Generation of Summaries that Appropriately and Adequately Express the Contents of Original Documents Using Word-Association Knowledge

Kazuki Takigawa\textsuperscript{a} Masaki Murata\textsuperscript{b} Masaaki Tsuchida\textsuperscript{c}
Stijn De Seager\textsuperscript{c} Kazuhide Yamamoto\textsuperscript{a} Kentaro Torisawa\textsuperscript{c}

\textsuperscript{a}: Nagaoka University of Technology
\textsuperscript{b}: Tottori University
\textsuperscript{c}: National Institute of Information and Communications Technology (Japan)
Our Goal

Making a **good short summary** from sentences.

---

**Example for our purpose**

A bomb went off.
Some people were killed.
This was triggered by rebel campaign.

**terror**

---

This summary word does not appear in the original document!
Criteria for Our Method

(i) The contents of the original document are associated by the summary.

(ii) The content that is not described in the original document is not associated by the summary.

- We use co-occurrence words as word-association knowledge.
- We paraphrase input by this knowledge, and outputs summarize.
Related Work

- Generally summarize method
  - A part of the original document is extracted.

- Summarization by paraphrasing
  - Kondo et al.[96]
    - Paraphrasing plural verbs into a verb
  - Our Method
    - Handling parts of speech other than verbs
Method

1. Obtaining candidate as summarize
2. Calculating the score for each candidate

Here, "score" indicates the validity that a candidate is a good summary.
The higher score a candidate has, the candidate is likely to be a better summary.

3. Arranging of Candidates
4. Outputting the candidate having the highest score
Obtaining candidate as summarize

1. We obtain all nouns from an input document.
2. We obtain co-occurring words with each of the obtained nouns.
3. Obtained co-occurring words → candidate

※co-occurring words:
   Top 50 nouns having high frequency as the co-occurring words of that noun.
Calculating the score

Assuming that the content of the input document is a set of correct data items…

< criterion (i) >

paraphrasing

This indicates the maximum number of correct data items a candidate can relate to without missing correct data items by using word-association knowledge related words.

similar to Recall

< criterion (ii) >

paraphrasing

This indicates the least number of incorrect data items a candidate can relate to by using related words.

similar to Precision
Calculating the score (2)

\[
Recall(c) = \frac{|RW(c) \cap IW|}{|IW|}
\]

\[
Precision(c) = \frac{|RW(c) \cap (IW \cup (\bigcup_{i \in IW} RW(i)))|}{|RW(c)|}
\]

\[
F\text{-measure } (c) = \frac{2 \times Recall(c) \times Precision_i(c)}{Recall(c) + Precision_i(c)}
\]

\[c\] is a candidate summary,

\[IW\] is a set of nouns obtained from an input document, and

\[RW(x)\] is a set of related words of a word \[x\].
Arranging of Candidates

Method 1: Arranging by Recall(c)
Method 2: Arranging by Precision(c)
Method 3: Arranging by F-measure(c)

※When there are candidates having the same score?

Other score gives priority to high one.
Evaluation Experiment

- We manually created 24 input documents for evaluation.
- The evaluation was performed by a test subject.

**<Evaluation Method>**

(1) Top 1  (2) Top 5  (3) Top 10  (4) MRR

TopX: indicates the ratio when one of the top X candidates is correct

\[
MRR = \frac{\sum_{i=1}^{M} 1/r_i}{M}
\]

- \( r_i \): The highest rank of input: \( i \)
- \( M \): The number of input

\{ strict : correct data \( \rightarrow \) only correct candidates \\
lenient : correct data \( \rightarrow \) candidates similar to a correct candidate \}
### Result

Arrangement in Recall is best method!

<table>
<thead>
<tr>
<th>Used Score</th>
<th>Top1</th>
<th>Top5</th>
<th>Top10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.17</td>
<td>0.29</td>
<td>0.38</td>
<td>0.22</td>
</tr>
<tr>
<td>Precision</td>
<td>0.00</td>
<td>0.17</td>
<td>0.25</td>
<td>0.06</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.08</td>
<td>0.25</td>
<td>0.33</td>
<td>0.16</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Used Score</th>
<th>Top1</th>
<th>Top5</th>
<th>Top10</th>
<th>MRR</th>
</tr>
</thead>
<tbody>
<tr>
<td>Recall</td>
<td>0.33</td>
<td>0.58</td>
<td>0.75</td>
<td>0.45</td>
</tr>
<tr>
<td>Precision₂</td>
<td>0.00</td>
<td>0.25</td>
<td>0.46</td>
<td>0.10</td>
</tr>
<tr>
<td>F-measure</td>
<td>0.17</td>
<td>0.58</td>
<td>0.71</td>
<td>0.34</td>
</tr>
</tbody>
</table>
Example

Input:
プライバシーを守るため、
個人情報を保護するように設定を行った。
(In order to defend privacy, I performed setting so as to protect individual information.)

The top 5 outputted summaries:
1. セキュリティ (security) 2. ヘルプ (help)
3. セキュリティー (security) 4. ポリシー (policy)
5. 保護 (protect)
Method for two-word summary

1. We generate a one-word summary $c_1$ using our method.
2. We make a set of $c_1$ and $ec$.
3. We calculate score of a set of $c_1$ and $ec$.
4. The set having the highest score is outputted.

$ec$: a candidate except $c_1$
Example of two-word Summary

Input:
犯人が銃を撃ち、銃弾が被害者に当たり殺害された。
(A criminal used a gun, and a suffered person was hit by a bullet and killed.)

The outputted summary
Word 1: 発砲 (shooting or fire)
Word 2: 殺害 (be killed)
We proposed a new method for generating summaries that appropriately and adequately express the contents of the original documents using word-association knowledge.

In the proposed method, we used two criteria (precision and recall).
We obtained the best results when using recall.
we described our idea in the case of a two-word summary.