Hybrid Text Summarization Method based on the TF Method and the LEAD Method

Kai ISHIKAWA   Shinichi ANDO   Akitoshi OKUMURA
C&C Media Research, NEC Corporation
4-1-1 Miyazaki Miyamae-ku Kawasaki-shi Kanagawa 216-8555, Japan
{ishikawa, ando, okumura}@ccm.cl.nec.co.jp

Abstract

This paper describes a hybrid text summarization method based on a TF-based sentence extraction method and a LEAD sentence extraction method. The LEAD method is known to be effective than other methods for document summarization of newspapers in lower summarization (output-to-input) ratio. In order to combine the LEAD method with the TF method, we used a rectangular distribution function that determines the importance of sentences according to their position in a document. With our method, the importance of a sentence is determined by multiplying the TF-based score and the distribution function. We conducted open test evaluation using the formal run test data of sentence extraction sub-task in NTCIR-2 Workshop TSC task (30 newspaper articles). The proposed method was tested by the average values of F-measure for 10%, 30%, and 50% summaries, and proved 34.1% for TF method, 39.1% for LEAD method, and 42.4% for the proposed method.

Keywords: *TF, LEAD, headline, hybrid, position, distribution.*

1 Introduction

Currently, most sentence extraction methods used for automatic text summarization are based on the calculation of sentence importance. Sentences are ranked according to importance values, and the upper-ranked sentences are extracted and used to compose a summary. In other words, sentence extraction for automatic text summarization can be derived from evaluating the importance value of sentences within a document. Okumura, et al. suggests that the following seven elements are useful in calculating the importance of a sentence within an article [4].
(1) Frequency of keyword appearance in an article.
(2) Position of a sentence in an article or in a paragraph.
(3) Title or headline of an article.
(4) Text structure based on the relationship between sentences.
(5) Key expressions that appear in an article.
(6) Relationships between sentences or words in an article.
(7) Similarities between sentences in an article.
The TF method [2] is an example of how information such as that from (1) above can be utilized, and is the earliest known method to be used for automatic text summarization since research began in this area. The LEAD method is an example of how information such as that from (2) above can be used, and is known to be particularly effective for summarization of newspaper articles. Headline information from journalistic articles falls into category (3). For the sentence extraction task (TSC) given at NTCIR-2 Workshop, the Mainichi newspaper articles were assigned as input data for sentence extraction. The combination of different summarization methods or features is one of most important subjects in recent research on summarization [1]. In this paper, we describe a hybrid text summarization method based on both the TF method and the LEAD method. Information from (1) through (3) above is utilized in this method.

2 Hybrid text summarization method

2.1 Basic TF-based importance weight

This section describes three methods that are used to calculate sentence importance: the TF method, the LEAD method, and our hybrid method. First, we will describe how the sentence importance value is calculated using the TF method. With the TF method, the importance value IW(s) of a sentence is given as follows:

$$IW_{TF}(s) = \sum_{\{t\} \in s} f(t) \qquad (1)$$

Here, $\{t\} \in s$ refers to the set of terms in a sentence s, and $f(t)$ refers to the frequency that a term appears in an article. The importance value of a term t is given by $tf = f(t)$, and the importance of a sentence is determined by the summation of the importance value of each term in the sentence. This calculation usually utilizes content words or keywords as the set of terms.

2.2 Utilization of headline information

Based on this TF method in determining the importance weight of a sentence, we took the title or headline information into consideration. It is plausible to think that terms appearing in a title or a headline are highly important. The importance weight based on this hypothesis $IW_{Head-TF}(s)$ can be given as follows:

$$IW_{Head-TF}(s) = \sum_{\{t\} \in s} \alpha(t) \cdot f(t), \tag{2}$$

where,

$$\alpha(t) = \begin{cases} A & \text{if } t \in headline \\ 1 & \text{otherwise} \end{cases} .$$

Here, the importance weight of a term is given by $A \cdot f(t)$, where A is a real number greater than 1 when the term appears in the headline; if the term does not appear in the headline, the weight is given by $f(t)$.

2.3 Combination with the LEAD method

Third, we combined the LEAD method with the importance value based on the TF method and the headline information $IW_{Head-TF}(s)$. The LEAD method is a method used for determining important sentences by extracting the leading sentences in a text. This method is known for its effectiveness in summarizing newspaper articles because important sentences tend to appear in the first few sentences of a newspaper article.

To combine this LEAD method with the importance value of a sentence based on the TF method, we used the following importance value $IW_{Proposed}(s, i)$ for a sentence s and its position i ($= 1, 2, \ldots$) in the article:

$$IW_{Proposed}(s, i) = \beta(i) \cdot IW_{Head-TF}(s), \tag{3}$$

where,

$$\beta(i) = \begin{cases} B & \text{if } 1 \leq i \leq N \\ 1 & \text{if } i > N \end{cases} .$$

Here, $\beta(i)$ is a rectangular function that models the distribution of important sentences according to their position in the article. B is a real number greater than 1.

3 Evaluation

3.1 Sentence extraction for summary

For the sentence extraction task presented at NTCIR Workshop 2 (Text Summarization Challange), the Mainichi newspaper articles were assigned as input data for sentence extraction. The input data consists of headline and a body of text, with sentence and paragraph separators attached.

For the NTCIR-2 TSC task, the evaluation of automatic text summarization results was carried out twice: once in a dry run, and once in a formal run. Both runs were conducted in the same manner. The test set for both runs was composed of 30 newspaper articles, and when performing the task for each article, the summary output was required to be in three different summarization ratios: 10%, 30%, and 50%. In this paper, we evaluate our system based on the test set used for the sentence extraction task (TSC) of NTCIR-2 Workshop. The test set article data and summarization results for each task were provided to the participants by NTCIR-2 Workshop.

3.2 Proposed and baseline methods

To determine the efficiency of our hybrid method, we evaluated the summarization results using the following five methods:

TF The TF method using the importance value in Equation (1).

Head-TF The TF method with headline information using the importance value in Equation (2), where the parameter $A = 20$ is used.

Proposed The proposed hybrid method using the importance value in Equation (3), where the parameters $A = 20$, $B = 10$, and $N = 3$ are used.

Hyb-LEAD The hybrid LEAD method (in contrast to the **Proposed** hybrid method) using LEAD for N sentences and Head-TF for other sentences up to threshold, where the parameters $A = 20$ and $N = 3$ are used.

LEAD The LEAD method that extract leading sentences up to the given threshold.

Here, in the evaluation of importance weight, words from a certain part of speech, i.e. common nouns, proper nouns, and Sa-Hen (nouns), were used as a set of terms. The parameters $A = 20$, $B = 10$, and $N = 3$ that were used were heuristically defined according to the average of the F-measure values of the summarization results for the 10%, 30%, and 50% ratios.

We used the following three steps to perform important sentence extraction for evaluation based on the previous five methods;

Step 1 Obtain word sequence with part of speech tags for each sentence in the input newspaper articles by automatic morphological analysis.

Step 2 Calculate the importance weight IW for each input sentence.

Step 3 Rank sentences according to their importance values IW. Upper-ranked sentences are extracted under the condition of the summary ratio and composed into a summary.

3.3 Evaluation with F-measure

We compared summaries obtained by previous five methods, i.e., TF, Head-TF, Proposed, Hyb-LEAD, and LEAD, to show the efficiency of our proposed method. To measure summary accuracy, we used the following F-measure;

$$F = \frac{1}{\alpha \cdot \frac{1}{P} + (1-\alpha) \cdot \frac{1}{R}} \tag{4}$$

Here, the parameter α is a real number $(0 < \alpha < 1)$, P denotes precision, and R denotes recall. Here, the precision is given by $P =$ number of correct sentences in generated summary / total number of sentences in generated summary, and the recall is given by $R =$ number of correct sentences in generated summary / total number of sentences in answer summary. Here, the relation $R = P = F$ holds true, because the number of sentences to be extracted from each article is equivalent to the number of sentences in the answer summary in this task. Consequently we will only discuss the F-measure in the subsequent discussions.

In Table 1, the F-measure values of the summaries obtained in the dry-run test set are shown. Each row shows results of importance weight calculation method for the five methods and each column refers to the summarization ratios of 10%, 30%, and 50%. The numerical values show the averages and their standard deviations of the F-measures for the summaries generated for the 30 articles.

In the table, we find that the Proposed and the Hyb-LEAD methods, both of which are hybrid methods based on the TF and LEAD, are superior to the other methods, i.e. TF, Head-TF, and LEAD for all summarization ratios. The F-measure values of the TF method and the HEAD-TF method for 10% summary are 0.190 and 0.291. This suggests that terms appearing in a headline are effective in sentence score calculation. However, the values are both less than that of the LEAD method 0.417. The TF method alone and the HEAD-TF method are inferior to the LEAD method for the lower summarization ratio. On the contrary, the Head-TF method is superior to the LEAD method for the summarization ratios of 30% and 50%.

Table 1. F-measure values of the dry-run test set.

	Summarization ratio		
	10%	30%	50%
TF	0.190 ±0.215	0.485 ±0.143	0.743 ±0.087
Head-TF	0.291 ±0.267	0.538 ±0.121	0.764 ±0.079
Proposed	**0.446** ±0.268	**0.569** ±0.138	**0.767** ±0.078
Hyb-LEAD	0.442 ±0.271	0.571 ±0.137	0.770 ±0.079
LEAD	0.417 ±0.226	0.510 ±0.138	0.749 ±0.118

Table 2. F-measure values of the formal-run test set.

	Summarization ratio		
	10%	30%	50%
TF	0.119 ±0.178	0.353 ±0.131	0.551 ±0.092
Head-TF	0.095 ±0.141	0.405 ±0.127	0.573 ±0.102
Proposed	**0.251** ±0.283	**0.447** ±0.136	**0.574** ±0.110
Hyb-LEAD	0.252 ±0.279	0.445 ±0.139	0.569 ±0.119
LEAD	0.276 ±0.310	0.367 ±0.198	0.530 ±0.110

These results provide a qualitative explanation of why the methods Proposed and Hyb-LEAD, which are both hybrid methods based on the TF method with headline information and the LEAD method, are superior to other methods for all of the summarization ratios. To obtain summary with lower ratio, the importance weight provided by LEAD-based method takes precedence over the one by TF-based effect, while for summary with higher ratio, the importance weight given by TF-based method is given priority over the other. The proposed hybrid method makes full use of advantages by both sides.

In comparison to the other methods, the difference between the F-measure values of Proposed and Hyb-LEAD is very small. The Proposed method is based on the TF method, in which the leading $N = 3$ sentences are not always extracted according to the TF-based importance weight. On the other hand, the Hyb-LEAD method is based on the LEAD method in which the leading $N = 3$ sentences are always extracted without restriction. The small difference in the F-measure values of these two methods was due to the low accuracy of the TF-based importance weight.

The F-measure values of the summaries obtained in the formal-run test set are shown in Table 2. Just as with the results for the dry-run test set in Table 1, the averages and standard deviations of the F-measures for the 30 summaries generated are shown here in Table 2. Compared to the F-measure values for the dry run (see Table 1), the overall values are lower. However, the order of methods according to the F-measure values is unvarying, and the advantage of combined methods, Proposed and Hyb-LEAD, is apparent. Looking at the results for the summarization ratio of 10%, the LEAD method proved to be the most effective of all, and the F-measure values of the two hybrid methods are much smaller. This result seems to stem from the low value obtained by the Head-TF method, which is TF-based and uses headline information. While the Head-TF method is superior to the TF method in the dry run test set for all of the summarization ratios, the Head-TF method is inferior to the TF method in the formal run test set for the 10% ratio, despite its good results for the 30% and 50% ratios. This suggests that the headline description style may have been different for the dry run and formal run test sets.

3.4 Evaluation with random baseline

Here, we discuss the causes of smaller F-measure values as a whole for the formal run set compared to that of the dry run set. The outcome suggests that the articles in the formal set are more complex to be summarized either by the TF method, the LEAD method, or the Proposed method. The ratio of articles that cannot be readily summarized is likely to be higher in the formal run set.

To analyze the difference between the dry run and the formal run results, we compared the F-measure values of random extraction for dry run and formal run test sets. The F-measure value of summary by random extraction can be obtained by theoretical calculation. When we have N sentences in an article and Np (positive integer) of them are important sentences, the number of important sentences extracted by random extraction of n sentences results to a hyper-geometric distribution $HG(n, p; N)$. Here $p(0 < p < 1)$ is a rational number corresponding to a summarization ratio. We use $n = Np$ because the number of important sentences in answer summary and the one extracted by automatic summarization are the same.

The probability $f(k|n, p, N)$, where k is the number of correct sentences in randomly extracted n sentences satisfying $\tilde{k} \sim HG(n, p; N)$, is given as follows:

$$f(k|n,p,N) = \frac{\begin{pmatrix} Np \\ k \end{pmatrix}\begin{pmatrix} N(1-p) \\ n-k \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{5}$$

Table 3. Test set statistics of the dry-run test set.

$\bar{N} = 23.87$	Summarization ratio		
	10%	30%	50%
True ratio $\bar{p}$	0.151	0.439	0.724
# of sentence $\bar{n}$	3.60	10.37	16.97
$E(\tilde{F}_{Random})$	0.151	0.439	0.724
$\sigma(\tilde{F}_{Random})$	0.178	0.118	0.060

Table 4. Test set statistics of the formal-run test set.

$\bar{N} = 33.10$	Summarization ratio		
	10%	30%	50%
True ratio $\bar{p}$	0.105	0.315	0.536
# of sentence $\bar{n}$	3.40	10.03	16.93
$E(\tilde{F}_{Random})$	0.105	0.315	0.536
$\sigma(\tilde{F}_{Random})$	0.160	0.124	0.086

The expectation $E(\tilde{F})$ and the variance $V(\tilde{F})$ of F-measure ($F = \frac{k}{n}$) are obtained as follows:

$$E(\tilde{F}) = \sum_{k=0}^{n} \frac{k}{n} \cdot f(k|n,p,N) = p \tag{6}$$

$$V(\tilde{F}) = \sum_{k=0}^{n} \left(\frac{k}{n} - E(\tilde{F}) \right)^2 \cdot f(k|n,p,N) \tag{7}$$

$$= \frac{p(1-p)}{n} \cdot \frac{N-n}{N-1} \tag{8}$$

To obtain these F-measure values defined in the above formulas, we utilized the statistic values $\bar{N}$, $\bar{n}$ and $\bar{p}$ of dry run and formal run test sets shown in tables 3 and 4. The expectation values $E(\tilde{F}_{Random})$ and the standard deviations $\sigma(\tilde{F}_{Random}) = \sqrt{V(\tilde{F}_{Random})}$ for random extraction are obtained using the formulas (6)-(8).

Comparing with the values in tables 1 and 2, we found again F-measure values as a whole for the formal run set smaller compared to the dry run set. The difference in F-measure values between the formal run set and the dry run set seems to be caused by the difference in true summarization ratio between them. We evaluated true gain with each extraction method comparing with that of random extraction.

The following difference of the averaged k (the number of correctly extracted sentences) for each extraction method and the random extraction are compared in tables 5 and 6.

$$\Delta\bar{k} = \bar{k} - \bar{k}_{Random} = \bar{n} \cdot \bar{F} - \bar{n} \cdot E(\tilde{F}_{Random}) \tag{9}$$

Table 5. Gain of $\bar{k}$ for the formal-run test set.

$\bar{N} = 23.87$	Summarization ratio		
	10%	30%	50%
$\bar{p}$	0.151	0.439	0.724
$\bar{n}$	3.60	10.37	16.97
$\bar{k}_{Random}$	0.544	4.55	12.29
$\Delta\bar{k}_{TF}$	0.140	0.477	0.322
$\Delta\bar{k}_{Head-TF}$	0.504	1.027	0.679
$\mathbf{\Delta\bar{k}_{Proposed}}$	**1.062**	**1.348**	**0.730**
$\Delta\bar{k}_{Hyb-LEAD}$	1.048	1.369	0.781
$\Delta\bar{k}_{LEAD}$	0.958	0.736	0.424

Table 6. Gain of $\bar{k}$ for the formal-run test set.

$\bar{N} = 33.10$	Summarization ratio		
	10%	30%	50%
$\bar{p}$	0.105	0.315	0.536
$\bar{n}$	3.40	10.03	16.93
$\bar{k}_{Random}$	0.357	3.159	9.074
$\Delta\bar{k}_{TF}$	0.048	0.381	0.254
$\Delta\bar{k}_{Head-TF}$	-0.034	0.903	0.626
$\mathbf{\Delta\bar{k}_{Proposed}}$	**0.496**	**1.324**	**0.643**
$\Delta\bar{k}_{Hyb-LEAD}$	0.500	1.304	0.559
$\Delta\bar{k}_{LEAD}$	0.581	0.522	-0.102

Here, $\bar{k}$ and $\bar{k}_{Random}$ are the averaged numbers of correctly extracted sentences by each method and by random extraction, so the difference $\Delta\bar{k}$ refers to the true gain of each method. The difference 1.000 means that the method is better than random extraction by one correct sentence gained in each summary. A negative value means that the method is statistically equal to or even worse than the random extraction.

Compared to the values for dry-run set, the values are lower as a whole for formal-run set. Especially, utilization of headline information in 10% summary and the LEAD method in 50 % summary don't seem to be effective.

3.5 Further discussion

Here, we use the article (DOCNO:980208039) as an example of documents that cannot be readily summarized. With the proposed method, F-measure values of 10%, 30%, 50% summaries were 0.000, 0.333, 0.400, which are the lowest in the formal run set. The expectation values of F-measure with random extraction, that are equal to the summarization ratios, are 0.1, 0.3, 0.5 for 10%, 30%, 50% summaries. For this, we can conclude that the proposed method has no significant effect on summarization of this particular article.

The article is an editorial on political tug of war between the ruling and opposition parties concerning the Lower House Election. Comparing with the answer summaries, the leading three sentences don't appear in the 10% and 30% summary. Only one sentence appears in the 50% summary. The LEAD method is not effective for this kind of articles.

From the headline ”[社説 (editorial)] 衆院補選 (Lower House by-election) 党勢拡大の機会を生かせ (Make full use of the opportunity to enhance the party prestige)”, the terms ”衆院 (Lower House)”, ”補選 (by-election)”, ”党勢 (the party prestige)”, ”拡大 (enhance)”, and ”機会 (opportunity)” are extracted and considered as more (20 times) important in the TF-based sentence score calculation. However, none of the words appeared in the 10% answer summary, and only the terms ” 衆院 (Lower House)” and ”補選 (by-election)” appeared in sequence in two sentences of the 30%, 50% summaries.

The three sentences of the 10% answer summary mean almost the same as the one expressed in the headlines, however the words used are not the same. When we determine the key sentences of the article based on the headline, we need to analyze the meaning of each word and ”read between the lines”. These are the difficulties which we encounter when summarizing editorial articles.

4 Conclusion

In this paper, we described a hybrid text summarization method based on the TF method and the LEAD method, and we conducted open test evaluation using the formal run test data of sentence extraction sub-task in NTCIR-2 Workshop text summarization task TSC (30 newspaper articles). The proposed method was tested by the average values of F-measure for 10%, 30%, and 50% summaries, and proved 34.1% for the TF method, 39.1% for LEAD method, and 42.4% for the proposed method.

On the other hand, extraction using the TF method, the LEAD method and the hybrid method, all which apply surface information are not so effective for some types of input article text, as discussed in the evaluation section. One solution to increase the strength of summary extraction is not to fix the combination of these methods as stated in this paper, but rather to dynamically select the optimum combination of method according to context and types of target text that are automatically determined.

In extracting sentences out of topical documents such as news articles, the 5W1H Information [3] can be combined with surface information which was exclusively utilized by the methods stated in this paper. For instance, we feed the headline to extract its 5W1H information and measure its correlation with 5W1Hs within the article's body and analyze their similarity.

The result can help the extraction of key sentence according to the meanings of each text line. In calculating the relations to match the factors of 5W1Hs, semantic concordance and distance between synonym groups also become important. It's also critical to build lexicon that covers synonymous words in specific domain.

Automatic assessment of input text type, dynamic switching of sentence extraction methods and practical application of 5W1H information all remain to be our future topics.

References

[1] C. Aone, M. E. Okurowski, and J. Gorlinsky. Scalable summarization using robust nlp and machine learning, 1998. In proceddings COLING-ACL'98.

[2] H. P. Luhn. The automatic creation of literature abstracts. *IBM Journal of Research and Development*, 2(2):159–165, 1958.

[3] A. Okumura, T. Ikeda, and K. Muraki. Text summarization based on information extraction and categorization using 5w1h. *Journal of Natural Language Processing*, 6(6):27–44, 1999.

[4] M. Okumura and H. Nanba. Automated text summarization: A survey. *Journal of Natural Language Processing*, 6(6):1 – 26, 1999.