4th International Conference on Web Research
April 25-26, 2018; Tehran, IRAN
ICWR2018
کنفرانس بین المللی وب پژوهی
چهارمین
۵ و ۶ اردیبهشت ۱۳۹۷؛ تهران، ایران

Query-oriented Text Summarization using Sentence Extraction Technique

Mahsa Afsharizadeh
Faculty of Engineering
The University of Kashan
Kashan, I.R. Iran
mafsharizade4@gmail.com

Hossein Ebrahimpour-Komleh
Faculty of Engineering
The University of Kashan
Kashan, I.R. Iran
ebrahimpour@kashanu.ac.ir

Ayoub Bagheri
Faculty of Engineering
The University of Kashan
Kashan, I.R. Iran
a.bagheri@kashanu.ac.ir

Abstract— Today there is a huge amount of information from a lot of various resources such as World Wide Web, news articles, e-books and emails. On the one hand, human beings face a shortage of time, and on the other hand, due to the social and occupational needs, they need to obtain the most important information from various resources. Automatic text summarization enables us to access the most important content in the shortest possible time. In this paper a query-oriented text summarization technique is proposed by extracting the most informative sentences. To this end, a number of features are extracted from the sentences, each of which evaluates the importance of the sentences from an aspect. In this paper 11 of the best features are extracted from each of the sentences. This paper has shown that use of more suitable features leads to improved summaries generated. In order to evaluate the automatic generated summaries, the ROUGE criterion has been used.

Keywords: query-oriented summarization, natural language processing, text mining, extractive summarization.

I. INTRODUCTION

The huge growth of information in various sources, including the World Wide Web, news articles, e-books and emails, has left mankind with a huge amount of information. The busy lifestyle of humans in the modern world has also minimized the time available for discovering information from this massive volume. This has led to the emergence of a kind of contradiction in the modern society of today's world. On the one hand, the lack of time and, on the other hand, the need to be aware of various information due to job and social needs, leads to requiring methods to facilitate access to information in the shortest possible time.

Because of this issue, the fields such as text mining, natural language processing and artificial intelligence also have come together and seek to find a solution to this problem. The result of the researchers' efforts in these fields has led to the emergence of an interesting and important topic called automatic text summarization. Automatic summarization of text documents in the shortest possible time has made it possible for people to access the most important content. Manual text summarization requires a large number of specialized people in different fields and spent a great deal of time and effort in this direction. It's not possible to produce summaries for texts without people specialized in different fields. These people, having enough knowledge and experience in the context of those texts, are able to do text summarization using the power of thought and reasoning. This suggests that automating this operation by the machine is a very useful process and requires the use of various information from areas such as artificial intelligence, natural language processing, and text mining.

Automatic text summarization faces some challenges. For example, extracting the proper features for sentences and words in the text has a significant impact on the performance of the summarization system, leading to select the most appropriate sentences that contain the most important information about the main subject of the text. Text summarization methods are divided into two categories: extractive and abstractive. Extractive summarization extracts important sentences from source documents and group them together to generate summary. Abstractive summarization creates a brief useful summary by generating new sentences. In this paper we propose an extractive technique for query-oriented summarization.

The extractive summarization should identify and extract the important sentences in a document. So far, various methods have been used in the extractive summarization method. One of these methods is TF-IDF [1]. This is a numerical criterion that indicates the importance of a word in a document among a corpus of documents. TF shows the frequency of a word in a document. IDF is a measure that reduces the weight of frequently occurring words in the corpus and increases the weight of words that rarely occur. The words with high TF-IDF value, have a strong relationship with the document in which they are located. Many people have used TF-IDF to measure the importance of sentences in extractive summarization, for example [2], [3], [4], and [5].

Another method used in extractive summarization is fuzzy logic in which, scoring sentences can be done using fuzzy logic. At first, appropriate features are extracted from the text. Then, with regard to the extracted features for each sentence, a score is calculated using the fuzzy system. There are also some works done in this area, including [6], [7], [8], [9], and [10].

Extractive summarization is also done using graph-based methods. In this method, each sentence from the document is considered as a vertex in the graph. If there is a common semantic relationship between the two sentences, they will be connected and weighted through the edges. A graph-based ranking algorithm is used to decide the importance of a vertex in a graph. The most important vertices in this graph are considered as important sentences and included in the summary. Among those who have worked in this field can be mentioned [11], [12], [13], and [14].

Another method used in extractive summarization is LSA[i]. The LSA method is an algebraic statistical method that derives the meaning and similarity of sentences based on words' information [15]. The idea of using the LSA method for summarizing text documents was introduced in 2001 [16]. The LSA method is an automated technique for extracting relationships Between words in text documents. In this method, a Term-by-Sentence matrix is first created from the original text. Then SVD is applied to it. This action leads to the discovery of hidden dimensions that are related to the various topics discussed in the document. Finally, the resulting matrices are used to identify and extract the most important sentences. These were some of the work done in the field of extractive summarization.

The rest of the paper is as follows: part II describes the proposed extractive summarization method. Part III expresses the experimental results. Finally, part IV concludes the paper.

II. The Proposed Extractive Summarization

In this paper a query-oriented extractive summarization method is proposed. The extractive summarization method extracts the sentences containing useful information and displays them in summary. The main challenge in these methods is how to identify and select important sentences in a document. For this purpose, a score is given to each sentence based on the extracted features from them. Then the sentences are ranked according to these score values. The proposed Scheme is shown in Figure 1. In this paper we extend *Ahuja* and *Anand* work [17] by adding some other useful features [18]. Our proposed scheme is consisted of five steps, each of them is explained below.

A. Data Preparation

Data preparation is the first step in the proposed method. We apply the proposed summarization system to DUC 2007 corpus [19]. This corpus has 45 clusters. Each of them has 25 text documents. There are various text formats in these documents. There are plenty of redundant characters in them. In this step, the sentences are extracted from raw text with different formats.

B. Text Pre-processing

Text pre-processing is an important step in all text processing tasks. It contains some important parts such as tokenization, stop words removal, stemming and Part Of Speech Tagging[ii].

1) Tokenization

There are a lot of tokens that are not worthy of content such as question mark, surprise mark, comma and so on. For this purpose, in the tokenization process, these meaningless tokens are deleted from the text and the sentences are broken down to meaningful tokens. These meaningful tokens are separated from each other by the white space or punctuations.

2) Stop Words Removal

Stop words are the words that are frequently repeated in the text but they are not meaningful like "a", "an", "the" and so on. clearing the text of these words is the task of stop words removal step.

3) Stemming

Stemming step reduces the words into their stems. For example, the words "is", "am" and "are" are transformed to their stem, that is, "be".

4) POS tagging

This step is used to specify the word category. Therefore, the words are categorized into the groups like *nouns*, *verbs*, *adjectives*, *adverbs* etc.

C. Feature extraction

After text pre-processing step, it's time to extract the appropriate features from the text.

In this paper, the extractive summarization technique is used. The most informative sentences are identified and extracted from the text. The main challenge in this technique is to decide which sentences are the most important. For this purpose, some useful features should be extracted from each sentence and a score is assigned it based on its feature values. Then the high ranked sentences are selected to be present in the summary.

There are a number of features in the *Ahuja* and *Anand* work for the purpose of extractive summarization [17]. Choosing the appropriate features has a great impact on the performance of the summarization. We use from these features for summarizing DUC 2007 corpus [19]. In order to enhance the performance of the summarization, we use a number of query dependent features that leads to higher ROUGE values [18]. The experimental results show this improvement. In this paper, we use from eleven appropriate features. The features are listed below:

1- *Document Feature*
2- *Sentence Position*
3- *Normalized Sentence Length*
4- *Numerical Data*
5- *Proper Noun*

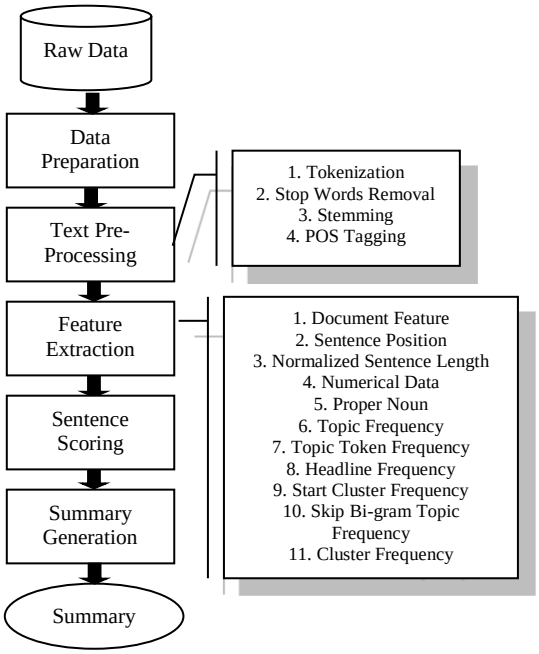

Figure 1. The proposed summarization method

6- *Topic Frequency*
7- *Topic Token Frequency*
8- *Headline Frequency*
9- *Start Cluster Frequency*
10- *Skip Bi-gram Topic Frequency*
11- *Cluster Frequency*

Document Feature: This feature calculates a weight for a sentence based on the total content of its document. It is sum of the weights of separate words in the sentence.

Sentence Position: The main idea behind this feature is that the sentences that appear in the beginning and end of the text, are usually more important than the sentences that appear across the text. This feature is calculates based on (1):

$$Sent\_pos(x) = \frac{N - x + 1}{N} \qquad (1)$$

In this equation, N is the total number of sentences in the document and x is the index of the interested sentence.

Normalized Sentence Length: This feature says that long sentences are more informative than short ones. It assigns a weight to a sentence based on the ratio of its length to the length of the longest sentence in its document.

Numerical Data: The sentence with the numerical data is probably an informative sentence. It is computed by dividing the count of numerical data in the sentence to the length of sentence.

Proper Noun: Sentences with proper nouns are likely to carry valuable information. Therefore, they are a good choice for presence in the summary. This feature is computed based on the ratio of the count of total proper nouns in the sentence to the length of the sentence.

Topic Frequency: This feature calculates the relative frequency of the content words (set of the words after stop words removal procedure) in the sentence in its topic description.

Topic Token Frequency: It is similar to *Topic Frequency* feature. The only difference between them is that *Topic Token Frequency* is computed on all the tokens instead of only the content words.

Headline Frequency: This feature calculates the relative frequency of the content words in the sentence on the set of headlines in the cluster.

Start Cluster Frequency: It calculates sum of the relative frequencies of the content words in the sentence on the set of the first 100 words in its document and all of the documents in its cluster.

Skip Bi-gram Topic Frequency: This feature is like to *Topic Frequency* feature. But it is computed on the skip bigrams rather than unigrams. A bigram is a sequence of two adjacent words in the text. A skip bigram is a bigram that allow gaps between its pair of words.

Cluster frequency: It computes sum of the relative frequencies of the content words in the sentence in the total content words in its cluster.

D. Sentence Scoring

All of the feature values of a sentence are calculated in the previous step (11 features). In the *Sentence Scoring* step, a score is assigns to each sentence based on a linear function of its feature values. Total score for a sentence s is calculated by (2):

$$Score(s) = \sum_{i=1...m} w_i . f_i(s) \qquad (2)$$

The first five weights are the same as the original paper [17]. The rest of the weights are experimentally set as: 0.9, 0.6, 0.5, 0.4, 0.4 and 0.2, respectively.

E. Summary Generation

After the *Sentence Scoring* stage, all of the sentences are ranked based on their scores. Top ranked sentences are selected to generate the summary. Considering that the summary length should be 250 words maximum.

III. Experimental Results

In this paper, DUC 2007 corpus is used for summarization purpose. DUC 2007 corpus is consisted of 45 clusters each one has a set of 25 relevant text documents. Each cluster is about a topic. The purpose is to generate a fluent 250-word summary for each cluster. Each topic and its document cluster have given to 4 different NIST assessors. These assessors have created a 250-word summary of the document cluster. These

multiple reference summaries are used in the evaluation of summary content. The comparison results for ROUGE measure (ROUGE-1 to ROUGE-4) between *Ahuja* [17] and the proposed method are shown in TABLE I. The comparison results in in this table show that the proposed method has improved the average recall, average precision and average F-measure.

The comparison results for ROUGE-L, ROUGE-W, ROUGE-S and ROUGE-SU are shown in Figure 2 to Figure 5, respectively. The proposed method in all of the figures show higher results in compare to *Ahuja* method [17]. This is because of enriching the feature set by appending some query oriented features to the primary feature set. So the extracted features will be both informative and query relevant. The experimental results in this paper have shown the improvement of the proposed method in comparison with *Ahuja* method [17]. The comparison of ROUGE-2 average recall and ROUGE-SU4 average recall results for our proposed model with *Ahuja* model and some peer systems that had participated in DUC evaluations are shown in Table II. For better visualizing this comparison, the results of Table II. are shown in a chart. Figure 6 shows this chart.

IV. CONCLUSION

In this paper a query-oriented text Summarization technique using sentence extraction is proposed. In the extractive summarization technique, the most informative sentences in the text are identified and selected to attend in the summary. To identify sentences containing valuable information, a set of appropriate features are extracted from the text. Whatever the extracted features of the sentences are more appropriate, the most informative sentences are more accurately identified and the quality of the generated summary improves.

TABLE I. ROUGE COMPARISON RESULTS BETWEEN PROPOSED METHOD AND AHUJA

		Avg-Recall	Avg-Precision	Avg-F
ROUGE-1	***Ahuja***	0.36480	0.35181	0.35808
	Proposed	0.36790	0.36116	0.36439
ROUGE-2	***Ahuja***	0.06887	0.06667	0.06774
	Proposed	0.07579	0.07467	0.07521
ROUGE-3	***Ahuja***	0.02012	0.01957	0.01984
	Proposed	0.02469	0.02439	0.02453
ROUGE-4	***Ahuja***	0.00939	0.00915	0.00926
	Proposed	0.01229	0.01212	0.01220

TABLE II. EVALUATION RESULTS (ROUGE-2 AND ROUGE-SU) BETWEEN PROPOSED METHOD AND SOME OTHER METHODS

Result / Method	Average ROUGE-2 Recall	Average ROUGE-SU Recall
Peer 16	0.03813	0.07385
Peer 1	0.06039	0.10507
Peer 27	0.06238	0.11884
Ahuja	0.06887	0.12922
Peer 6	0.07135	0.12517
Proposed	0.07579	0.13023

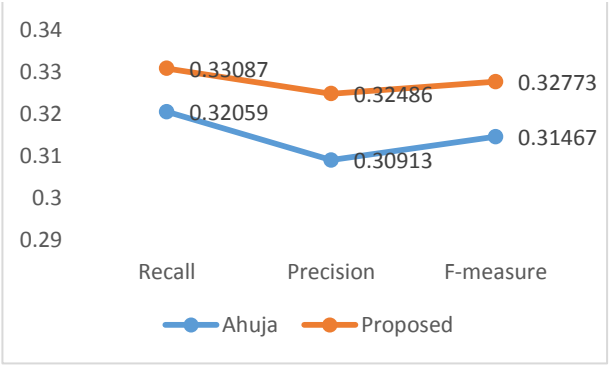

Figure 2. ROUGE-L results for Ahuja and proposed method

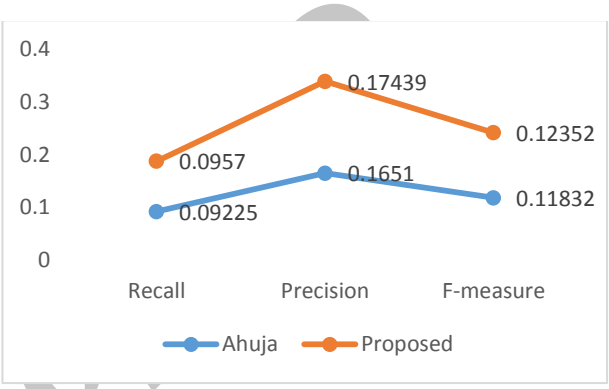

Figure 3. ROUGE-W results for Ahuja and proposed method

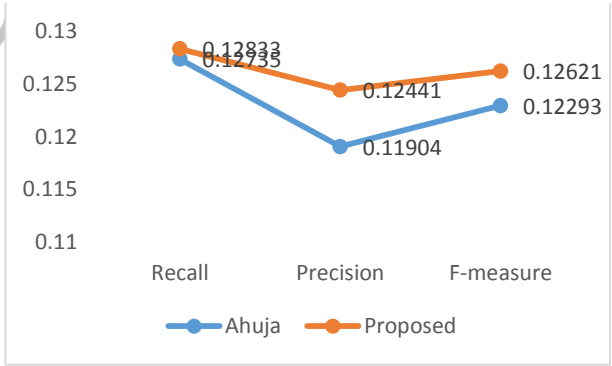

Figure 4. ROUGE-S results for Ahuja and proposed method

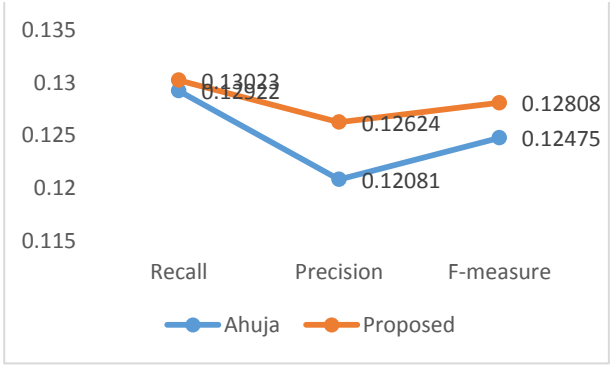

Figure 5. ROUGE-SU results for Ahuja and proposed method

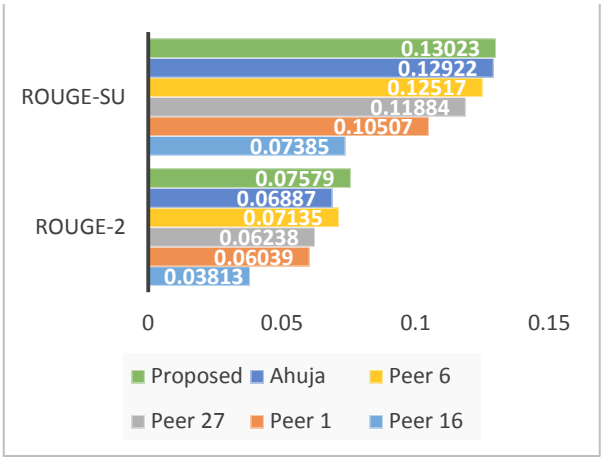

Figure 6. ROUGE Results between Proposed Method and Some other Methods

Each of these features takes the importance of sentences from a different perspective. In this paper, the proposed method by *Ahuja* has been improved by appending some appropriate query based features. The extracted features by *Ahuja* can be used for generic summarization but they are not sufficient for query based summarization. They are similarity between sentence to its document, the position of the sentence in the document, the length of the sentence, the existence of numerical data and the presence of proper nouns in the sentence. In order to identify both the informative and query relevant sentences, the feature set is enriched with a number of other appropriate features. The first set of features can identify informative sentences and the second set of appropriate features will help to extract the query relevant sentences. These features are related to the topic and headlines. Furthermore, skip bigrams are also considered in addition to the unigrams. Finally, by using more complete set of convenient features, better results in summarization is achieved. The results of the experiments show this improvement.

REFERENCES

[1] Ramos, J. *Using tf-idf to determine word relevance in document queries*. in *Proceedings of the first instructional conference on machine learning*. 2003.

[2] García-Hernández, R.A. and Y. Ledeneva. *Word sequence models for single text summarization*. in *Advances in Computer-Human Interactions, 2009. ACHI'09. Second International Conferences on*. 2009. IEEE.

[3] Sarkar, K. *An approach to summarizing Bengali news documents*. in *proceedings of the International Conference on Advances in Computing, Communications and Informatics*. 2012. ACM.

[4] Baralis, E., et al., *Mwi-sum: A multilingual summarizer based on frequent weighted itemsets.* ACM Transactions on Information Systems (TOIS), 2015. **34**(1): p. 5.

[5] Jayashree, R., S. Murthy, and B.S. Anami. *Categorized Text Document Summarization in the Kannada Language by Sentence Ranking*. in *Intelligent Systems Design and Applications (ISDA), 2012 12th International Conference on*. 2012. IEEE.

[6] Babar, S. and P.D. Patil, *Improving performance of text summarization.* Procedia Computer Science, 2015. **46**: p. 354-363.

[7] Ghalehtaki, R.A., H. Khotanlou, and M. Esmaeilpour. *A combinational method of fuzzy, particle swarm optimization and cellular learning automata for text summarization*. in *Intelligent Systems (ICIS), 2014 Iranian Conference on*. 2014. IEEE.

[8] Hannah, M.E., T. Geetha, and S. Mukherjee. *Automatic extractive text summarization based on fuzzy logic: a sentence oriented approach*. in *International Conference on Swarm, Evolutionary, and Memetic Computing*. 2011. Springer.

[9] Modaresi, P. and S. Conrad. *From Phrases to Keyphrases: An Unsupervised Fuzzy Set Approach to Summarize News Articles*. in *Proceedings of the 12th International Conference on Advances in Mobile Computing and Multimedia*. 2014. ACM.

[10] Suanmali, L., M.S. Binwahlan, and N. Salim. *Sentence features fusion for text summarization using fuzzy logic*. in *Hybrid Intelligent Systems, 2009. HIS'09. Ninth International Conference on*. 2009. IEEE.

[11] Mihalcea, R. *Graph-based ranking algorithms for sentence extraction, applied to text summarization*. in *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*. 2004. Association for Computational Linguistics.

[12] Malliaros, F.D. and K. Skianis. *Graph-based term weighting for text categorization*. in *Advances in Social Networks Analysis and Mining (ASONAM), 2015 IEEE/ACM International Conference on*. 2015. IEEE.

[13] Litvak, M. and M. Last. *Graph-based keyword extraction for single-document summarization*. in *Proceedings of the workshop on Multi-source Multilingual Information Extraction and Summarization*. 2008. Association for Computational Linguistics.

[14] Cheng, K., Y. Li, and X. Wang. *Single Document Summarization Based on Triangle Analysis of Dependency Graphs*. in *Network-Based Information Systems (NBiS), 2013 16th International Conference on*. 2013. IEEE.

[15] Landauer, T.K., P.W. Foltz, and D. Laham, *An introduction to latent semantic analysis.* Discourse processes, 1998. **25**(2-3): p. 259-284.

[16] Gong, Y. and X. Liu. *Generic text summarization using relevance measure and latent semantic analysis*. in *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval*. 2001. ACM.

[17] Ahuja, R. and W. Anand, *Multi-document Text Summarization Using Sentence Extraction*, in *Artificial Intelligence and Evolutionary Computations in Engineering Systems*. 2017, Springer. p. 235-242.

[18] Toutanova, K., et al. *The pythy summarization system: Microsoft research at duc 2007*. in *Proc. of DUC*. 2007.

[19] *DUC 2007*. 2007.

[i] Latent Semantic Analysis
[ii] POS tagging