Overview of Architecture

Retrieval-based method

- Query Generation
  A query used for similarity search is generated from the given input data by using the analyzers provided in Solr. In Run 1, the place names and person names were removed from the words in the theme as the search query.

- Document Retrieval
  Similarity search by Okapi BM25 is conducted for only comment texts in the Yahoo! news comments data. The each query was weighted. (Refer to the table on the left)

<table>
<thead>
<tr>
<th>Query type</th>
<th>Weight</th>
</tr>
</thead>
<tbody>
<tr>
<td>Comment text</td>
<td>1.0</td>
</tr>
<tr>
<td>Title in Yahoo! Topics</td>
<td>9.1</td>
</tr>
<tr>
<td>Theme</td>
<td>3.5</td>
</tr>
</tbody>
</table>

Generation-based method

Why do we use the visual information?

- The image or video may contain more detailed information than texts.

How can the visual information be used without image input?

1. Visual association from the input text
2. Fusion between textual information and associated visual information
3. Response text generation based on the fused information.

Learning Method

Prior train 1: Extraction of context vectors between textual and visual information

The model has two LSTM encoder, one fusion layer, and one decoder with attention. C_t is calculated by following equation at the fusion layer. The trained model is used to extract the correspondence between the textual and visual information.

C_t = W_{ext}C_t^ext + W_{vis}C_t^vis + b_c

Prior train 2: Learning for visual association

The associative encoder (LSTM) is trained, which inputs the textual context vector C_t^ext extracted in step 1 and outputs the visual context vector C_t^vis corresponding to the input utterance text.

C_t^vis = LSTM(C_t)

Final step: Generation of response text via association

Figure 2 shows a network used in step 3. Learning is performed in the network where the visual encoder in step 1 is replaced with the associative encoder trained in step 2.

Conclusion

In the retrieval-based methods, it was found that the accuracy is improved by excluding place names and person names from article themes.

In the generation-based method, we could not find enough evidence from the results using the evaluation data of STC to show that the associated visual information would accelerate to generate more appropriate responses. Newly, we conducted additional experiment based on the problems found in the evaluation results. As a preliminary result, it was confirmed that visual information seemed to work effectively in several examples.