KitAi-PI: Summarization System for NTCIR-14 QA Lab-PoliInfo

Satoshi Hiai, Yuka Otani, Takashi Yamamura
and Kazutaka Shimada

Department of Artificial Intelligence,
Kyushu Institute of Technology
Contents

• Introduction and Objective
• Outline of Our System
• Training Data Construction
• Formal Run
• Summary
Contents

• Introduction and Objective
• Outline of Our System
• Training Data Construction
• Formal Run
• Summary
Introduction – Assembly Minutes Summarization

• Two types of summarization methods
  • Abstractive: Use of expressions not contained in the source text
  • Extractive: Use of expressions in the source text

• Assembly minutes corpus
  • A summary consists of expressions contained in a speech

<table>
<thead>
<tr>
<th>Summary</th>
<th>Assembly member speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>被災地そして日本の未来のため 東京は先頭に立つべき。知事の所見は。</td>
<td>U.1 我々が生きている日本列島は、数限りない天変地異に見舞われてきました。</td>
</tr>
<tr>
<td>The same expressions</td>
<td>U.2 被災地のため、そして、日本の未来のため に、東京は先頭に立つべきと考えます が、知事の所見を伺います。</td>
</tr>
</tbody>
</table>

Summary generation with an extractive approach
Introduction – Extractive Summarization

• Extraction of a set of important utterances
  • Supervised method usually shows better performance than unsupervised method
  • Use of a machine learning method
    • Construction of importance prediction model
• Problem
  • Given assembly minutes data do not contain importance information for each utterance
Objective

- Automatic training data construction
  - Hypothesis
    - An utterance with high similarity to a sentence in a summary is more important

Summary

被災地そして日本の未来のため
東京は先頭に立つべき。知事の所見は。

Assembly member speech

<table>
<thead>
<tr>
<th>U.1</th>
<th>我々が生きている日本列島は、数限りない天変地異に見舞われてきました。</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.2</td>
<td>被災地のため、そして、日本の未来のため、東京は先頭に立つべきと考えますが、知事の所見を伺います。</td>
</tr>
</tbody>
</table>

No importance information for utterances
Objective

- **Automatic training data construction**
  - **Hypothesis**
    - An utterance with high similarity to a sentence in a summary is more important

<table>
<thead>
<tr>
<th>Summary</th>
</tr>
</thead>
</table>
| 被災地そして日本の未来のため
東京は先頭に立つべき。知事の所見は。 |

<table>
<thead>
<tr>
<th>Assembly member speech</th>
</tr>
</thead>
<tbody>
<tr>
<td>U.1</td>
</tr>
<tr>
<td>我々が生きている日本列島は、数限りない天変地異に見舞われてきました。</td>
</tr>
<tr>
<td>U.2</td>
</tr>
<tr>
<td>被災地のため、そして、日本の未来のために、東京は先頭に立つべきと考えますが、知事の所見を伺います。</td>
</tr>
</tbody>
</table>

We can apply a machine learning method
Contents

• Introduction and Objective
• Outline of Our System
• Training Data Construction
• Formal Run
• Summary
Outline of Our System

- Training data construction
- Training utterance importance prediction model
- Utterance extraction with trained model

Speech
Uttrunce.1
Uttrunce.2
...

Reference summaries

Training data
Speeches and Importance scores of utterances

Importance prediction model

Generated summary
Uttrunce.2: 0.8

Importance
Prediction
model

Speech
Uttrunce.1: 0.4
Uttrunce.2: 0.8
...
Contents

• Introduction and Objective
• Outline of Our System
• Training Data Construction
• Formal Run
• Summary
Training Data Construction – Assignment of Importance Scores

- Automatic assignment of an importance score to each utterance using a word similarity
  - We regard a word similarity as an importance score

<table>
<thead>
<tr>
<th>Assembly member speech</th>
<th>Summary</th>
</tr>
</thead>
<tbody>
<tr>
<td>Utterance 1</td>
<td>0.123</td>
</tr>
<tr>
<td>Utterance 2</td>
<td>0.900</td>
</tr>
<tr>
<td>Utterance 3</td>
<td>...</td>
</tr>
<tr>
<td>Utterance N</td>
<td>0.201</td>
</tr>
<tr>
<td>Utterance 1</td>
<td>0.820</td>
</tr>
<tr>
<td>Utterance 2</td>
<td>0.110</td>
</tr>
<tr>
<td>Utterance 3</td>
<td>...</td>
</tr>
<tr>
<td>Utterance N</td>
<td>0.221</td>
</tr>
</tbody>
</table>

- Evaluation of similarity measures
  - e.g. cosine similarity, edit distance, ...
Training Data Construction – Evaluation of Similarity Measures

- Given corpus: 529 speeches (7,226 utterances)
- Training data: 477 speeches (6,551 utterances)
- Development data: 52 speeches (675 utterances)

**Similarity measures**
- Cosine similarity between BoWs
- Edit distance

Evaluation of Similarity Measures

- Training data of each similarity measure
- Reference summaries

Importance prediction model

Speech ▸ Generated Summary
Training Data Construction – Similarity Measures

• Cosine similarity between bag-of-words
• Edit distance
  • We adopt $1 - (\text{the distances})$ as the similarity measure
• ROUGE-1 similarity score
  • We use word unigram overlap
• Cosine similarity between sentence embeddings
  • Two methods to generate sentence embeddings
    • Average of word embeddings generated with word2vec
    • Sentence embedding generated with doc2vec
• Average of all the similarity measures
Training Data Construction – Result of Similarity Measures Evaluation

- Evaluation of generated summaries

<table>
<thead>
<tr>
<th>Similarity measure</th>
<th>Rouge N1</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cosine similarity between bag-of-words</td>
<td>0.333</td>
</tr>
<tr>
<td>Edit distance</td>
<td>0.338</td>
</tr>
<tr>
<td>ROUGE-1 similarity score</td>
<td>0.341</td>
</tr>
<tr>
<td>Cosine similarity between sentence embedding (Word2vec)</td>
<td>0.306</td>
</tr>
<tr>
<td>Cosine similarity between sentence embedding (Doc2vec)</td>
<td>0.316</td>
</tr>
<tr>
<td>Average of all of the similarity measures</td>
<td><strong>0.349</strong></td>
</tr>
</tbody>
</table>

Average of all the similarity measures is adopted on the formal run
Contents

• Introduction and Objective
• Outline of Our System
• Training Data Construction
• Formal Run
• Summary
Settings for Formal Run

- Importance prediction model
  - Features
    - BoW, sentence position in the speech, speaker of the speech
  - Support vector regression (SVR)
- Our methods for the formal run
  - w/ sentence compression
    - We applied a sentence compression on the basis of simple rules
  - w/o sentence compression
## Result on Formal Run – ROUGE Scores

<table>
<thead>
<tr>
<th>Surface form</th>
<th>Recall</th>
<th>F-measure</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>N1</td>
<td>N2</td>
</tr>
<tr>
<td>w/o sentence compression</td>
<td>0.440</td>
<td>0.185</td>
</tr>
<tr>
<td>w/ sentence compression</td>
<td>0.390</td>
<td>0.174</td>
</tr>
<tr>
<td>OtherSysAve</td>
<td>0.282</td>
<td>0.096</td>
</tr>
</tbody>
</table>

*OtherSysAve: the average scores of all the submitted runs of all the participants*

- Our methods outperformed OtherSysAve on all the scores
- F-measure of Rouge N4 of the method with sentence compression was **the best score**
  - It can generate summaries containing important phrases
Result on Formal Run – Participants Assessment

• Quality question scores

<table>
<thead>
<tr>
<th></th>
<th>Content</th>
<th>Formed</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>X=0</td>
<td>X=2</td>
<td></td>
</tr>
<tr>
<td>w/o sentence compression</td>
<td>0.856</td>
<td>1.134</td>
<td>1.732</td>
</tr>
<tr>
<td>w/ sentence compression</td>
<td>0.788</td>
<td>1.035</td>
<td>1.308</td>
</tr>
<tr>
<td>OtherSysAve</td>
<td>0.423</td>
<td>0.603</td>
<td>1.655</td>
</tr>
</tbody>
</table>

• The method w/o the sentence compression step outperformed OtherSysAve on all the scores

• The formedness score of the method with sentence compression was lower than OtherSysAve

The improvement of the sentence compression step is important future work
Summary

• KitAi-PI: extractive summarization system
  • Automatic training data construction
  • Applying the supervised machine learning method
• The formal run result showed the effectiveness of our method
  • Summaries containing important phrases but ill-formed ones
• The improvement of the sentence compression step is important future work