A New Effective Criterion to Select Sentences in Extractive Text Summarization

Maryam Kiabod, Mohammad Naderi Dehkordi, Sayed Mehran Sharafi
Department of Computer Engineering, Najafabad Branch, Islamic azad University, Isfahan, Iran
m_kiabod@sco.iaun.ac.ir

Abstract- **This paper introduces a new criterion to get better performance in selecting the most important sentences of text for extractive text summarization. There are two kinds of criteria to find the most relevant sentences of text: statistical criteria and semantic relations between text sentences. The proposed technique is a statistical criterion. The idea behind our approach is to consider the position of sentence words relative to words in the sentence occurring in title and keywords. We evaluate this criterion in combination with other statistical criteria. The results show that using this criterion in selecting the most important sentences of text has good results.**

Index Terms- **Extractive Text Summarization, keywords, Sentence Extraction, Text mining**

I. Introduction

The amount of information grows rapidly on the web. As a result, we need text summarization systems to save time and access the main concept of the text in a short time.

Text summarization is the process of reducing the length of the original text. Text summarization techniques are classified into two categories: extractive and abstractive. Extractive category selects important sentences of text and concatenates them to form the summary, while abstractive category derives the main concept of the original text.

Natural language processing is a tool that is used in abstractive text summarization approach. This technique applies semantic relations between words to determine the main concept of text. Extractive approach forms the summary of text based on characteristics of sentences. Some of these criteria are: number of words in the sentence occurring in title [1], sentence length [2], sentence position [3], and number of numerical data [4]. After calculating these criteria, they are combined to compute scores of the sentences.

Most of these criteria are statistical. In spite of simplicity of these criteria, they eventuate in good results. Also, they do not need an external database to determine the scores of text sentences. In this article, we introduce a new statistical criterion. We evaluate this criterion in choosing the most important sentences of text. The results show that this criterion is useful to select the most relevant sentences of text. The rest of this paper is as follow. Section 2 provides a review of previous works on text summarization systems. Section 3 presents our technique. Section 4 describes experimental results and evaluation. Finally, we conclude and suggest future work in section 5.

II. Literature Review

Automatic text summarization dates back to fifties. In 1958, Luhn [5] created text summarization system based on weighting sentences of a text. He used word frequency to specify topic of the text. There are some methods that consider statistical criterions. Edmundson [6] used Cue method (i.e. "introduction", "conclusion", and "result"), title method and location method for determining the weight of sentences. Statistical methods suffer from not considering the cohesion of text.

Kupiec, Pederson, and Chen [7] suggested a trainable method to summarize the original text. In this method, number of votes collected by the sentence determines the probability of being included the sentence in summary.

Another method includes graph approach proposed by Kruengkrai and Jaruskululchi [8] to determine text title and produce summary. Their approach takes advantages of both the local and global properties of sentences. They used clusters of significant words within each sentence to calculate the local property of sentence and relations of all sentences in text to determine global property of text.

Beside statistical methods, there are other approaches that consider semantic relations among words. These methods need linguistic knowledge. Chen, Wang, and Guan [9] proposed an automated text summarization system based on lexical chain. Lexical chain is a series of interrelated words in a text. WordNet is a lexical database includes relations among words such as synonym, hyponymy, meronymy, and some other relations.

Svore, Vander Wende and Bures [10] used machine learning algorithm to summarize text. Eslami, Khosravyan D., Kyoomarsi, and Khosravi proposed an approach based on Fuzzy Logic [11]. Fuzzy Logic does not guarantee the cohesion of the summary of text. Halavati, Qazvinian, Sharif H. applied Genetic algorithm in text summarization system [12]. Genetic Algorithm also is used in improving content selection in automatic text summarization by Khosraiyan, Kumarci, and Khosravi [13]. It was based on statistical tools. Latent Semantic Analysis [14] is another approach used in text summarization system. Abdel Fattha and Ren [15] proposed a technique based on Regression to estimate text features weights. In regression model a mathematical function can relate output to input variables. Feature parameters were considered as input variables and training phase identifies corresponding outputs.

There are some methods that combine algorithms, such as, Fuzzy Logic and PSO [2]. Salim, Salem Binwahla, and Suanmali [16] proposed a technique based on fuzzy logic. Text features (such as similarity to title, sentence length, and similarity to keywords, etc.) were given to fuzzy system as input parameters.

III. Methodology

We use extractive text summarization to form the summary. Extractive method is one of the methods in text summarization technique. This method extracts the most relevant sentences of text based on their scores. These scores are calculated by considering the sentence characteristics and combining them to compute the sentence score. After that, sentences with the highest scores are selected and concatenated to form the summary. These characteristics include sentence position [3], sentence length [2], and sentence-to-centroid cohesion [5]. These characteristics are statistical criterions. The benefit of using these criteria is that they are independent of an external database to determine the scores.

This step involves four sections: preprocessing, the criterions for calculating scores of sentences, calculating sentence score, and sentence selection. We explain these sections in the next four sections. The proposed criterion will be introduced in the second section.

[1] Preprocessing

The first step in our technique involves preparing text document to be analyzed by text summarization algorithm.

In this stage, we perform sentence segmentation, sentence tokenization, part of speech tagging, removing stop words, and word stemming. Sentence segmentation separates text document into sentences. Then, sentence tokenization is applied to separate input text into individual words. We use part of speech tagging to recognize types of text words. Stop words are the words with less importance in identifying the important content of text. As a result, we remove stop words of text. Finally, word stemming removes prefixes and suffixes of each word.

[2] The Criterions for Calculating Scores of Sentences

In this section, we explain the criterions that we used in calculating sentence score. We use five characteristics which others used in selecting the most important sentences of text. Then we introduce our proposed criteria. These characteristics are as follow:

1. Term frequency

Term frequency is a criteria used to determine significant words of text [5]. The results of using this criterion has shown that words with higher term frequency are more important and having greater chances to be included in the most relevant sentences of text.

To calculate this criterion, we use part of speech tagging to remain nouns of text and remove every other type of the words from text. Then, we calculate term frequency of each word and normal it by dividing this score by total number of text words.

2. Position score

Baxendale [17] showed that if the position of sentence within the paragraph is some fixed position, the sentence is suitable to be selected for summary. Experimental results corroborated the fact that in most of the paragraph, the topic sentence was the first sentence and in some paragraphs, the topic was the last sentence of paragraph. As a result, the first and last sentence of paragraph would be appropriate choices as summary of text. It means that sentences located at the first and last paragraph of text are more important than others. So, we divide text into three sections and fix the position score of sentences at the first and the last paragraph to 0.66 and at the second section to 0.33.

3. Sentence length

Sentence length is a criterion which is used in identifying the best sentences for summary. This criterion is calculated by dividing number of words occurring in the sentence by number of words occurring in longest sentence of text [2].

4. Numerical data

Numerical data is another effective criterion in selecting sentences of text for summary. We give value 1 to sentences that include a numerical data.

5. Proper name

Proper name is another criterion which is used in determining sentences for summary. The value 1 is given to the sentences that contain proper name [18].

6. The proposed criterion

In this stage, we introduce a statistical criterion and use it in combination with other criterions discussed earlier to calculate the score of sentence.

This criterion is based on the distance between a word in the sentence and words in that sentence occurring in title and keywords. The distance is the subtraction of the position of word in the sentence and the position of words in the sentence occurring in title and keywords. We show that sentence words with higher distance to words of the sentence occurring in title and keywords have less importance and have fewer chances to be included in important sentences of text. We conclude that the words with shorter distance are more important and having greater chances to be included in significant words of text. This criterion is as follow.

$$the - new - criterion = \frac{1}{(shortest - dis) + 1} \qquad (1)$$

where shortest_dis is the shortest distance between a word of the sentence and words of the sentence occurring in title and keywords.

We give value 1 to words of the sentence included in title and keywords. It means that these words do not have any distance to words occurring in title and keywords of text (the shortest_dis is zero for these words of text). As a result, these words get the complete score. To compute this criterion, first, we calculate this score for each word of the sentence. Then, for a sentence, we get the sum of scores of all words of the sentence. Finally, this score is normalized by the number of words included in the sentence

[3] Calculating Sentence Score

In this step, we calculate two scores for each sentence of text.

The first one is the combination of five criterions. These criterions include term frequency, sentence position, sentence length, numerical data, and proper name. This score has been shown in equation (2).

$$first - sentence - score = \frac{tf+pos+len+NumData+PropName}{5} \quad (2)$$

We add our proposed criterion to the equation (2) to calculate the second score for each sentence of text. The second score has been shown in equation (3).

$$second - sentence - score = \frac{tf+pos+len+NumData+PropName+ProposedCriteria}{6} \quad (3)$$

where tf is the sum of normalized term frequency of sentence words, pos is the score of sentence position within the text, len is score of sentence length, NumData displays if the sentence contains a numerical data (1 is given to sentences with numerical data), ProperName shows the existence of proper name in the sentence, and Proposed Criteria is the score of our proposed criterion.

[4] Sentence Selection

After calculating score of text sentences, it is time to select the best sentences for summary. For this reason, first, we rank text sentences according to their scores in decreasing order. Then, the first n-top sentences are selected as the most important sentences of text. Finally, we concatenate them together to form the summary. The value n depends on the compression rate. The higher compression rate leads to a shorter summary. We fix the compression rate to 0.8.

IV. TESTING & ANALYSIS

In this step, we evaluate the performance of considering the new criteria in selecting sentences for summary. For this reason, we calculate the first score (the equation (2)) and the second score (the equation (3)) for each sentence of text. Then, we compare these two scores together to evaluate the effectiveness of proposed criterion.

We use DUC2002[1] as the test dataset to evaluate our criteria. We perform three criteria to evaluate the effectiveness of the proposed criterion: Precision, Recall, and F-measure. Precision is the fraction of retrieved instances that are relevant, while Recall is the fraction of relevant instances that are retrieved. F-measure is a combination of Recall and Precision. These criterions have been shown in equation (4), equation (5), and equation (6) [19].

$$precision - rate = \frac{|\{relevant\ sentences\} \cap \{retrieved\ sentences|}{|\{retrieved\ sentences\}|} \quad (4)$$

$$recall\ rate = \frac{|\{relevant\ sentences\} \cap \{retrieved\ sentences\}|}{|\{relevant\ sentences\}|} \quad (5)$$

$$F - measure = \frac{2*precision\ rate*recall\ rate}{precision\ rate+recall\ rate} \quad (6)$$

The results have been shown in Figure 1, Figure 2, and Figure 3. The numerical results have been shown in Table 1. The results show that this criterion is effective in selecting the most relevant sentences of text.

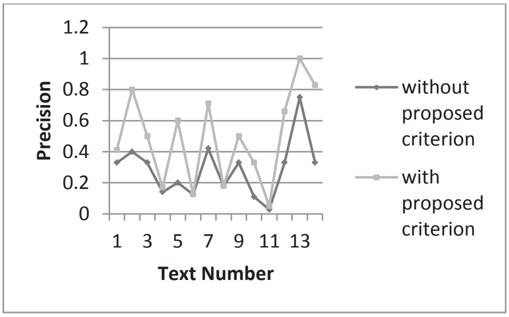

Figure 1: The precision score with and without the proposed criterion

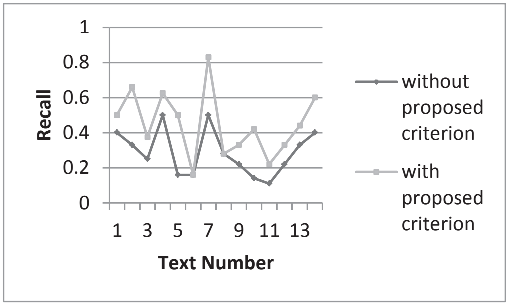

Figure 2: The recall score with and without the proposed criterion

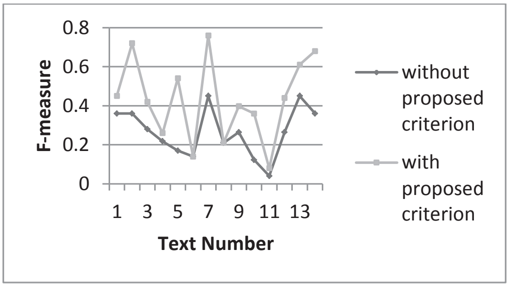

Figure 3: The F-measure score with and without the proposed criterion

[1] www.nlpir.nist.gov

Table 1
The numerical results of precision, recall and F-measure

Set no.	Without the proposed criterion			With proposed criterion		
	P	R	F	P	R	F
D061j	0.33	0.4	0.36	0.41	0.5	0.45
D062J	0.4	0.33	0.36	0.8	0.66	0.72
D071F	0.33	0.25	0.28	0.5	0.375	0.42
D072F	0.14	0.5	0.218	0.17	0.625	0.26
D074B	0.2	0.16	0.17	0.6	0.5	0.54
D075B	0.125	0.16	0.14	0.125	0.16	0.14
D083A	0.42	0.5	0.45	0.71	0.83	0.76
D091C	0.18	0.28	0.21	0.18	0.28	0.21
D092C	0.33	0.22	0.264	0.5	0.33	0.397
D098e	0.14	0.11	0.123	0.33	0.42	0.36
D0102e	0.027	0.11	0.04	0.05	0.22	0.08
D106g	0.33	0.22	0.264	0.66	0.33	0.44
D110h	0.75	0.33	0.45	0.83	0.44	0.61
D113h	0.33	0.4	0.36	0.83	0.6	0.9
Average	0.288	0.283	0.327	0.478	0.447	0.449

V. CONCLUSION

In this article, we proposed a new statistical criterion which is effective in choosing the sentences of text for summary. The originality of the criterion lies on the shortest distance between the word and words of the sentence occurring in title and keywords. We showed that words with less distance with words occurring in title and keywords are more important and are included in the most important sentences of text. The results show that considering this criterion in selecting the most relevant sentences of text is useful and increases the precision of choice.

In future, we intend to evaluate the effectiveness of the mathematical functions on this criterion. Also we evaluate this criterion in combination of some other statistical criterions, such as, cue phrases.

REFERENCES

[1] G.Salton, C.Buckley, "Term-weighting approaches in automatic text retrieval", Information Proceeding and Management 24, 1988, 513-523.Reprinted in: Sparck-Jones, K.; Willet, P. (eds).Readings in I.Retreival, Morgan Kaufmann, 1997, pp.323-328.
.

[2] L.Suanmali, , M. Salem Binwahlan, and N. Salim , “sentence Features Fusion for Text Summarization using Fuzzy Logic”, IEEE,pp.142-145, 2009.

[3] H.P.Edmundson, “New methods in automatic extraction”, journal of the ACM, 1969, pp.264-285.

[4] C.Y.Lin, "Training a selection function for extraction" in Proc. 8th int. conf. Information and knowledge management, Kansas City, Missouri, United States,1999, pp.55-62.

[5] H.P.Luhn, ”The Automatic Creation of literature abstracts”, IBM journal of Research Development, 1958, pp.159-165.

[6] H.P.Edmundson, “New methods in automatic extraction”, journal of the ACM, 1969, pp.264-285.

[7] J.Kupiec, j.Pedersen, AND F.Chen, “A trainable document summarizer”, in Proc. 18th ACMSIGIR Conf., 1955, pp.68-73.

[8] C.Jaruskululchi, C.Kruengkrai, “generic text summarization using local and global properties of sentences”, 2003 IEEE/WIC int. conf. web intelligence, pp.13-16.

[9] Y.Chen, X.Wang, L.V.YI.Guan, “Automatic text Summarization Based on Lexical chains”, in Advances in Natural Computation, 2005, pp.947-951.

[10] K.Svore, L. Vanderwende, and C.Bures, “Enhancing single-document summarization by combining Ranknet and third-party sources”, In Proc. EMNLP-CoNLL.

[11] F., Kyoomarsi, H., Khosravi, E. Eslami, and P.Khosravyan Dehkordy, “Optimizing Text Summarization Based on Fuzzy Logic”, in Proc. 7th IEEE/ACIS Int. Conf. Computer and Information Science, IEEE, University of shahid Bahonar Kerman,2008,pp.347-352.

[12] V.Qazvinian, L.Sharif Hassanabadi, R.Halavati, “Summarization Text with a Genetic Algorithm-Based Sentence Extraction”, International Journal of Knowledge Management Studies (IJKMS), 2008, vol.4, no.2, pp.426-444.

[13] P.Khosraviyan Dehkordi, F.Kumarci, H. Khosravi, “Text Summarization Based on Genetic Programming”, International Journal of Computing and ICT Research, 2009.

[14] S.Hariharan, “Multi Document Summarization by Combinational Approach”, International Journal of Computational Cognition, 2010, vol.8, no.4, pp.68-74.

[15] M.Abdel Fattah, and F.Ren, “Automatic Text Summarization”, Proceedings of World of Science, Engineering and Technology, 2008, vol.27, pp.195-192.

[16] L.Suanmali, N. Salim, and M.Salem Binwahlan, “fuzzy swarm based text summarization”, journal of computer science, 2009, pp.338-346.

[17] Baxendale, P, Machine-made Index for Technical Literature –An Experiment‘, IBM Journal of Research Development, Vol. 2, No.4, 1958, pp. 354-361.

[18] M.Hassel, “exploitation of named entities in automatic text summarization for Swedish”, in proc. NODALIDA 03-14th Nordic conf. computational linguistics, may 2003.

[19] Y.Y.Chen, O.M.Foong, S.P.Uong, I.Kurniawan, “Text Summarization for Oil and Gas Drilling Topic”, Proceeding of world academy of science and technology, vol.32, 2008, pp.37-40