of the Candidates with Query Suggestion and Query Log Data

In this step, we merged the candidates extracted in step 1 with the following six data, which were provided from the organizers.

- Query suggestion by Google
- Query suggestion by Yahoo
- Query suggestion by Bing
- Co-click queries
- Co-session queries
- Co-topic queries

1 http://165.242.101.30/~nanba/task/project_next/ (in Japanese)
2 http://nlp.ist.i.kyoto-u.ac.jp/index.php?cmd=read&page=JUMAN&alias%5B%5D=%E6%97%A5%E6%9C%AC%E8%AA%9E%E5%BD%A2%E6%85%8B%E7%B4%A0%E8%A7%A3%E6%9E%90%E3%82%B7%E3%82%B9%E3%83%86%E3%83%A0JUMAN
We show the procedure of the mergence in Figure 1. We merged the candidates using equation 1.

Table 1. Linguistic Patterns for Extracting Subtopic Candidates from Web Documents.

<table>
<thead>
<tr>
<th>pattern</th>
<th>freq.</th>
<th>pattern</th>
<th>freq.</th>
</tr>
</thead>
<tbody>
<tr>
<td>TS</td>
<td>574</td>
<td>S&lt;NP&gt;, T&lt;NP&gt;</td>
<td>5</td>
</tr>
<tr>
<td>T&lt;NP&gt;の&lt;NP&gt;S</td>
<td>191</td>
<td>S,T</td>
<td>5</td>
</tr>
<tr>
<td>ST</td>
<td>126</td>
<td>S&lt;NP&gt;;S&lt;NP&gt;T</td>
<td>5</td>
</tr>
<tr>
<td>S&lt;NP&gt;; &lt;NP&gt;T</td>
<td>45</td>
<td>T&lt;NP&gt;; は, &lt;NP&gt;;S</td>
<td>4</td>
</tr>
<tr>
<td>S&lt;NP&gt;;・&lt;NP&gt;T</td>
<td>24</td>
<td>T&lt;NP&gt;; から, &lt;NP&gt;;S</td>
<td>3</td>
</tr>
<tr>
<td>T&lt;NP&gt;をS</td>
<td>23</td>
<td>T&lt;NP&gt; (&lt;NP&gt;;S</td>
<td>3</td>
</tr>
<tr>
<td>TS</td>
<td>13</td>
<td>S&lt;NP&gt;;と&lt;NP&gt;T</td>
<td>3</td>
</tr>
<tr>
<td>T&lt;NP&gt;;を&lt;NP&gt;S</td>
<td>12</td>
<td>S&lt;NP&gt;;, &lt;NP&gt;T</td>
<td>3</td>
</tr>
<tr>
<td>S&lt;NP&gt;; &lt;NP&gt;T</td>
<td>9</td>
<td>T&lt;NP&gt;;の&lt;NP&gt;;S</td>
<td>3</td>
</tr>
<tr>
<td>T&lt;NP&gt;;は&lt;NP&gt;S</td>
<td>8</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Score}(\text{candidate}) = \sum_{i=1}^{n} W_i \text{ occur}(\text{candidate}, d_i) \quad (1)
\]

Here, \(\text{occur}(\text{candidate}, d)\) is a 0-1 function that indicates whether \(\text{candidate}\) occurs in \(d\) \((\text{occur}(\text{candidate}, d)=1)\) or not \((\text{occur}(\text{candidate}, d)=0)\). \(d_1 - d_n\) indicate candidates from web documents, query suggestions using Google, Yahoo, and Bing, and query logs using co-click, co-session, and co-topic, respectively. \(W_1=1/3\), and \(W_2 - W_r = 1/6\). As a result of this mergence, we obtained 1,683,456 candidates (for 100 topics) in total.

Then, we conducted clustering for grouping similar subtopics. We assumed that similar subtopics tend to retrieve similar documents. Based on this idea, we retrieved documents using topic-subtopic pairs, and used the top 10 documents with the highest relevance score as features for clustering. We used 44,280 web documents provided by organizers, and constructed an IR system using INDRI\(^1\). For the clustering, we used bayon\(^2\). We set the number of clusters as 10, because we were allowed to submit 10 subtopics at most for each topic. Among subtopic candidates in each cluster, we selected the candidate having the highest score calculated by equation 1.

### 2.3 Identification of the Vertical Intent for Each Subtopic

Our method of identifying the vertical intent for each subtopic consists of the following two steps.

- Identification of content types (vertical intents) of each web page
- Identification of the vertical intent for each subtopic

We describe the details of each step in the following sections.

#### 2.3.1 Identification of Content Types of Each Web Page

Features for machine learning

We identified content types (vertical intents) of each web page using a machine learning-based approach. We used frequencies of the following cue phrases in a document as features for machine learning.

- **Cue phrases for encyclopedia in web pages:** とは (is defined), って (is defined), 事典 (encyclopedia), 辞典 (dictionary), 攻略 (hints-and-tips), 解説 (commentary), 講座 (lecture), 作り方 (how to), すれば良い (recommend to), してください (had better to do), します (do), されます (was done), 手順 (procedure), 編集 (edit), 差分 (view history), 最終更新 (last update), 新規 (new), wiki, 参考 (reference), 出典 (reference), 資料 (material)
- **Cue phrases for encyclopedia in URLs:** jp.wikipedia, dic, weblio.jp, tools
- **Cue phrases for news in web pages:** ニュース (news), 新聞 (news article), トピックス (topics), 日経 (Nikkei), 速報 (breaking news), お知らせ (news), 紹介 (introduction), 開催 (release), 配信 (news distribution), 報じる (report), テレビ (TV), 記事 (article), 更新 (updated date), Introduction, Staff, Cast, Story, Character, Trailer, スタッフ (staff), キャスト (cast), ストーリー (story), あらすじ (outline), 登場人物 (characters), 予告 (trailer), 試合結果 (result of games), 成績 (score)
- **Cue phrases for news in URLs:** news, headlines
- **Cue phrases for image in web pages:** ムービー (movie), 動画 (video), 画像 (image), 壁紙 (wallpaper), 待ち受け (standby screen), 画面 (display)

---

\(^1\) http://www.lemurproject.org/indri/

\(^2\) https://code.google.com/archive/p/bayon/
Experimental settings

For classifying multiclasses using binary classifier SVM, we used the one-vs-rest classification method and a linear kernel. We used TinySVM\(^2\) as a machine learning package. We conducted 5-fold cross validation. We used recall and precision for evaluation.

Results

We show the experimental results in Table 3.

<table>
<thead>
<tr>
<th>Content Type (Vertical Intent)</th>
<th>Recall</th>
<th>Precision</th>
</tr>
</thead>
<tbody>
<tr>
<td>Encyclopedia</td>
<td>0.726</td>
<td>0.794</td>
</tr>
<tr>
<td>News</td>
<td>0.319</td>
<td>0.625</td>
</tr>
<tr>
<td>Image</td>
<td>0.803</td>
<td>0.994</td>
</tr>
<tr>
<td>QA</td>
<td>0.486</td>
<td>0.607</td>
</tr>
<tr>
<td>Shopping</td>
<td>0.267</td>
<td>0.492</td>
</tr>
<tr>
<td>Web</td>
<td>0.586</td>
<td>0.627</td>
</tr>
<tr>
<td>Average</td>
<td>0.591</td>
<td>0.735</td>
</tr>
</tbody>
</table>

We classified 44,280 web documents into six content types, and used them for the identification of the vertical intent for each subtopic.

2.3.2 Identification of the Vertical Intent for Each Subtopic

We identified vertical intents using the k-NN method. We retrieved the top 500 documents using a topic and subtopic pair as a query by the INDRI-based search engine (see Section 2.2). Then, we determined the category that appears most frequently in the top 500 documents as the vertical intent of the subtopic.

3. EXPERIMENTS

3.1 Data and Evaluation

We used 100 topics for the Japanese subtask to evaluate our system. Our system was evaluated by I-rec, D-nDCG@10, D#-nDCG@10, V-score, and QU-score.

3.2 Results and Discussion

We submitted one result provided by our system NEXTI. The experimental results are shown in Table 4. Our system outperformed other systems in all evaluation.

---

\(^2\) http://chasen.org/~taku/software/TinySVM/
Table 4. Experimental Results.

<table>
<thead>
<tr>
<th>Evaluation Metrics</th>
<th>Scores</th>
</tr>
</thead>
<tbody>
<tr>
<td>I-rec@10</td>
<td>0.6535</td>
</tr>
<tr>
<td>D-nDCG@10</td>
<td>0.5535</td>
</tr>
<tr>
<td>D#-nDCG@10</td>
<td>0.6035</td>
</tr>
<tr>
<td>V-score(Q-Run only)</td>
<td>0.6507</td>
</tr>
<tr>
<td>QU-score(Q-Run only)</td>
<td>0.6271</td>
</tr>
</tbody>
</table>

In the evaluation by QU-score, our system obtained the best scores in 35 topics, while obtained the worst in 8 topics. In the evaluation by I-rec, our system obtained the best scores in 40 topics, while obtained the worst scores in 18 topics.

To investigate the effectiveness of our method, we compared the 15 systems in detail. Figure 2 shows the number of topics over the threshold value (I-rec score) for each system. In the figure, for example, the number of topics over 0.9 I-rec score in our system is 18, which is the highest among all systems. When the threshold value is over 0.6, our system greatly outperformed others.

4. CONCLUSION

In this paper, we proposed a method that identified subtopics with their vertical intents for a given topic. From the experimental results, our system obtained the best performance among all participant groups.

REFERENCES


Figure 2. Comparison of 15 systems by I-rec.