AHP TECHNIQUES FOR PERSIAN TEXT SUMMARIZATION

Seyyed Mohsen Tofighy[1], Ram Gopal Raj[2], Hamid Haj Seyyed Javadi[3]
[1, 2] Faculty of Computer Science and Information Technology University of Malaya, Kuala Lumpur, Malaysia
[3]Department of Math and Computer Science, Shahed University, Tehran, Iran
Email: Tofighy_sm@yahoo.com, ramdr@um.edu.my, h.javadi@shahed.ac.ir

ABSTRACT

In recent years, there has been an increasing amount of information on the web. Some of essential resources to shorten text documents use summarization technologies. In this paper, we present an AHP technique for Persian Text Summarization. This proposed model uses analytical hierarchy as a base factor for an evaluation algorithm and improves the summarization quality of Persian language text. The weighting and combination methods are two main contributions of the proposed text evaluation algorithm.

Keywords: Text Evaluation, Summary, Traditional Features

1.0 INTRODUCTION

The proliferation of available information on the internet has led to it becoming an integral part of human life. However, many users do not have enough time to read so much information especially text, which results in users often resorting to reading abstracts and headlines instead. It is not easy for users to manually summarize these large documents. Nonetheless, it is important for each user to generate summaries as a means of saving time. The goal of automatic text summarization is to condense the source text into a shorter version of itself while still preserving its informational content and overall meaning, [1]. Determining the informational content of text is a complex process as often involves processes such as stemming, [2], information dissemination, [3], as well as other models that are used to compare information, [4]. Currently there are two primary approaches used in text summarization: extractive and abstractive, [5]. Extractive summarization methods are focused on identifying the most relevant sentences within a given source document. This results in the final summary consisting of fixed components, sentences of the original source document. As opposed to this, abstractive type summarization initially attempts to understand the text in order to eventually summarize the individual sentences. This often results in the summary consisting of sentences that are somewhat different from the original source text as most of the sentences would have been summarized. While this is the ideal goal of abstractive text summarization, an ideal summarizer remains unavailable as it involves natural language processing (NLP), especially in terms of the semantic representations involved. The problem is acerbated in the Persian language as many of the required language resources that are required for the NLP are not available currently.

The use of AHP for evaluation and decision-making was studied by many authors. Despite the differences in the approaches taken, each technique has its merits. AHP is based on pair wise comparisons using ratio scales to indicate the summarization accuracy performance. In this paper we present the use of AHP for evaluation and selecting sentences based on their weights, and concurrently adjust the method to make it easier to use. It should be noted that the Persian language differs from the English language both morphologically and semantically. Our method consists of two main parts, one to calculate the weights for the sentences and another to evaluate those sentences which should result in an optimal summary of the Persian text to be generated.

2.0 RELATED WORK

Initial work done on text summarization was started almost fifty years ago, [6].Text summarization in its infancy consisted of reading the original document and attempting to understand the contents, and after that generating a short document of the content. The automatically summarized text was generated by a machine that assessed the importance of the information within the input document, based on a user's or application's needs, [7]. The earliest research on automatic summarization consisted of selecting sentences from a source document

based on the term frequencies to measure sentence relevance, [8], sentence positions in a paragraph, [9], and sentence similarity, [10]. Sentences are included in the summary if the words in the sentence have sufficiently high scores. Most supervised extractive methods used currently focus on utilization of powerful machine learning algorithms that can properly combine these features.

Other approaches consist of statistical analysis, generally based assessing the structure of the text via discourse analysis combined with training algorithms that use human generated summaries to estimate the importance probabilities of sentences from the source document. The importance probabilities would then be used to determine if a sentence should be included in the summary.

The use of Bayesian models in text summarization systems is popular due to its simplicity, [11]. Other work, [12], claimed that the corpus-trained featuring weights increase accuracy, an assertion that was supported by [13]. This model handles each sentence individually, which results in main connection between the sentences being ignored. Genetic algorithms can be used to calculate the weights of each sentence in the summary as shown by [14].

Despite its benefits, statistical methods have their shortcomings when used for text summarization such as need for human supervision when dealing with ambiguous words; misunderstood rhetoric, construing non-text objects and synonyms and other context dependent terms. Nonetheless statistical approaches to text summarization are still considered useful, [15]. Recent research on text summarization has overcome some of the problems of statistical approaches by combining them with other approaches. For example, [16] presented an automatic text summarization system combining both a statistical approach and fractal theory to summarize documents.

3.0 ADAPTATION OF TRADITIONAL FEATURES

As mentioned earlier, extractive type summarizations are used to identify important sentences within the original text and put them together to create a summary. For the process of selecting important sentences, effective features denoting the relevance of sentences are determined. In this section, there are several common features that have been considered for sentence selection weighting of sentences but we use six features for scoring sentences as follows:

F1= Word frequency.
Frequency usage of each word is calculated after removal of stock words such as 'a', 'the', 'an' and so on. These words are often the most frequent words in sentences but have little semantic impact on a sentences meaning. The most common measure used to calculate the word frequency is *tf* and *idf* method such as (1).

$$w_i = tf_i * log\frac{N}{n} \quad \textbf{(1)}$$

Where tf_i is the frequency of term t_i at sentence, N is the number of sentences and n is number of sentences that contain the term t_i. The Sentence Thematic score can be calculated as (2), where k is number of words in sentence.

$$\text{F1}=\frac{\sum_{\text{i=1}}^{\text{k}} \text{W}_\text{i}}{\text{Max}\,(\sum_{i=1}^{k} w_i)} \quad \textbf{(2)}$$

F2= Keywords in the sentence.
Keywords are usually words that have the highest occurrences within a sentence. It can be calculated as the ratio of the number of thematic words that occur in the sentence to maximum number of key words in the sentence, as (3).

$$\text{F2} = \frac{No.of\ Keyword\ in\ Sentence}{Max(No.Keyword)} \quad \textbf{(3)}$$

F3= Headline Word.
The third feature is the headline word that is based on the hypotheses that the title contains the subject of the document summarized in its words and the sentence is highly relevant to the document. The headline feature score of sentence s is calculated as (4).

$$F3 = \sum_{s \in path\ from\ main\ title\ to\ Sentence} \frac{No.of\ Title\ word\ in\ s\ of\ sentence}{No.of\ Word\ in\ Title} \quad \textbf{(4)}$$

F4= Cue Word.
These are pragmatic words in a sentence which indicate that a given sentence is carrying an important message in the document. The cue score for each sentence is calculated as in (7).

$$\mathrm{F4} = \frac{No.Cue\ word\ in\ s\ of\ sentence}{Max(No.Cue\ Word\ in\ paragraph\ of\ sentence\ )} \quad \textbf{(7)}$$

F5= Sentence Location
This feature based on the assumption that first and end sentences of a paragraph are the most important. Position score is calculated as (6), where *n* is the number of sentences of paragraph which sentence is located it and *i* is ordinal number of sentence that regarding its position among other sentences.

$$F5 = \mathrm{Max}(\frac{1}{i}, \frac{1}{n-i+1}) \quad \textbf{(6)}$$

F6= Sentence Length.
This feature is useful to penalize sentences that are too short, such as these sentences are not expected to belong to the summary, which is calculated as in (7).

$$F6 = \frac{No.Word\ occurring\ in\ sentence\ x}{No.Word\ occurring\ in\ longest\ sentence} \quad \textbf{(7)}$$

4.0 ANALYTIC HIERARCHY PROCESS

One of the most useful methods for decision-making is the Analytical Hierarchy Process (AHP). This method was developed by Saaty in 1980, [17]. AHP is a powerful method the uses a multiple criteria decision-making approach which can generally assist in selection issues to help decide which alternatives are most suitable for defined problems. Most of the decision issues are multi-criteria: appropriate 'Length', maximize 'Term Weight', maximize 'keyword', appropriate 'position', maximize 'Headline' and 'Cue' and etcetera. There are some situations where making an incorrect decision may be too costly and its impact may be uncertain, for example not selecting a key sentence or word resulting in a loss of critical information. Evaluating a decision requires that it is considered based on its weights and its scores. A sentence's importance must be assessed in terms of the values mentioned above and how important these values themselves are for a given decision, [18].

This makes AHP very suitable in assisting in making the appropriate decisions and making powerful and accurate text summarization systems. AHP is a structured technique that is very useful in the process of complex decision-making for problems which involve singling out and offering one out of many possible decisions. It should be noted that in most instances, AHP does not focus on one exclusively "correct" decision, but rather chooses a decision which proves to be the most adequate or the most useful based on the user's needs. Just as a human decision-maker bases a judgment on knowledge and experience, in order to make decisions correctly,

thus does the AHP approach base the decisions made on the factors that are specified. The benefit of this approach is that it categorizes tangible and intangible factors in a structured manner, and provides an organized yet relatively uncomplicated solution to the decision-making problems, [19].

The simplest hierarchy in decision-making consists of three levels: The top level, as usual, is the goal to be achieved by the decision; on the second level is constituted by the six groups of factors as defined by the text summarization techniques; the third level is constituted by the strategies that should be evaluated and compared, [20]. A graphical representation of the hierarchy is presented in the following Fig. 1:

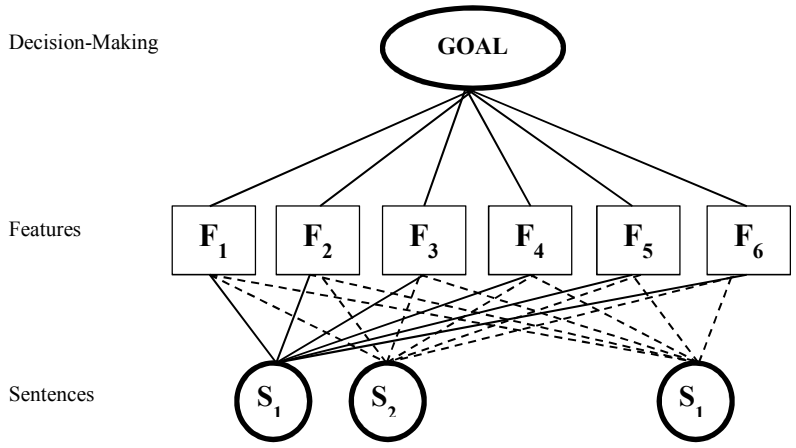

Fig. 1. AHP structure for text summarization

4.1 The use of AHP for summary generation

Table 1 illustrates our proposed algorithm based on AHP method. As seen in Table 1 … F1, F2… F6 are features of the summarization found in the source sentences. As mentioned in earlier our summarization method comprises six factors: Length, Term Weight, keyword, position, Headline and Cue; F1, F2 F3, F4, F5 and F6 respectively.

The initial impression is that the features are not independent but they are in fact thus. To begin we create an initial matrix for the summarization where the principal diagonal contains entries of '1's, as each factor is as important as itself. In this way to make the pairwise comparison, as shown in Table 1 below.

Table 1. Pairwise comparison matrix of features

	F1	F2	F3	F4	F5	F6	W
F1	1	α					W_{F_1}
F2	1/α	1					W_{F_2}
F3			1				W_{F_3}
F4				1			W_{F_4}
F5					1		W_{F_5}
F6						1	W_{F_6}

We then consider the number of potential sentences, S1...n. Subsequently we extract six features for pairwise comparisons but this time in terms of how well S1...n perform in based on the six criteria, F1...6, as seen in Table 2.

Table 2. Pairwise comparison matrix of sentences

Fkϵ(1...6)	S1	S2	.	Sn	W
S1	1	α			W1

S2	1/α	1			W2
.			1		.
Sn				1	Wn

The relative importance or weights (w) for features (Fk) and sentences (Sn), where; n (1, ..., N) is the number of sentences and k (1, ... , 6) is the number of features, are obtained from pair-wise comparisons. This is performed in order to the degree of preference for each summarization option. The relative importance of each sentence is obtained by comparing these sentences with all other features. From the weights calculated, a matrix as seen in Table 3 is generated for each individual k (1, ... , 6) known as AHP matrix.

Table 3. Matrix of sentences weights for each feature

	F1	F2	F3	F4	F5	F6
S1	W11	W12	W13	W14	W15	W16
S2	W21	W22	W23	W24	W25	W26
.	.	.	.	.	.	.
Sn	Wn1	Wn2	Wn3	Wn4	Wn5	Wn6

After specifying the weights for each feature and pair within the comparison matrix for each sentence based on the specified features, the final weight of each sentence is calculated using the following formula(8):

$$W_{S_i} = \sum_{j=1}^{6} W_{F_j} * W_{ij} \quad (8)$$

After calculating each sentence weight, determining a sentence's final priority is possible which is then used to select the sentences with highest priority.

We define the compression ratio, R, as the number of sentences of summary divided by the total number of sentences within the original document, [16] that can be determinate by formula (9).

$$R = \frac{\#\text{ sentence of output}}{\#\text{ sentence of input}} \quad (9)$$

5.0 EXPERIMENTAL RESULTS

The evaluation our method is obtained by comparing the experimental summarization results with human summaries that were produced manually, [21]. Recall is taken as a measure of the informational components of the original text that are correctly extracted and Precision is taken as a measure of the components of extracted information that are correct, [22]. Hence the Recall and Precision are formulated as follows:

$$\text{Recall} = \frac{\text{Sentence Count}(\text{Source Document} \cap \text{Summary})}{\text{Sentence Count}(\text{Source Document})} \quad (10)$$

$$\text{Precision} = \frac{\text{Sentence Count}(\text{Source Document} \cap \text{Summary})}{\text{Sentence Count}(\text{Summary})} \quad (11)$$

$$\text{F} - \text{Measure} = \frac{2 * \text{precision} * \text{Recall}}{\text{precision} + \text{Recall}} \quad (12)$$

We evaluated the performance of our method using several different sets of Persian news texts. This process was performed using two different proportions of the original text: 30% and 40%. We calculate the precision and recall of our work and other approaches. The results are summarized in Table 4. Our method is still at an experimental stage and we are working toward increasing its precision.

Table 4 shows the results of our system as compared with other two systems. In this comparison, a higher number indicates better performance.

Table 4. system evaluation – F-Measure

Compression Rate	Our System	FarsiSum	Fractal
30% - 40%	0.71	0.53	0.69

6.0 CONCLUSION

The demand for automatic text summarization systems has been increasing. This is due to the proliferation of available information mostly via the internet resulting in a situation where users and information seekers have been overwhelmed. This has led to users seeking out shorter or summarized versions of the available full length documents. However, current automatic text summarization methods and the resulting system have had limited accuracy leading to users not being able to rely on the generated summarizations fully. Our solution to the accuracy problem as presented in this paper is to utilize a hybrid method of text summarization. Our method utilizes AHP, which has thus far been used mostly in group decision-making, in order to evaluate the summarizations performed on each sentence within a given set of text. This allows our method to perform the summary in a complex and multifaceted manner based on several key features, based on AHP generated weights of sentences. The benefit of our Hybrid text summary model is that it provides a mechanism to consider multiple affective sentence features in the summarization process. The evaluation of our method as compared to other available methods for summarizing Persian text proved that our method has better performance than those methods. This summarization model is a new approach to text summarization and as part of our future work we plan to incorporate, other decision-making techniques such as the Technique for Order of Preference by Similarity to Ideal Solution (TOPSIS).

REFERENCES

[1] Kumar J. Y., Salim N.,"Automatic Multi Document Summarization Approaches", *Journal of Computer Science* Vol. 8 No.1,2012, pp. 133-140.

[2] R.G. Raj, S. Abdul-Kareem. "A Pattern Based Approach for the Derivation of Base Forms of Verbs from Participles and Tenses for Flexible NLP". *Malaysian Journal of Computer Science,* 2011, 24(2): 63 – 72.

[3] R.G. Raj, S. Abdul-Kareem. "Information Dissemination and Storage for Tele-Text Based Conversational Systems' Learning". *Malaysian Journal of Computer Science,* 2009, 22(2): 138 – 159.

[4] R.G. Raj, V. Balakrishnan. "A Model for Determining The Degree of Contradictions in Information". Malaysian Journal of Computer Science, 2011, 24(3): 160 – 167.

[5] Radev, D.R., E. Hovy, K. McKeown,"Introduction to the Special Issue on Summarization". *J. Comput. Linguistic,* Vol. 28, 2002, pp. 399- 408.

[6] Huang C. C., Chu P. Y., Chiang Y. H., "A Fuzzy AHP application in government-sponsored R&D project selection", *Omega*, Vol. 36, 2008, pp. 1038-1052.

[7] Móro R., "Combinations of Different Raters for Text Summarization", *Information Sciences and Technologies Bulletin of the ACM*, Slovakia, Vol. 4, No. 2,2012, pp. 56-58 575. Mani, I., G. Klein, D. House, L. Hirschman, T. Firmin, and B., "a text summarization evaluation", *Natural Language Engineering,* Viol. 8, 2002, pp. 43–68.

[8] Luhn, H. P., "The Automatic Creation of Literature Abstracts", *IBM Journal of Research and Development,* Vol. 2, 1958, pp. 159-165.

[9] Baxendale, P. B., "Machine-Made Index for Technical Literature: An Experiment:, *IBM Journal of Research and Development,* Vol. 2, 1958, pp. 354-361.

[10] Gong Y., Liu X., "Generic text summarization using relevance measure and latent semantic analysis", *annual international ACM SIGIR*, Vol. 24, 2001, pp. 19-25.

[11] Kim, S.B., Han, K.S., Rim, H.C. and Myaeng, S.H., "Some Effective Techniques for Naïve Bayes Text Classification", IEEE Transactions on Knowledge and Data Engineering, Vol. 18, No. 11, 2006, pp. 1457-1466.

[12] Kupiec J., Pedersen J., and Chen F., "A trainable document summarizer. In Proceedings", *Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, Vol. 18, 1995, pp. 68– 73.

[13] Edmundson, H. P., "New Methods in Automatic Extracting", *Journal of the ACM,* Vol. 16, 1969, pp. 264-285.

[14] Yeh J. Y., Ke H. R., Yang W. P. & Meng I. H,, "Text summarization using a trainable summarizer and latent semantic analysis*", Information Processing and Management*, Vol. 41, 2005, pp. 75–95.

[15] Victoria M., "Statistical Approaches to Automatic Text Summarization", *Bulletin of the American Society for Information Science and Technology*, Vol. 30, No. 4, 2004.

[16] Tofighy M., Kashefi O., Zamanifar A., Javadi H. S., "Persian Text Summarizatioon Using Fractal Theory", *Communications in Computer and Information Science*, Vol. 252, 2011, pp. 651-662.

[17] Saaty T.L., "The Analytical Hierarchy Process, Planning, Priority", *Resource Allocation. RWS Publications*, 1980.

[18] Ristova E., Panov Z.,Ivanova T. S., "using the AHP methodology to evaluate strategic investment alternatives of new paradigms in information technology", *International Journal of Engineering Science and Technology (IJEST)*,Vol. 4, No. 2, 2012, pp. 710-715.

[19] Palcic I., Lalic B., "analytical hierarchy process as a tool for selecting and evaluating project", *Int j simul model* Vol. 8, 2009, pp. 16-29.

[20] Rahimpour C. B., Khodabandeh A. A., " AHP Techniques for Trust Evaluation in Semantic Web", *Journal of Advances in Computer Research*, Vol. 3, 2011, pp. 85-91.

[21] Sparck Jones K., " Automatic Summarizing: Factors and Directions", *In: Mani, I., Maybury, M. T. (eds.) Advances in Automatic Text Summarization*, Cambridge 1999, pp. 1-12. MIT Press.

[22] Amini M., Gallinari P., " The use of unlabeled data to improve supervised learning for text summarization", *ACM SIGIR,* Vol. 24, 2002, pp. 105–112.

BIOGRAPHY

Seyyed Mohsen Tofighy is a student in University of Malaya (UM) attached to the Department of Artificial Intelligent. He has just completed her Masters studies in the field of Software Engineering.

Ram Gopal Raj holds a PhD from the University of Malaya (UM) and is attached with the Department of Artificial Intelligence of the Faculty of Computer Science and Information Technology at UM. His research is mostly related to chatterbot technology and information search methods as well as knowledge representation schemes.

Hamid Haj Seyyed Javadi holds a PhD from Amirkabir University of Tehran (AUT). He is currently a professor assistant at the Department of Computer Science and Information Technology at Parand University. Him work deals with information security, information retrieval and data and knowledge engineering.