International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

Application of Machine Learning Techniques in Persian Text Summarization Systems

Sayede Azadeh Hosseinzadeh, Ruhollah Dianat, Mohammad Bahrani

Abstract **– With the rapid growth of the Internet and the increasing information and online resources to find the desired information from large volumes of information to users is difficult. The method proposed in this paper, a system based on neural network is summarized text maker. In the proposed system, the original sentences based on features such as similarity with title and similarities with the center of the text, sentence length, sentence position, positive and negative words, valuable keywords scores receive. These points can be extracted from the sentences in a training body. According to the training corpus of sentences, in summaries to be or not to be labeled manually. Neural Network Based on these corpus, trained and finally used in the neural network by receiving scores of sentences extracted from a test text, decides whether it should be included in the final summaries. Experimental results on a documents of database Hamshahri show that the proposed system is able to with the scale F about 0.67 summaries do.**

Keywords **– Sentence Extraction, Text Summarization, Similarity of Sentences, Neural Network, Machine Learning Algorithm.**

I. INTRODUCTION

Nowadays, with the rapid growth of data and information, finding appropriate information is very important and efficient. Find related documents with the information we require is difficult. How to access this information and research, they have always been one of the major problems researchers. For solving this problem users need tools that can be used to identify the appropriate information. Such as these tools is a text summarization techniques. Summarize text processing that takes a document as input and returns as output a shorter document that contains important parts of the text. Thus, the user can achieve to the information required at least time. The purpose of the good summary is providing a summary of the original text to display important part of it as a good summary of the readability and cohesion between sentences within it.In general summarizing text can be done in two ways:

• *Extractive summarization*

In these summaries, important sentences the same manner were expressed in the original text, identify and are copied literally in the text. In summary extraction, information retrieval techniques are used. Such as these techniques, we can frequency table of words and identify keywords that for select the important parts of the text are noted.

• *Abstractive Summarization*

This kind of summarization, summary sentences is taken from the original text.

This means that exactly sentences do not exist in the original text. This kind of summary that act much like the human mind contains statements that are not present in the original document but covers the basic concepts.

The purpose of this type summarize, understanding document using knowledge-based and producing of high quality summary.

The main objective of this study is to presentation of approach based on neural network.

This means that the preprocessing is done on the texts, based on the similarity of parameters such as, similarity with central sentence and evaluations are carried out by neural networks and finally the measure of precision and measure of recall and F will compare the evaluations.

Rest of the paper is organized as follows:

In Section 2, we review the work done by research groups described in text summarization.

In section 3, we implement the proposed method using neural networks is provided using the database Hamshahri, the proposed method will be evaluated.

Finally, in Section 4, the results of the research are suggested.

II. RELATED WORKS

The use of text summarization tools goes back to 1950. The first summarization system was introduced in 1958 by Ian [4]. He proposed a simple method based on word frequency. He maintained that criteria of frequent the words could be useful criteria for the diagnosis of sentences valuable and important.

Another Summarization System was proposed by Baksndl in 1962 [1]. In this system, a new parameter was used to score sentences in a document. This new parameter, the location parameter of sentences in the text.

Edmoundson In 1969 [2], used a new method based on assigning a public weight to each sentence of the text for the select important sentences. He is selected four the parameter position sentences in the text, the similarity of each sentences with title of text, frequency of word and words symbol for the value of the sentences.

In 1995 Kupiec and colleagues [3] were used of combined parameters such location sentence in paragraph, sentence length and word symbol to produce Extractive summarization techniques.

In recent years the use of evolutionary algorithms to generate summaries are taken into consideration. In research conducted in 2007 by Ghazvinian [7] involved three similar parameters such as, text readability and coherence and cohesion sentences.

Shareqi in 2008, [8] conducted by using the same parameters in harmony search algorithm summarization the documents.

In 2009[6], presented multi-document summarization the Persian text using a method based on clustering. And

proposed a new method based on clustering for multi-document summarization of Persian texts.

In 2009 [9] a paper was presented as a new approach for text summarization based on harmony search algorithm. The proposed method is a two-stage text summarization system is based on harmony search algorithm.

In 2011[5], presented a paper titled an approach based on fuzzy inference system for Persian text summarization. This paper presents a practical approach to selecting sentences value from a the original text.

III. The Proposed Method

A. The proposed method

In this paper, we have proposed a text summaries based on neural network mechanism that aims to improve the quality previous approaches and effort to produce a summary which is closer to man.

In other words, summary produced by proposed system, the summarize that there is good coherence between sentences and has good readability and covers important parts of the original text. Among the challenges we are faced in extractive summarization, this is what of the parameters is used to produce a summary that are most similar to human summaries. In addition, the influence of each parameter how much to select summary sentences. One approach to this challenge is the use of machine learning techniques.

There is a two-stage machine learning techniques: Training phase and test phase.

Training phase is when a computer model is trained by a set of sample data warehouse and test phase when the model is tested by other instances of the data warehouse. When learning a complete model, it can be used for other applications and evaluations. Using this technique, summarization system based on defined parameters teach for sentences that how sentences are appropriate in summarization. In fact, using a sample data set to learn the patterns inference and these patterns are used to select other sentences.

To evaluate the sentences need to have the proper tools and features. Correct selection of properties, resulting in a more qualitative to evaluation sentences and identifying sentences important and no significant. Also, the choice of tools to achieve this task is very important. Select the appropriate characteristics, it is important But more importantly, combining these features together is for valuation a sentence. Features are used in this stage, the features are independent of the values of these properties are only associated with the original sentences. At this stage the 7 features we used such as similarity with title and similarities with the center of the text, sentence length, sentence position, positive and negative words, valuable keywords. To evaluate the original document sentences, first text summarization system, received the original document, and does first some basic tasks, such as remove unnecessary tags and separation of the original document sentences and the document is ready to be processed then created for each sentence a vector with 7 properties. Then we use neural network to combine specifications. This way the score vectors for each sentence of the original text is given into neural network. The neural network, for each sentence score between 0 and 1 is assigned that indicates the degree of sentence importance. Finally, in the last component, that is sentence extraction component, the sentence that their score is lower than a specified threshold are identified. These sentences are sentences that contain important parts are not original document and do not probably in final summary. Therefore, they were removed from the original document and become original document to smaller document that are removed less important sentences. In other words, this document includes sentences that are valuable and important to cover the original content. The concepts discussed in the summarization system

Sentences weight systems: Summarization systems break primary documents to a number the sentence. Each of these, are allocated based on the number of occurrences of words weighted to. Tf-idf weighting system to calculate the weights of these sentences are used:

$$tf_{i,j} = \frac{freq_{i,j}}{Max_l\, freq_{i,j}} \tag{1}$$

$$idf_i = \log \frac{N}{n_i} \tag{2}$$

$tf_{i,j}$ is frequency word I in sentence j that based on the number of occurrences of word against maximum occurrences of the word in the text is calculated also documents which there is a word repeatedly, it may be less important. Thus, the "document inverse frequency" or "agent idf» is used together with the weight of words. Inverse word frequency used to calculate the important of words that were less important than usual words. In the formula, N the total number of sentences in the documents and ni is the number of sentences containing the word i.

For each word in the sentence Wi, j based on tf and idf are defined as follows:

$$w_{i,j} = tf_{i,j} \times idf_i \tag{3}$$

Similarity matrix: In this matrix,are calculated the similarity of each sentence with other sentences in the original document . The following formula is used to compute the similarity of sentences:

$$Sim(sj, si) = \frac{\vec{Sj}.\vec{Si}}{|\vec{Sj}| \times |\vec{Si}|} \tag{4}$$

$$Sim(s_j, s_i) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,j}}{\sqrt{\sum_{i=1}^{t} w^2{}_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2{}_{i,j}}} \tag{5}$$

We discusses by using cosine similarity to calculate the similarity between sentences in the original document. And the resulting values we can save in the similarity matrix similar to Figure 1, so can use it later.

International Journal of Electronics Communication and Computer Engineering
Volume 5, Issue 3, ISSN (Online): 2249–071X, ISSN (Print): 2278–4209

$$\begin{bmatrix} sim_{1,1} & sim_{1,2} & \dots & sim_{1,N} \\ sim_{2,1} & sim_{2,2} & \dots & sim_{2,N} \\ \dots & & & \\ \dots & & & \\ \dots & & & \\ sim_{N,1} & sim_{N,2} & \dots & sim_{N,N} \end{bmatrix}$$

Fig.1. matrix similar to the original document sentences

B. Feature extraction

1. Similarity with title: A good summary contain sentences that is similar to the original text title. For calculating the feature based on the vector model used of a document title as a query against all the original text sentences and calculating the similarity between the title vector and each sentence $sim\ (s_j, q)$ to give a score.

$$Sim(s_j, q) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,q}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2_{i,q}}} \quad (6)$$

The value of this attribute is normalized as follows:

$$Title\_Score_i = \frac{Sim(S_i, q)}{Max(Sim(S_i, q))} \quad (7)$$

However, this feature if close to one, indicating that sentence is more similar to a document title and more appropriate be to placement in the final summary and approach zero whatever lower similarities.

2. Valuable words: This feature is for identifying key words and valuable content of the original document. The words that most frequent (excluding the deterrent words) in the main document, the words are considered Valuable. Thus, first the words frequency table of the original document which deterrent words have been removed, be descending. And then n words of the high of table (n = 50 has been assumed here), select and put in a vector which Vthematic. Then calculated the similarity between this vector and each sentence in the original text based on Equation 8:

$$Sim(s_j, V_{thematic}) = \frac{\sum_{i=1}^{t} w_{i,j} \times w_{i,Vthematic}}{\sqrt{\sum_{i=1}^{t} w^2_{i,j}} \times \sqrt{\sum_{i=1}^{t} w^2_{i,Vthematic}}} \quad (8)$$

Final the value of this feature are normalized as follows:

$$Thematic\_score_i = \frac{Sim(S_i, V_{Thematic})}{Max(Sim(S_i, V_{Thematic}))} \quad (9)$$

However, the parameter score for a sentence further indicates that the sentence is a appropriate and important. But if the score obtained for sentence tends to be zero, ie, sentence is not include useful words of original text and probability placement it's in the final summary, low.

3. Position sentence: Determine the position of the sentences in the original document for identifying the most important part of it is useful. According to studies, the most important sentences in the text, are the first sentence or final sentences and are considered valuable information.

$$Pos\_Score_i = 1 - 0.5 * Sin(3.14 * i/(N-1)) \quad (10)$$

N number of all of the sentences in original text.
I number of sentence is in the main text.

Based on this relationship, if the sentence is placed at the beginning or end of a document, scores it is tends to 1. But if the sentence is located in the middle part of original document, scores will tend to zero and possible placement in final summary is low.

4. Sentence length: Usually long or short phrases are not suitable for the final summary. To account for this feature, we first calculated the length all sentences in the original text, then we calculated the mean lengths obtained that called Lavg. The scores of these features for each sentence in the original document are calculated according to equation 11:

$$Len\_Score_i = \begin{Bmatrix} 1 & x \geq 1 \\ x & x \prec 1 \end{Bmatrix} \quad (11)$$

$$x = \begin{Bmatrix} 1 & l_i = l_{avg} \\ \dfrac{0.5 \times l_{avg}}{|l_i - l_{avg}|} & else \end{Bmatrix} \quad (12)$$

The value of this feature for a sentence,if desire to a zero indicate that the sentence is very short or very long And probability placement its in the final summary, is low.

5. Similarities with the center of gravity: The purpose of this feature, finding the most important sentence, or in other words, the text central sentence. To account for this feature, we need to calculate the similarity between sentences. Then based on the similarity matrix which was previously made, we start working steps: First, the similarity matrix generated called Sim_Matrix. The value of this matrix Sim_matrix[I,j] Indicative the similarity between i, j. Then, the sum of each row are calculated for each sentence as follows:

$$Sim\_Sen_i = \sum_{j=1}^{n} Sim\_Matrix[i, j] \quad (13)$$

After computing, find the maximum value of Sim_Seni and record the location the sentence ie the index of sentence in the variable index_centroid.

Using equation 5, we calculate the similarity of each sentence in the original document with the sentence which its index is index_centroid.

$$centroidSen_i = \sum_{j=1}^{n} Sim(S_j, S_{index\_centroid}) \quad (14)$$

Final the value of this attribute are normalized using the following equation:

$$centroid\_Score_i = \frac{centroid\_Sen_i}{Max(centroid\_sen_i)} \quad (15)$$

Thus, if value of this feature for the sentence tends to 1 this means that the sentence with central sentence is similarity suitable and for placement in the final summary is appropriate, but if the sentence scores tend to be zero, indicating that the similarity with central sentence is low not be suitable for placement in the final summary.

6. Positive keywords: There are terms like "sum", "result", "Introduction" and etc show the importance of sentence.So if any of these terms in sentences, value of this parameter is 1 and the sentence rest 0 is assumed.

7. Negative keywords: There are terms like "other words", "example" and etc shows in the sentence, including this sentence explanation of the previous sentences And so it is relatively less important. if any of these terms in sentences, value of this parameter is 0 and the sentence rest 1 is assumed.

C. The Proposed Neural Network Architecture

In this paper, we use a feed forward multi-layer perceptron neural network to evaluate the summarization parameters we used.

The neural network has three layers: input layer, hidden layer and output layer. The number of units in the input layer depends on the number of input parameters. We will work with 7 characters. The parameters are similarity with title and similarities with the center of the text, sentence length, sentence position, positive and negative words, valuable keywords.

It should be noted that the number of units in the hidden layer neural network has a significant impact on performance.

IV. SYSTEM EVALUATION

To evaluation the system was implemented using a set of data is database Hamshahri.

A. System Evaluation Measures

The two major criterion for measuring the effectiveness of information retrieval systems include precision and recall.

1. Precision: These criterion indicate that whether the number of sentences extracted by summarization system are consistent with the sentence of human summaries and shows whether the number of sentences extracted by the system is good. The value of this criterion is calculated according to equation 16:

$$P = tp/(tp+fp) \quad (16)$$

2. Recall: Call criterion reflects the system how many good sentences, in a final summarization provided and is calculated according to equation 17. We then have:

$$R=(tp+fn)/tp \quad (17)$$

fp: sentences retrieved but irrelevant to human summarization.

tp: sentences retrieved associated with human summarization.

tn: not retrieve sentences, but related to Human summarization.

fn: not retrieve sentences and irrelevant to human summarization.

Other criterion used for evaluation of the F scale.

3. Measure F: Criterion is that the interaction between precision and recall, F measure is the harmonic mean of precision and recall is calculated according to equation 18:

$$F\ Scale = (recall * precision * 2) / (recall + precision) \quad (18)$$

In this paper we have used these three criteria.

B. Data Used To System Evaluation

To evaluate the performance of the system requires standard information body of sentence is Persian. For further progress in the field of summarization of databases Hamshahri for evaluating automated summarization of persian texts we used. This documentation contains 100 files of database Hamshahri and any document containing different lengths from 12 to 22. Summarization of the documents made by a person and are labeled sentences with the importance of placing 1 and less important with 0. The 100 files that are part of the news papers, the neural network training data are the number of sentences in this data is 1545. Then, for each sentences, seven features were extracted and results with labels 0 and 1 for training the neural network and used from the trained neural network to test. After this stage, the 20 files containing 397 sentences for testing were evaluated by 4 experts, and a summary of the document have been prepared. In other words, each the document have four experts and then apiece study the original document and such before labeled with 0 and 1. After that, we gave each of sentence 20 to file the neural network and the neural network input features for each sentence decided that there should be included in the summary (1) or not (zero) then summaries produced by the system with summarization the four person were compared.

C. Evaluation Results the Proposed System

In this section we discuss results obtained from the evaluation of the proposed system to text summarization. After determining Importance sentence of any document by 4 expert and results were calculated by the neural network and placing a threshold value of 30% the calculations based on compared to human summarization that there is in data set and summary that our system has produced based on the relationships defined in this section formulas 16, 17 and 18 are calculated. Here, we have used a range of 10 to 100 neurons for training data.For example, a total of 20 neurons the graph of performance evaluation as follows:

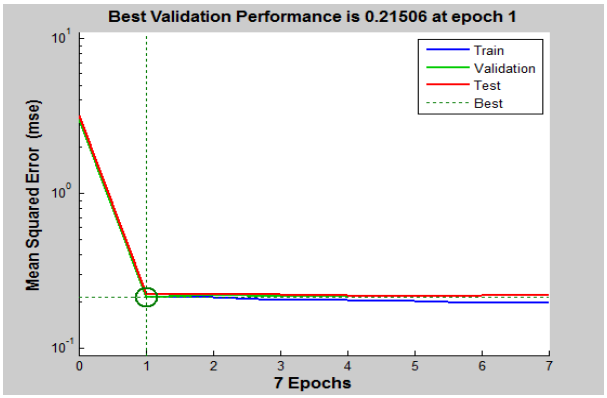

Fig.2. Figure assess the performance of 20 neurons in 7 Steps

Table 1: average results of evaluating proposed method by 4 expert

F measure	Recall measure	Precision measure	Neurons
0.65	0.88	0.52	10

0.67	0.96	0.51	20
0.64	0.85	0.52	30
0.63	0.82	0.52	40
0.63	0.80	0.52	50
0.63	0.84	0.51	60
0.61	0.72	0.53	70
0.63	0.76	0.53	80
0.66	0.89	0.52	90
0.64	0.85	0.51	100

As is observed, the criterion F in 20 neurons with the highest value is 0.67, which means the volume of training data, the value is acceptable.

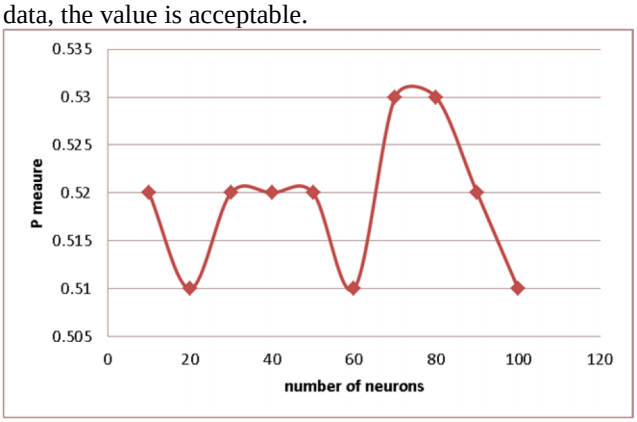

Fig.3. Relationship Diagram Precision measure the number of neurons

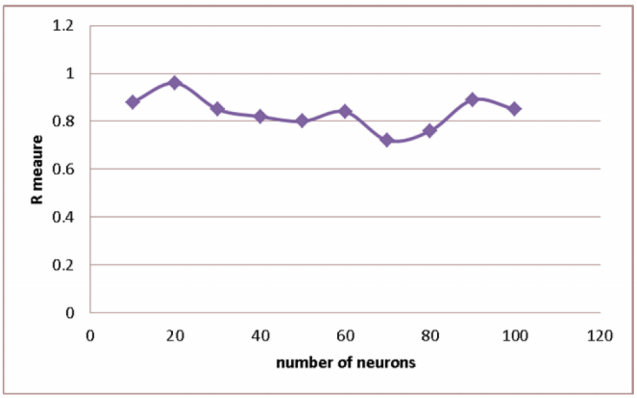

Fig.4. Relationship Diagram recall measure the number of neurons

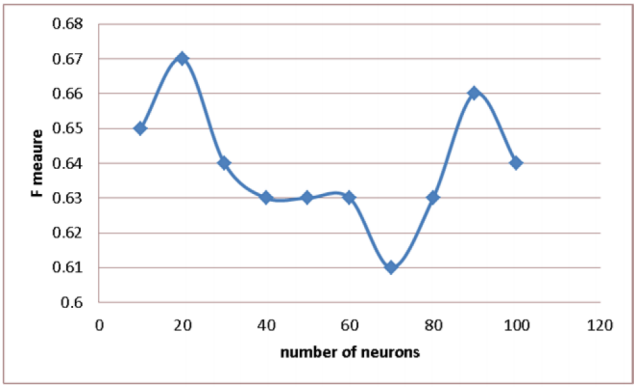

Fig.5. Relationship Diagram F measure the number of neurons

V. Conclusions and Future Work

Due to the increased volume of documents, development summarization systems that have high accuracy, it is necessary. Techniques used in this paper to use neural networks to generate extractive summarization. Select the parameters such as selected sentences for the summarization plays an important role in the efficiency of neural network. In this study of 100 documents that the news database Hamshahri used to for training the neural network. Experimental results on a Hamshahri database documents show that the proposed system is able do to scale F at about 0.67 summarization. In the proposed system, removing less valuable and less important sentences from the original document using neural networks, resulting in a more qualitative summary of the survey findings indicate that important.

For future work, the following are recommended:

• You can increase the amount of training data, better results may be seen. In fact, as more training data we have seen better results in evaluation criteria (criteria of precision, recall and criterion F) will.

• Expanding system for use in summarization of text of multi-document: Multi-document summarization system operates so that the number of articles or documents that are received on a single topic and a summary of the material in all of them documents input.

• The use of other learning algorithms in the context of a summarization: Can learning algorithms such as SVM algorithm and neural network combined with harmony search algorithm can be used to in a single document and Farsi multi-document summarization.

References

[1] Baxendale,P.(1962).”*Machine-made index for technical literature - an experiment”.IBM Journal of Research Developmen, 201-209.*

[2] Edmundson,H. P. (1969). “*New methods in automatic extracting”. Journal of the ACM*, 340-352.

[3] J. Kupiec, J. Pederson and F. Chen,(1995). “*A Trainable Document Summarizer*”, Proceedings of the 18th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, Seattle, Washington, 96-99.

[4] Luhn, H.P,. (1958). “*The automatic creation of literature abstracts”*. IBM Journal of Research Development, pp.159-165.

[5] Mansoorizadeh.M.,taqdiri.E(2009),”An approach based on fuzzy inference system for summarizing Persian text”,11th Iranian Conference on Fuzzy Systems,pp 5-10.

[6] Moshki.M.,Analuee.M(2009),” Multi-document summarization of Persian texts, using a method based on clustering”, 1th National Conference of Software Engineering,pp 1-12.

[7] Qazvinian, V.,Sharif Hassanabadi,L., Halavati,R(2008) .”*Summarization Text with a Genetic Algorithm-Based Sentence Extraction. International of Knowledge Management Studies (IJKMS)”*, Volume 4, Number 2: pp. 426-444.

[8] Shareghi,E.,Sharif,L (2008).”*Text Summarization with Harmony Search Algorithm-Based Sentence Extraction”, Proceedings of The 5th International Conference on Soft Computing as Transdisciplinary Science and Technology*, France. pp. 226-231.

[9] Zahabi.H.,Sharif.L (2011),” new approach based on a harmony search algorithm for summarizing text”, 15th Annual International Conference of Computer Society of Iran,pp 1-8.

AUTHOR'S PROFILE

Sayede Azadeh Hosseinzadeh
is Student at Department of Computer Engineering, Science and Research Branch of Bushehr, Islamic Azad University, Bushehr, Iran. Focus on intelligent systems, web and data mining.
Email: s.a.hoseinzadeh@gmail.com.

Dr. Rouhollah Dianat
is Assistant Professor at Department of Computer Engineering, Qom Center Branch, Qom University, Tehran, Iran.
Email: rouhollah.dianat@gmail.com

Dr. Mohammad Bahrani
is Assistant Professor at Department of Computer Engineering, Tehran Center Branch, Sharif University, Tehran, Iran.
Email: bahrami@ce.sharif.edu