Selection Informative Units for Extractive Summarization

METİN TURAN
Computer Engineering Department,
İstanbul Ticaret University,
Küçükyalı, İstanbul,
TURKEY

Abstract: - An Extractive Multi-Document Summarizer must select the most informative units and prevents duplication in extraction. In order to achieve this goal, a new technique, called "comprising at least one Representative Term at the Highest Frequency", called RTHF, is proposed in this work. The units which include representative terms, but with low frequencies are not considered for extraction (selection of the most informative units). On the other hand, these units which provide RTHF feature, precede other similar units in ranking (prevents duplication). The heuristic behind the RTHF is explained by probability. RTHF was experimented on a previously developed and tested paragraph- based Extractive Multi-Document Summarizer. The results show that it enhances the original system by 0.8% ~ 3.2% (Average-F values of ROUGE metrics).

Key-Words: - Document Summarization, Informative Units, TF-IDF, Paragraph Extraction, NLP, AI

Received: June 28, 2022. Revised: February 21, 2023. Accepted: March 8, 2023. Published: March 23, 2023.

1 Introduction

As Automatic multi-document summarization is a job of producing a summary from the bulk of documents. Although this is a result of the rapidly increasing amount of documents in public, objective is to produce summaries automatically which is more similar to the job done by human Summarizer fundamentally. A survey, including Extractive Multi-Document Summarizer (EMDS) approaches can be found in the article written by Kumar and Salim [1] or M.Sc. Thesis of Sizov [2].

A document is composed of small units such as sentences, paragraphs or text segments. Sentence is the most common unit in summary because it provides easy parsing and processing. Researchers have been suggested different techniques [3, 4] in order to select more relevant sentences.

There are comparatively a few studies that focus on the extraction of paragraphs in EMDS. The well-known research was done by Mitra and colleagues [5]. The latest system has been developed in a doctorate thesis [6]. The result of latter work highlights that paragraph-based summary can be effective as much as sentence based summary.

Document units can be identified by text features. Researches have been focused on discovering new text features. Pioneer of researchers was Edmundson [7] suggested three additional features (cue, title, location) to evaluate the sentence weights more accurately. After a long time, Kupiec proposed a system [8] which was based on the probability of features in text, such that Sentence Length Cutoff Feature, Fixed-Phrase Feature, Paragraph Feature, Thematic Word Feature, and Uppercase Word Feature. Another important work is done by Kumar and his colleagues [9] this decade. They calculated sentence popularity using word features such that cue words, stigma words and keywords. One of the important researches was done by Suanmali [10] who proposed a fuzzy system to score sentences using some features had been suggested in the literature (proper noun, thematic word, numerical data). Finally, Gupta and Lehal work is an example of the latest researches used feature based approach [11]. They investigated text mining technologies and exemplified applications in broad range.

Machine learning has been adopted to identify weights of sentences to be selected for a summary recently. For example, Binwahlan [12] used PSO technique in 2009. Pairs of documents and summaries were used for training in this technique. The other similar work was done by Bossard and Rodriges [13]. They used a genetic algorithm to determine the best weights for the features. Manne and Fatima [14] also suggested an HMM tagger to improve the quality of the summary by feature term identification.

The feature based technique is simple, however, it doesn't explain how the terms are related or they disperse through documents. Li and colleagues [15] used lexical chains and suggested a keyword

extraction algorithm, so that the shortcoming of the TF-IDF is partially prevented. A few techniques in the literature are also suggested to obtain representative terms (dispersion considered) instead of using all vocabulary exist in the document/s. The first approach [16] refers to the Helmholtz principle in Gestalt theory of physics and obtains a statistical value for each term. The terms above a threshold value are confirmed as representative terms. The second approach [17] is based on "inverse document frequency" (TF-IDF) values of terms. Furthermore, Litvak and Last's work [18] is an example of single document summary which uses a graph based approach to obtain representative words.

The other important technique studied is comparing the document structure. Marcu [19] proposed a method which captures the rhetorical structure of a document. It depends on a set of constraints and assumes text coherence. The rhetorical structure composition is also applied to the multi-documents by Yong-dong and colleagues [20]. Another work is done by Salton [21], who suggested paragraph-based extraction using the intra-document links between paragraphs. A text relationship map is finally produced. Okazaki [22] also proposed a similar approach applied to sentences interrelationships.

The purpose of this study is to devise a new technique in order to select more informative paragraphs through similar ones, so that minimizing information duplication and enhancing summary quality even for higher compression rates. The devised technique is called RTHF (comprising at least one Representative Term at the Highest Frequency). RTHF assures that a unit contains at least one representative term which frequency in this unit is the greatest for all units in the document/s. Moreover, if a unit includes lots of representative terms with low frequencies, it is accepted as garbage. By the way, this unit isn't considered for summarization anyway. Finally, RTHF units are ranked in order and extracted in sequence until the summary size is obtained.

Automatic multi-document summarization is actually a complex task requires both detection of the related segments in documents and selection of the more informative ones for extraction. Moreover, in which order the extracted segments should be presented is another issue. In this work, the successful paragraph based work [6] is extended to use representative term frequencies in order to select more informative paragraphs, called RTHF. The advantages of technique are being very simple (only vector operations) and applicable to any EMDS. Using only representative term frequencies is just enough to complete other tasks in an EMDS. Selection of informative segments has been worked. Furthermore, it can also be adapted to order segments in extraction by sorting the frequencies of representative words in selected segments (further work).

RTHF is a heuristic technique and announced first time in the literature. It was experimented on the paragraph-based EMDS [6]. Similar data set (DUC 2006) was used in experiments. Eventually, final ROUGE metrics were compared and discussed with the values announced in [6]. This study shows that using RTHF enhances ROUGE metrics between 1% and 3%.

2 Problem Formulation

Assume D is the whole document set and T is the set of all terms existing in D.

$D = \{ d_1, d_2, d_3, \dots , d_n \}$
$T = \{ t_1, t_2, t_3, \dots , t_m \}$

First of all, all stop words are removed. Later, the remaining terms are stemmed. Synonymous terms are evaluated together in frequency calculation. It is assumed that the terms in T are now independent.
In order to obtain meaningful words (T_r) to represent the documents, Term Dispersion Ratio (TDR) metric is proposed by Equation (1). This metric evaluates how common a word seen through documents. In this work, the TDR effect on meaningful word selection is experimented by different TDR values in Equation (2).

$$TDR_{t_j} = \frac{\sum_{i=1}^{n} \begin{cases} 1 & if\ t_j \in d_i \\ 0 & otherwise \end{cases}}{n} \qquad \forall\, t_j \in T, if\ TDR_{t_j} \geq TDR \tag{1}$$

$$\text{then } t_j \in T_r = \{ t_1, t_2, t_3, \dots , t_k \} \tag{2}$$

Some readers can be confused and mistakenly refer TDR as TF-IDF. However, TF-IDF formula is defined in Equation (3), where n is the document number in the data set and document frequency df_t is defined to be the number of documents in the collection that contain a term t.

$$idf_t = \log \frac{n}{df_t} \tag{3}$$

Thus the idf of a rare term is high, whereas the idf of a frequent term is likely to be low. However, TDR works in reverse manner. It is interested in

representative terms which frequently seen (over a ratio) in document collection or not (not representative term).

As a result, TDR is the minimum proportion to select a term to be representative. It is defined as the minimum number of documents must include a term. When a unit type (sentence, paragraph) is determined (a paragraph is used in EMDS), then all units in the D can be represented as follows:

U= { $u_1, u_2, u_3, \dots, u_v$ },

If the frequency of a representative term t_r in unit u_i is defined by relationship f(u_i, t_r), then unit term vector ($\overrightarrow{V_{u_i}}$) can be represented as follows:

$$\overrightarrow{V_{u_i}} = (\mathrm{f}(u_i, t_1), \mathrm{f}(u_i, t_2), \dots, \mathrm{f}(u_i, t_k))$$

For document $d_{k,,}$ the document center vector ($\overrightarrow{V_{d_k c}}$) is defined by the highest frequency of each representative term which is seen in this document (Equation (4)).

$$\overrightarrow{V_{d_k c}} = (\max\{\mathrm{f}(u_1, t_1), \dots, \mathrm{f}(u_l, t_1)\}, \dots, \max\{\mathrm{f}(u_1, t_k), \dots, \mathrm{f}(u_l, t_k)\}) \quad \text{where, } l \le v,\ u_i \in d_k \wedge t_j \in T_r \qquad (4)$$

Furthermore, $\overrightarrow{V_{Dc}}$ is defined by the highest frequency of each representative term which is seen in the all units of document set D (Equation (5)). It is called data set center vector.

$$\overrightarrow{V_{Dc}} = (\max\{\mathrm{f}(u_1, t_1), \dots, \mathrm{f}(u_v, t_1)\}, \dots, \max\{\mathrm{f}(u_1, t_k), \dots, \mathrm{f}(u_v, t_k)\}), \quad \text{where, } u_i \in U \wedge t_j \in T_r \qquad (5)$$

As soon as document center vectors and data set center vector are constructed, Euclidean distances are calculated between each document center vector ($\overrightarrow{V_{d_k c}}$) and data set center vector ($\overrightarrow{V_{Dc}}$). The units which are over 2σ distance are assumed outlier. The documents which are far away from the data set center vector ($\overrightarrow{V_{Dc}}$) are discarded (outlier documents). Others are now a candidate for summary.

The idea behind the outlier units can be explained by document structure. A well-written document generally consists of one topic and its sub-topics. A paragraph is expected to include one of these sub-topics. The process of selecting representative terms is actually an attempt to associate the term/s with a sub-topic. However, some paragraphs might include general information about the topic not a specific sub-topic (includes lots of infrequent representative terms and seems related to nearly all sub-topics). Although it seems a heuristic realization, it is a result of the entropy law given by Equation (6).

$$\text{Entropy} = -\sum_{i=1}^{n} p_i log_2 p_i \qquad (6)$$

Entropy implies the stability of the system. If entropy is zero, system is more stable. When entropy is getting closer to zero, then the unit is related to only a few terms. If a unit could be assigned to only one term, then it makes entropy zero, which would be the best result. Eventually filtering general units (related to the lots of representative terms) is suggested in this article. This type of units should be detected and they wouldn't be considered for future processing (extraction).

Moreover, units can contain lots of terms which may not be even representative terms, or only one representative term with low frequency. RTHF is a solution proposal to detect more informative units in the documents and can be defined as follows:

"A unit must contain at least one representative term which frequency in this unit is the most for all units in the document".

If a unit provides RTHF, it includes information about one sub-topic in detail. However, it doesn't still guarantee that unit is not a general one. We only decrease the probability of being a general unit. Moreover, RTHF helps us to decrease overlapping (data duplication) in the summary by selecting one or a few informative units for each topic in the document (depends on the compression rate).

The usage of RTHF is exemplified by paragraph vectors given in Table 1. The rows are the paragraphs (u_i) in the document and the columns are the representative terms (t_j).

Table 1. Example paragraph vectors.

	t_1	t_2	t_3	t_4	t_5	t_6	t_7	t_8
u_1	2	0	1	0	0	0	5	1
u_2	6	7	0	0	0	0	2	3
u_3	0	0	0	3	3	6	5	0
u_4	4	2	3	1	0	2	4	3
u_5	1	1	3	5	2	0	0	0
u_6	0	0	3	3	5	2	0	0

If RTHF is applied to the paragraph vectors in the Table 1 then u_1, u_2, u_3, u_5 and u_6 are selected for extraction. Although it is possibly a general paragraph (it includes nearly all T_r terms, but not at most for any one) in the document, not specific, it

would be even selected for a simple Matching Percent (MP) similarity measure, which is defined as percentage of count of seen representative terms over total count of representative terms. It can be shown as follows:

MP_{u_1} = 4 / 8 = 0.5,
MP_{u_2}= 4 / 8 = 0.5,
MP_{u_3} = 4 / 8 = 0.5,
MP_{u_4} = 7 / 8 = 0.875, **highest**,
MP_{u_5}= 5 / 8 = 0.675,
MP_{u_6}= 4 / 8 = 0.5.

On the other hand, when RTHF is applied, then $u_1$and u_3 have similar term frequencies for t_7, so that both of them are candidates for extraction. Moreover u_2 has two representative terms at most (t_1, t_2) and it is selected. However u_4 has no representative term at most frequency, so that it is not going to be selected.

3 Proof of RTHF

The effect of the TDR can be defined by the intersection property of the set theory. The condition for a term being a member of representative terms set (T_r) is the minimum number of documents it must be seen, named r. r can be defined as upper integer obtained from the multiplication of n by TDR given by Equation (7). Then the total number of combinational sets (C(n, r)) can be given by Equation (8).

$$r = \lceil n * TDR \rceil \quad (7)$$
$$C(n,r) = n! / r!\ (n-r)! = \text{k} \quad (8)$$

On the way i ϵ Z^+ and i ≤ k, define an index i on Equation 8, where i = $\{1, 2, 3, \dots, k\}$. Then members of combinational set can be expressed by $C_i^{(n,r)}$ notation. If it is openly stated, $C_i^{(n,r)}$describes the set that includes document numbers of ith combination in the set of r combinations of n documents.
Let's try to explain it with an example. Assume we have got following 4 documents and accept TDR is 0.5 for simplification. Then $C_i^{(n,r)}$ values are as follows:

D = $\{d_1, d_2, d_3, d_4\}$,
where, $r = \lceil 4 * 0.5 \rceil = 2$ and C(4,2) = 6

$C_1^{(4,2)}$ = { d_1, d_2}, $C_2^{(4,2)}$ = { d_1, d_3},
$C_3^{(4,2)}$ = { d_1, d_4}, $C_4^{(4,2)}$ = { d_2, d_3},
$C_5^{(4,2)}$ = { d_2, d_4}, $C_6^{(4,2)}$ = { d_3, d_4}.

Representative terms (r_i) in the combinational set $C_i^{(n,r)}$ are determined by the intersection of document center vectors of documents ($\overrightarrow{V_{d_k c}}$) these are in the combinational set ($C_i^{(n,r)}$). If all the documents in the combinational set have a nonzero frequency for a term in their document vector, then this term is selected as representative term. Otherwise it is not selected. In order to achieve this goal, first of all, the elements of $\overrightarrow{V_{d_k c}}$ document center vector are transformed into binary numbers by using the following sign function.

$$sgn(u) = \begin{cases} 1, & if\ u > 0 \\ 0, & if\ u = 0 \\ -1, & if\ u < 0 \end{cases}$$

Noting that because all of frequencies are non-negative, thus, the vector members are either one or zero. Sign function applied document center vector is represented by $\overrightarrow{sgn_{V_{d_k c}}}$.
By the way, the frequencies are removed, and the existence of a representative term in a document is only considered (1 means existence and 0 means non-existence). Then, combinational set $C_i^{(n,r)}$vectorial computation is done as defined in the Equation (9).

$$\overrightarrow{T_i^{(n,r)}} = \bigcap_{d_k \in C_i^{(n,r)}} \overrightarrow{sgn_{V_{d_k c}}} \quad (9)$$

Intersection operator in the Equation (9) results in a vector presents the terms which are member of all the documents within combinatorial set $C_i^{(n,r)}$. At that point, $\overrightarrow{T_i^{(n,r)}}$ vector is used to construct representative terms set ($T_{r_i}^{(n,r)}$) by applying the following rule.

$$T_{r_i}^{(n,r)} = \begin{cases} t_i \in T_{r_i}^{(n,r)}, & if\ t_i = 1 \text{ in } \overrightarrow{T_i^{(n,r)}} \\ t_i \notin T_{r_i}^{(n,r)}, & if\ t_i = 0 \text{ in } \overrightarrow{T_i^{(n,r)}} \end{cases}$$

Finally, all representative terms set for document collection is calculated by the Equation (10), as result of the union of $T_{r_i}^{(n,r)}$ sets obtained above.

$$T_r^{(n,r)} = \bigcup_{i \in \{1,\dots,C(n,r)\}} T_{r_i}^{(n,r)} \quad (10)$$

Let's continue with an example. Assume the following document center vectors are given in the Table 2. Assume TDR is 0.5.

Table 2. Example document center vectors.

	t_1	t_2	t_3	t_4	t_5	t_6
$\overrightarrow{V_{d_1c}}$	2	0	3	0	0	1
$\overrightarrow{V_{d_2c}}$	0	2	1	0	0	0
$\overrightarrow{V_{d_3c}}$	2	0	0	0	0	1
$\overrightarrow{V_{d_4c}}$	3	0	0	3	2	0

First of all, the document center vectors are converted into binary numbers vectors using sign function. It is given in the Table 3.

Table 3. Document center vectors are converted into binary numbers

	t_1	t_2	t_3	t_4	t_5	t_6
$\overrightarrow{sgn_{V_{d_1c}}}$	1	0	1	0	0	1
$\overrightarrow{sgn_{V_{d_2c}}}$	0	1	1	0	0	0
$\overrightarrow{sgn_{V_{d_3c}}}$	1	0	0	0	0	1
$\overrightarrow{sgn_{V_{d_4c}}}$	1	0	0	1	1	0

Then, $\overrightarrow{T_\iota^{(n,r)}}$ vector values are computed using Equation (9). It is exemplified for $C_1^{(4,2)}$ combination below.

$$\overrightarrow{T_1^{(n,r)}} = \overrightarrow{sgn_{V_{d_1c}}} \cap \overrightarrow{sgn_{V_{d_2c}}}$$

All results are given in the Table 4.

Table 4. $\overrightarrow{T_\iota^{(n,r)}}$ computations of combinational set $C_i^{(n,r)}$.

	t_1	t_2	t_3	t_4	t_5	t_6
$\overrightarrow{T_1^{(n,r)}}$	0	0	1	0	0	0
$\overrightarrow{T_2^{(n,r)}}$	1	0	0	0	0	1
$\overrightarrow{T_3^{(n,r)}}$	1	0	0	0	0	0
$\overrightarrow{T_4^{(n,r)}}$	0	0	0	0	0	0
$\overrightarrow{T_5^{(n,r)}}$	0	0	0	0	0	0
$\overrightarrow{T_6^{(n,r)}}$	1	0	0	0	0	0

Later, $T_{r_i}^{(n,r)}$ sets of representative terms are produced from the vectors given Table 4.

$T_{r_1}^{(n,r)} = \{t_3\}$,
$T_{r_2}^{(n,r)} = \{t_1, t_6\}$,
$T_{r_3}^{(n,r)} = \{t_1\}$,
$T_{r_4}^{(n,r)} = \{\emptyset\}$,
$T_{r_5}^{(n,r)} = \{\emptyset\}$,
$T_{r_6}^{(n,r)} = \{t_1\}$,

Finally $T_r^{(n,r)}$ set is constructed by union of $T_{r_i}^{(n,r)}$ sets above.

$T_r^{(n,r)} =$
$T_{r_1}^{(n,r)} \cup T_{r_2}^{(n,r)} \cup T_{r_3}^{(n,r)} \cup T_{r_4}^{(n,r)} \cup T_{r_5}^{(n,r)} \cup T_{r_6}^{(n,r)}$
$T_r^{(n,r)} =$
$\{t_3\} \cup \{t_1, t_6\} \cup \{t_1\} \cup \{\emptyset\} \cup \{\emptyset\} \cup \{t_1\} =$
$\{t_1, t_3, t_6\}$

Although other combinations between r and n must be considered, it provides the subset relationship between representative terms sets ($T_r^{(n,n)} \subseteq \ldots \subseteq T_r^{(n,r+1)} \subseteq T_r^{(n,r)}$), so that final representative terms set (T_r) can be defined by Equation (12) in a simple form instead of Equation (11).

$$T_r = \cup_{v \in \{r, r+1, \ldots, n\}} T_r^{(n,v)} \quad (11)$$

$$T_r = T_r^{(n,r)} \quad (12)$$

Let's consider the probability of a term (t_i) seen in the document set to be a member of T_r set. Assume terms (m terms) appearing within a document set with equal probability (1/m). If r is 1 then all terms are member of $T_r^{(n,1)}$ set. However, consider an r value, then the probability of selection k terms within m terms is given by Equation (13).

$$\mathrm{P}\,(T_r^{(n,r)} \mid t_i) = \frac{|T_{r_i}^{(n,r)}|}{m} = \frac{k}{m} \quad (13)$$

The probability obtained at Equation (13) is dependent to TDR by inverse ratio. When TDR increases, then the probability $\mathrm{P}(T_r^{(n,r)} \mid t_i)$ decreases. It is a result of the subset relationship between representative terms ($T_r^{(n,n)} \subseteq \ldots \subseteq T_r^{(n,r+1)} \subseteq T_r^{(n,r)}$). This can be realized heuristically and modeled by limit which is given by Equation (14).

$$\lim_{r \to n} \frac{|T_{r_i}^{(n,r)}|}{\mathrm{m}} = 0 \quad (14)$$

It can be obtained from Equation 14 that in the case of infinite document set then T_r would be an empty set. In other means, all units would not include a

common term all together. As a result, the count of selected representative terms can be controlled by arranging the r value. r is also determined by TDR that means it plays an important role for selecting representative terms in a controlled way.
On the other hand, the metrics called precision (P) and recall (R) which are defined by Equation (15) and Equation (16) respectively, must be increased for the success of EMDS. In order to increase precision and recall, then the summary must include most relevant paragraphs.

$$P = \frac{|\{\text{relevant paragraphs}\} \cap \{\text{retrieved paragraphs}\}|}{|\{\text{retrieved paragraphs}\}|} \quad (15)$$

$$R = \frac{|\{\text{relevant paragraphs}\} \cap \{\text{retrieved paragraphs}\}|}{|\{\text{relevant paragraphs}\}|} \quad (16)$$

In order to increase the relevant paragraph number in the summary, RTHF plays an important role. It determines general paragraphs these include many members of T_r with low frequencies.
If the model is simplified to understand heuristic, assume document set is composed of y paragraphs and T_r includes x terms. If each paragraph contains only one term of T_r at most frequency then the following general paragraphs counts would be determined.

$$\begin{pmatrix} If\ x \geq y & 0 \\ If\ x < y & \left\lceil \left(1 - \left(\frac{x}{y}\right)\right) * y \right\rceil \end{pmatrix}$$

The actual effect of RTHF is expected at higher TDR. The results obtained in [6] also support this idea (The best scores are marked for the 75% TDR).

4 Experiments

Experiments are applied to the same DUC2006 corpus. The system summaries are limited to 250 words and extraction is paragraph-based.
ROUGE [23] is used to evaluate RTHF model. We focused on the F_Score metric which is given by Equation (17). It is the harmonic mean of Equations (15) and (16).

$$\text{F-Score} = 2 * \frac{P * R}{P + R} \quad (17)$$

The model was run for three TDR values (25%, 50%, 75%) on each data set (50 data sets) of DUC 2006 corpus. The average of ROUGE metrics was calculated for these TDR values separately.

Abbreviations, Average_R, Average_P and Average_F, used on Table 5 are average recall, average precision and average F_score respectively. Moreover, the maximum value of each row is marked bold to enhance readability.
Table 5 compares the best ROUGE metrics announced in [6] and the RTHF applied similar EMDS for different TDR's.

Table 5. Comparison of EMDS [6] and RTHF for different *TDR*'s

	The values for [6]	RTHF system TDR (25%)	TDR (50%)	TDR (75%)
ROUGE-1				
Average-R	0.60830	0.59882	0.61308	**0.62069**
Average-P	0.57537	**0.57653**	0.57016	0.57277
Average-F	0.58993	0.58595	0.58935	**0.59472**
ROUGE-2				
Average-R	0.38602	0.37872	0.39090	**0.39865**
Average-P	0.36308	0.36390	0.36317	**0.36775**
Average-F	0.37175	0.37021	0.37558	**0.38191**
ROUGE-3				
Average-R	0.32035	0.31241	0.32333	**0.33144**
Average-P	**0.30734**	0.29972	0.30011	0.30560
Average-F	0.30816	0.30514	0.31050	**0.31744**
ROUGE-4				
Average-R	0.28037	0.27335	0.28298	**0.29069**
Average-P	0.26259	0.26211	0.26248	**0.26791**
Average-F	0.26961	0.26691	0.27165	**0.27835**
ROUGE-L				
Average-R	0.54041	0.52834	0.54380	**0.55423**
Average-P	0.50852	0.50840	0.50529	**0.51118**
Average-F	0.52031	0.51682	0.52251	**0.53090**
ROUGE-				
Average-R	0.14879	0.14563	0.14988	**0.15237**
Average-P	0.27307	0.27290	0.27108	**0.27376**
Average-F	0.19156	0.18950	0.19256	**0.19551**
ROUGE-				
Average-R	0.36260	0.35148	0.36784	**0.37583**
Average-P	0.32567	**0.32691**	0.31931	0.32134
Average-F	0.33931	0.33544	0.33858	**0.34415**

It is clear that RTHF model enhances all ROUGE metrics. Table 6 implies that RTHF model enhances results over 1.0% in general.

Table 6. Enhancing percentage of Average-F metric by RTHF

	Enhancing
ROUGE-1	
Average-F	**+0.8%**
ROUGE-2	
Average-F	**+2.7%**
ROUGE-3	
Average-F	**+3.0%**
ROUGE-4	
Average-F	**+3.2%**
ROUGE-L	
Average-F	**+2.0%**
ROUGE-W	
Average-F	**+2.0%**
ROUGE-SU4	
Average-F	**+1.4%**

5 Conclusion and Further Works

A Summarizer is a tool composed of phases and each phase uses a different technique. In other words, to develop a successful automatic summarizer it requires techniques, working all together in harmony.

This work is directed to mark general paragraphs and preventing them to be a candidate for summary. RTHF responsibility is to achieve extra filtering on units after representative term selected. RTHF forces a unit to contain at least a term of $T_r$which frequency in this unit is the most for all units in the same document.

It is applied to the existing EMDS. The results show that RTHF is a successful feature to select more informative paragraphs. Moreover, it produces the best value for higher TDR (75%) as theoretically explained. RTHF is a unit based approach so it could be applied successfully to other extractive unit types (sentence, segment).

On the other hand, this technique has a drawback. That is how to select enough representative terms to produce summary (depends on compression rate). In other words, the relationship between TDR and compression rate should be established.

It is obvious that RTHF prevents general paragraphs to be selected for the summary. On the other hand, the model still suffers from the MP which scores low value for paragraphs these are only included one member of T_r at most and a few members of T_r's.

By the way, RTHF is a sharp feature which means selecting the best one. Selecting paragraphs these have at least one member of T_r over term average in the document units would be better.

References:

[1] Kumar YJ, Salim N. Automatic multi document summarization approaches. J Computer Sci 2012; 8: 133-140.

[2] Sizov G. Extraction-based automatic summarization - theoretical and empirical investigation of summarization techniques. MSc, Norwegian University, Norwegian, Oslo, 2010.

[3] Nenkova A, McKeown K. A survey of text summarization techniques. In: Aggarwal CC, Zhai C-X, editors. Mining Text Data, USA: Springer US, 2012. pp. 43-76.

[4] Das D, Martins AFT. A survey on automatic text summarization. 2007; Language Technologies Institute, Technical Report.

[5] Mitra M, Singhal A, Buckley C. Automatic text summarization by paragraph extraction. In: Workshop on Intelligent Scalable Text Summarization; 11 July 1997, Madrid, Spain. pp. 39-46.

[6] Turan M, Sönmez C, Ganiz, MC. The benchmark of paragraph and sentence extraction summaries using outlier document filtering based multi-document summarizer. Inf Technol Control 2014; 43: 433-439.

[7] Edmundson HP. New methods in automatic extracting. J ACM 1969; 16: 264-285.

[8] Kupiec J, Pedersen JO, Chen F. A trainable document summarizer. In: Proceedings of the18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval; 1995; Scattle WA, USA: ACM. pp. 68-73.

[9] Kumar PA, Kumar KP, Rao TS, Reddy PK. An improved approach to extract document summaries based on popularity. Lect Notes Comput Sc 2005; 3433: 310-318.

[10] Suanmali L, Salim N, Binwahlan MS. Fuzzy logic based method for improving text summarization. Int J Comput Sci Inf Secur 2009; 2: 1-6.

[11] Gupta V, Lehal GS. A survey of text mining techniques and applications. J Emerg Techol Web Intell 2009; 1: 60-76.

[12] Binwahlan MS, Salim N, Suanmali L. Swarm based text summarization. J Comput Sci 2009; 5: 338-346.

[13] Bossard A, Rodrigues C. Combining a multi-document update summarization system with a genetic algorithm. In: Hatzilygeroudis I, Prentzas J, editors. Smart Innovation, Systems and Technologies. Berlin, Germany: Springer, 2011. pp.71-87.

[14] Manne S, Fatima SS. An extensive empirical study of feature terms selection for text summarization and categorization. In: CCSEIT-12; 26-28 Oct 2012; Coimbatore, India. pp. 606-613.

[15] Li X, Wu X, Hu X, Xie F, Jiang Z. Keyword extraction based on lexical chains and word co-occurance for Chinese news web page. 2008 IEEE International Conference on Data Mining Workshops; 15-19 Dec 2008; Pisa, Italy: IEEE. pp. 744-751.

[16] Balinsky H, Balinsky A, Simske S. Document sentences as a small world. International Conference on Systems, Mans and Cybernetics; 9-12 Oct 2011; Los Alamitos, CA, USA: IEEE. pp. 2583-2588.

[17] Wang M, Xi G, Wang X, Li C, Zhang Z. Multi-document summarization based on word feature mining. International Conference on Computer Science and Software Engineering; 12-14 Dec 2008; Wuhan, China: IEEE. pp. 743-746.

[18] Litvak M, Last M. Graph-based keyword extraction for single-document summarization. MMIES '08 Proceedings of the Workshop on Multi-Source Multilingual Information Extraction and Summarization; 23 August 2008; Manchester, UK: ACM. pp. 17-24.

[19] Marcu D. Discourse trees are good indicators of importance in text. Advances in Automatic Text Summarization, MIT Press, 2009. pp. 123-136.

[20] Yong-dong X, Xiao-long W, Tao L, Zhi-ming X. Multi-document summarization based on rhetorical structure: sentence extraction and evaluation. IEEE International Conference on Systems, Man and Cybernetics; 7-10 Oct 2007; Montreal, Canada: IEEE. pp. 3034-3039.

[21] Salton G, Singhal A, Mitra M, Buckley C. Automatic text structuring and summarization. Inform Process Manag 1997; 32: 53-65.

[22] Okazaki N, Matsuo Y, Matsumura N, Ishizuka M. Sentence extraction by spreading activation through sentence similarity. IEICE Trans Inf Syst 2003; E86D: 1686-1694.

[23] Lin C-Y. ROUGE:A package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out(WAS); 25-26 July 2004; Barcelona, Spain: Association for Computational Linguistics. pp. 74-81.

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

Metin Turan, implemented algorithm, proved theory and carried out the experiments.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding

Conflict of Interest

The author has no conflict of interest to declare.