Multi-document Text Summarization: SimWithFirst Based Features and Sentence Co-selection Based Evaluation

Md. Mohsin Ali, Monotosh Kumar Ghosh, and Abdullah-Al-Mamun
Department of Computer Science and Engineering
Khulna University of Engineering & Technology, Khulna – 9203, Bangladesh
e-mail: mohsin_kuet_cse@yahoo.com, mono_cse@yahoo.com, and mamun_kuet04@yahoo.com

Abstract— Document summarization is an emerging technique for understanding the main purpose of any kind of documents. To visualize a large text document within a short duration and small visible area like PDA screen, summarization provides a greater flexibility and convenience. In this paper we study various text summarization techniques e.g. RANDOM, LEAD and MEAD. Then, we propose two techniques for both single and multi document text summarization. One is adding a new feature SimWithFirst (Similarity With First Sentence) with MEAD (Combination of Centroid, Position, and Length Features) called CPSL and another is the combination of LEAD and CPSL called LESM. Finally we simulate and compare the results of new techniques with conventional ones called MEAD with respect to some evaluation techniques. Simulation results demonstrate that CPSL shows better performance for short summarization than MEAD and for remaining cases it is almost similar to MEAD. Furthermore, simulation results demonstrate that LESM also shows better performance for short summarization than MEAD but for remaining cases it does not show better performance than MEAD.

Keywords—CPSL, LESM, MEAD, SimWithFirst, Multi-document Summarization, Sentence Co-selection.

I. INTRODUCTION

Text summarization is the process of creating a summary of one or more text documents. That is, we may summarize a large amount of news from different sources [1]. The technique of automatic text summarization is indispensable for dealing with summarization problem. Many summarization techniques and their evaluation methods have been developed for this purpose. Such techniques are RANDOM [3], LEAD [3], MEAD [4] etc. which are used to generate the summary. MEAD is the recent toolkit for summarization.

This paper presents two methods that incorporate new features based on ‘the similarity with first’ to improve the summarization of multiple documents as well as single document.

This paper is organized as follows: Section II discusses the various existing techniques of document summarization. Section III discusses two proposed methods of text summarization. Section IV describes the experimental results of different summarization techniques using different evaluation techniques. Finally, Section V presents the conclusions of this paper.

II. EXISTING TECHNIQUES OF TEXT SUMMARIZATION

A. RANDOM based

The RANDOM based technique [3] is the simplest of all the other as it randomly selects lines from the source document, depending upon the compression percentage and put them inside the summary. In this technique, a random value between 0 and 1 is assigned to each sentence of the document. A threshold value for length of the sentence is provided. We will assign a score of 0 to all sentences that do not meet this length cutoff. Finally we choose required sentences according to assigned highest score for desired summary.

B. LEAD based

LEAD based technique [3] is a technique in which first or first and last sentence of the paragraph are chosen depending upon the compression rate (CR) and it is very good for news articles as they have the main theme set in the first lines of the articles. So, it can be reasonable that n% sentences are chosen from beginning of the text [5] e.g. selecting the first sentence of each document, then the second sentence of each, etc. until the desired summary is constructed. This method is called LEAD based method for summarization. In this technique we assign a score of 1/n to each sentence, where n is the sentence number in the corresponding document file. This means that the first sentence in each document will have the same scores; the second sentence in each document will have the same scores, etc. We also provide a threshold value for sentence’s length [5]. The sentences with lengths less than the specified value are thrown out.

C. MEAD based

MEAD is a centroid-based extractive summarizer that scores sentences based on sentence-level and inter-sentence features which indicate the quality of the sentence as a summary sentence [4]. It then chooses the top-ranked sentences for inclusion in the output summary. MEAD extractive summaries score sentences according to certain sentence features - Centroid [6], Position [6], and Length [5]. In this technique the score of a sentence is calculated using the following formula [5].
The sentence's score is computed as follows [1] [5].

\[
scord(S_j) = \begin{cases} 
\sum (w_c * C_i + w_p * P_i) & \text{If } \text{Length}(S_i) > \text{Threshold} \\
0 & \text{If } \text{Length}(S_i) < \text{Threshold} 
\end{cases}
\]

Here, \( w_c \) = The weight for the Centroid feature.  
\( w_p \) = The weight for the Position feature.  
\( C_i \) = The calculated Centroid value for \( i \)th sentence.  
\( P_i \) = The calculated Position value for \( i \)th sentence.  
\( S_i \) = The \( i \)th sentence of the document.  
\( i \) = Sentence number within the cluster (1 \( \leq \) \( i \) \( \leq \) \( n \)).  
\( n \) = Number of sentences in a single or multiple text documents.

The highest value of the scored sentence is taken in the extract file. Thus the MEAD based summary is generated. The default weights for Centroid and Position are both 1. The default Length cutoff is 9. We use the default value in our experiment.

III. PROPOSED METHODS OF TEXT SUMMARIZATION

A. CPSL technique

CPSL technique is the combination of SimWithFirst feature with MEAD. MEAD is described in the previous sections. Now, SimWithFirst feature will be discussed here. It is also known as First-sentence overlap [1].

Normally the first sentence is very important in any document. So we can take the first sentence as a main topic of the summary. Then we compute the similarity score of every sentence with the first sentence and take the highest score as the most similar sentence with the first. The overlap value is computed as the inner product of the sentence vectors for the current sentence \( i \) and the first sentence of the document. Thus the cosine similarity between a sentence at position \( i \) and the first sentence in the document is calculated as follows [1] [5].

\[
F_i = \overline{S_i} \overline{S_1}
\]

So, SimWithFirst computes the cosine similarity between a sentence \( S_i \) and the first sentence \( S_1 \) in the document and we combine this with MEAD. MEAD now decides which sentence to include in the summary on the basis of sentence’s score. The sentence’s score is computed as follows [1] [5].

\[
\text{SCORE}(s) = \sum (w_c * C_i + w_p * P_i + w_f * F_i)
\]

Where, \( w_f \) is the weight for the SimWithFirst (First-sentence overlap) feature which is set to 1 by default (but in MEAD it is 0) and \( i \) is same as that of MEAD. Then the highest scored sentences are included into summary according to the predefined CR.

B. LSEM technique

LSEM technique is the combination of LEAD and CPSL. Initially, we compute summary according to LEAD and CPSL techniques. Then, we extract only common sentences from these two summaries. After that, we apply LEAD based technique on the unmatched sentences of the summary to obtain the summary of desired percentage. We sort the unmatched sentences and then take the first unmatched sentence of the first document and check the desired percentage; if desired percentage is not achieved, we take the first unmatched sentence of the second document. If desired percentage is not achieved after taking all first unmatched sentences of the documents, we take last unmatched sentence of the first document and again last unmatched sentence of the second document and so on. After taking all last sentences of all documents we take the second sentence of the first document and thus proceed. Here, we take first and last sentence of each document, because the consecutive sentence may indicate similar subject. So, we try to take the sentences from the unmatched sentences which refer to the various subjects. Moreover, last sentence of a document may the conclusion of the document and it may focus the future guideline according to focus on the present situation. So, we think the last sentence may be important. As we sort the unmatched sentences so that we can achieve this technique by taking first and last sentence in turn.

IV. EXPERIMENTAL RESULTS

There are three general types of evaluation measures: (1) co-selection, (2) content-based similarity and (3) relevance correlation. Co-selection measures include precision and recall [4] of co-selected sentences, relative utility [4] and kappa [2]. Co-selection methods have some restrictions: they only work for extractive summarizers. Two manual summaries of the same input do not in general share many identical sentences. The weakness of co-selection measures with several content-based similarity measures are addressed in [4]. The similarity measures that are used are word overlap, longest common subsequence and cosine. The simulation is carried out in MEAD Toolkit-3.11 [5] with PERL language in LINUX environment on DUC 2004 experimental datasets [1].

Precision and recall is defined as follows [4]:

\[
P_{J_2}(J1) = \frac{A}{A + C}
\]

\[
R_{J_2}(J1) = \frac{A}{A + B}
\]

In our experiment, each set of documents which is compared has the same number of sentences and also the same number of sentences are extracted; thus \( P = R \).

We obtain the graph of Fig. 1 for five summarization techniques for various CRs using precision and recall method.
Fig. 1 shows that RANDOM technique is the worst technique. Below 30% and above 60% CR LEAD is the best technique. But, for remaining CRs it is lower than MEAD, CPSL and LESM; because it selects the first sentence and last sentence until desired summary generated. Below 30% CR MEAD is not better than LEAD, but above 30% CR MEAD is better than LEAD. Below 25% CR CPSL is better than MEAD, but above 25% CR it is not better than MEAD. Below 25% CR LESM is better than all other techniques except LEAD, but above 25% CR LESM is not better than MEAD and CPSL, because it is the combination of LEAD and CPSL. As MEAD is better for short summarization and good for average summarization than all other technique, so on an average this kappa technique evaluates MEAD as better solution for summarization. Without loss of generality, suppose that three judges are asked to build extracts of a single article 4 [1]. As an example, Table I shows the weights of the different sentences (note that no CR needs to be specified; from the data in the table, one can generate summaries at arbitrary CRs).

The inter-judge agreement measures, to what extent each judge satisfies the utility of the other judges by picking the right sentences [1]. In the example, with 50% summary, Judge 1 would pick sentences 1 and 2 because they have the maximum utility as far as he is concerned. Judge 2 would select the same two sentences, while Judge 3 would pick 2 and 4. The maximum utilities for each judge are 18 (= 10 + 8), 19 and 17 respectively. How well Judge 1’s utility assignment satisfies Judge 2’s utility need? Since they have both selected the same sentences, Judge 1 achieves 19/19 (1.00) of Judge 2’s utility. However, Judge 1 only achieves 13/17 (0.765) of Judge 3’s utility. We can therefore represent the cross-judge utility agreement $J_{i,j}$ as an asymmetric matrix (e.g., the value of $J_{1,2}$ is 0.765 while the value of $J_{2,1}$ is 13/18 or 0.722). The values $J_{i,j}$ of the cross-judge utility matrix for $r = 50\%$ are shown in Table II [1].
TABLE I. Illustrative example

<table>
<thead>
<tr>
<th></th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sentence 1</td>
<td>10</td>
<td>10</td>
<td>5</td>
</tr>
<tr>
<td>Sentence 2</td>
<td>8</td>
<td>9</td>
<td>8</td>
</tr>
<tr>
<td>Sentence 3</td>
<td>2</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>Sentence 4</td>
<td>5</td>
<td>6</td>
<td>9</td>
</tr>
</tbody>
</table>

TABLE II. Cross-judge utility agreement (J)

<table>
<thead>
<tr>
<th></th>
<th>Judge 1</th>
<th>Judge 2</th>
<th>Judge 3</th>
<th>Overall</th>
</tr>
</thead>
<tbody>
<tr>
<td>Judge 1</td>
<td>1.000</td>
<td>1.000</td>
<td>0.765</td>
<td>0.883</td>
</tr>
<tr>
<td>Judge 2</td>
<td>1.000</td>
<td>0.765</td>
<td>1.000</td>
<td>0.883</td>
</tr>
<tr>
<td>Judge 3</td>
<td>0.722</td>
<td>0.789</td>
<td>1.000</td>
<td>0.756</td>
</tr>
</tbody>
</table>

Figure 3. Comparison graphs of five summarization techniques using relative utility method

The performance of each judge (J_i) against all other judges by averaging for each Judge i all values in the matrix J_{ij} where i ≠ j are computed in [1]. These numbers indicate that Judge 3 is the outlier. Finally, the mean cross-judge agreement J is the average of J_{i,j} for i = 1...3. In the example, J = 0.841. J is like an upper bound on the performance of a summarizer (it can achieve a score higher than J only when it can do a better job than the judges).

We evaluate the summarization techniques using the modern evaluation method called relative utility and the results of this experiment are shown in Fig. 3. It shows that RANDOM technique is the worst technique because it selects the sentences randomly from the documents. Below 20% CR LEAD is the best technique and above 20% CR MEAD is not better, because it selects the first sentence, second sentence and so on. Below 30% CR MEAD is not better than LEAD, but above 30% CR MEAD is better than LEAD. So, on an average MEAD is better than LEAD. Again, below 35% CR CPSL is better than MEAD and above 35% CR CPSL is almost equal to MEAD. Below 20% CR LESM is better than CPSL, but above 20% CR it is not better than CPSL, because it is the combination of LEAD and CPSL. So, below 35% CR, the performance of CPSL is better than MEAD and above 35% CR it is almost equal to MEAD. So, on an average, the performance of CPSL is better than all other techniques for text summarization with respect to relative utility.

V. CONCLUSIONS

In this paper we have studied and analyzed RANDOM, MEAD, and LEAD based techniques for multi-document text summarization. Then, we have proposed two new techniques CPSL and LESM. Experimental results demonstrate that the kappa evaluation method supports MEAD technique for average text summarization. But precision and recall and relative utility based evaluation methods support our proposed CPSL technique as a best technique for multi-document texts summarization on an average CR. CPSL is better than MEAD for short summarization and above 35% CR the performance of CPSL is almost similar to MEAD. So, we conclude that CPSL is the best technique for multi-document text summarization in average CR and it is also applicable for single document text summarization. Although LESM is best for short summarization, but CPSL is better than LESM. The reason behind this that LESM includes LEAD which shows worst performance on an average CR due to its bad feature of selecting only the first or first and last sentences of the documents.

It can be further improved by adding other new features. Moreover, CPSL technique can be applied for (1) Bangla language and multimedia summarization, (2) Human speech summarization and (3) Class lecture material preparation according to student’s quality.

REFERENCES