Multi-Document Text Summarization Based on Multiple Linear Regression

Suhad Malallah*,PhD(Asst.Prof.) Zuhair Hussein Ali**,M.Sc.(lecturer)

Abstract

Due to the huge amount of information on the internet makes text summarization growth rapidly. Text summarization is the process of selecting important sentences from documents with keeping the main idea of the original documents. Features considered the basis of text summarization. In this paper a method for assigning a weight to selected features was developed which is depend on building a mathematical model using Multiple Linear Regression which estimate the weights between dependent and independent variables. The proposed model is evaluated using dataset supplied by the Text Analysis Conference (TAC-2011)_for English documents. The results were measured by using Recall-Oriented Understudy for Gisting Evaluation(ROUGE. The obtained results support the effectiveness of the proposed model

Keywords: weight feature, Muliple Linear Regression, dependent and independent variables.

* University of Technology
** College of Education, Al- Mustansiriya university

1. Introduction

According to the fast development of information-communication technologies, enormous quantity of documents have been created and put together in the World Wide Web. The huge amount of documents makes it difficult for the user to get useful information [1] . To deal with such problem of information overload, Automatic Text Summarization (ATS) has been used as a solution. ATS is the process of generating a single document summary from a set of documents or from a single document without losing its main ideas [2]. This process helps users to the general review of all related documents and interested issues with understanding the main content of the summarized documents; this process also helps to reduce the time needed to get these briefs. Rely on the amount of document to be summarized ATS can be classified as a Single Document summarization (SDS) or Multi Document summarization (MDS). In SDS only one document can be summarized into shorter one, whereas in MDS a set of related documents with same topic is summarized into one shorter summary[3]. Summarization methods, also, can be classified as abstractive summarization and extractive summarization. Aabstractive summarization depends on Natural Language Processing (NLP) strategies, which request deep understanding of NLP techniques to analyze the documents sentences and paragraphs, since some changes have to be done to the selected sentences. Whereas in the extractive summarization, no change is applied to the sentences which are selected to be included in the final summary[4]. Thus abstractive summarization seems to be more difficult and time-consuming than extractive summarization [5]. Also summarization can be categorized as query summarization and generic summarization. In the query based summarization a summary was generated according to the user query, where the documents searched to match with the user query [6]. While generic summarization creates a summary which include the main content of the documents. One of the most challenges for the generic summarization is that no topic or query available for the summarization process [7].

The fundamental objective of document summarization is the picking of suitable and pertinent sentence from the input documents. A technique to acquire the suitable sentences is assigned a weight for each sentence which indicates the salience of a sentence for choosing

to the summary and then selecting the top ones [8]. In this paper a method for extracting generic MDS for English text was proposed which depend on extracting seven features for each sentence in the documents, then a mathematical model used for assigning a weight for each feature. The mathematical model based on Multiple Linear Regression (MLR) to estimate the relationship between dependent and independent variables. The estimated parameters represent the weights of the selected features. We have utilized Text Analysis Conference (TAC-2011) dataset are used to assess the summarized results.

2. Related Works

ATS reduces a large number of text documents to a smaller set of sentences which explain the main ideas of these documents. Specialists in NLP are more interested to discover new methods for summarizing and explore a variety of models to come up with perfect summarization. In this section we investigate some of these methods [9].

Binwahlan,Salim, and Suanmali in 2009 [10]. Suggested a method for calculating the weights for the selected features. They used five different features, Where the first two features are combined more than one simple feature to produce one structural feature, while the three remaining features are simple features. These five selected features produced as input to the particle swarm optimization (PSO) which used to train these features and assign a weight to each one of them. Their results showed that structural features got average weight higher than simple features. Abuobieda in 2012 [11]. Suggested a method which based on selection of five features these features are sentence position, sentence length, numerical information, thematic words and title feature. The pseudo genetic algorithm was used to train the dataset and assign a weight to each feature. Their results showed that the importance of each feature as in the following order Title feature, sentence position, thematic words, sentence length and numerical information. Ghalehtaki in 2014 [12] a set of features extracted for each sentence. These features used as input to the combination model which consist of Cellular Learning Automata (CLA), PSO and fuzzy logic. The CLA was used to calculate the similarity between sentences to reduce the redundancy. While the PSO was used to set a weight for each feature, then The fuzzy logic used to give scores to the sentences, these scored sentences were arranged in descending order, the sentence with higher score was selected to be

included in the created summary. Saleh and Kadhim in 2016 [13] proposed a method which based on formulating the problem of ATS as a multi-objective optimization (MOO). Where there is two main objective function redundancy reduction and content coverage.

The redundancy reduction was computed using cosine similarity between each sentence in the dataset, whereas the content coverage was computed using the cosine similarity between each sentence with the mean of sentences centrality. Evolutionary Algorithm used to combine these two objective functions which aim to minimize the first objective and maximized the second objective function. A good results obtained from their proposed method.

3. The Proposed Method

In this paper an MLR is used for MDS. This method depends on extracting seven features for each sentence in the dataset, these extracted features produced as input to MLR model to obtain weights for each one of them. There are two phases in this method: Training phase and testing phase. In the training phase a set of document used to build a model and extract weights for each feature. Whereas in the testing phase the extracted weights used directly with the sentence score to get final summary. Figure (1) shows the block diagram of the proposed method, with five main steps:

1- Preprocessing
2- Features Extraction
3- Constructing an MLR model and assign weights for each feature
4- Sentence Ranking and ordering
5- Remove Redundancy

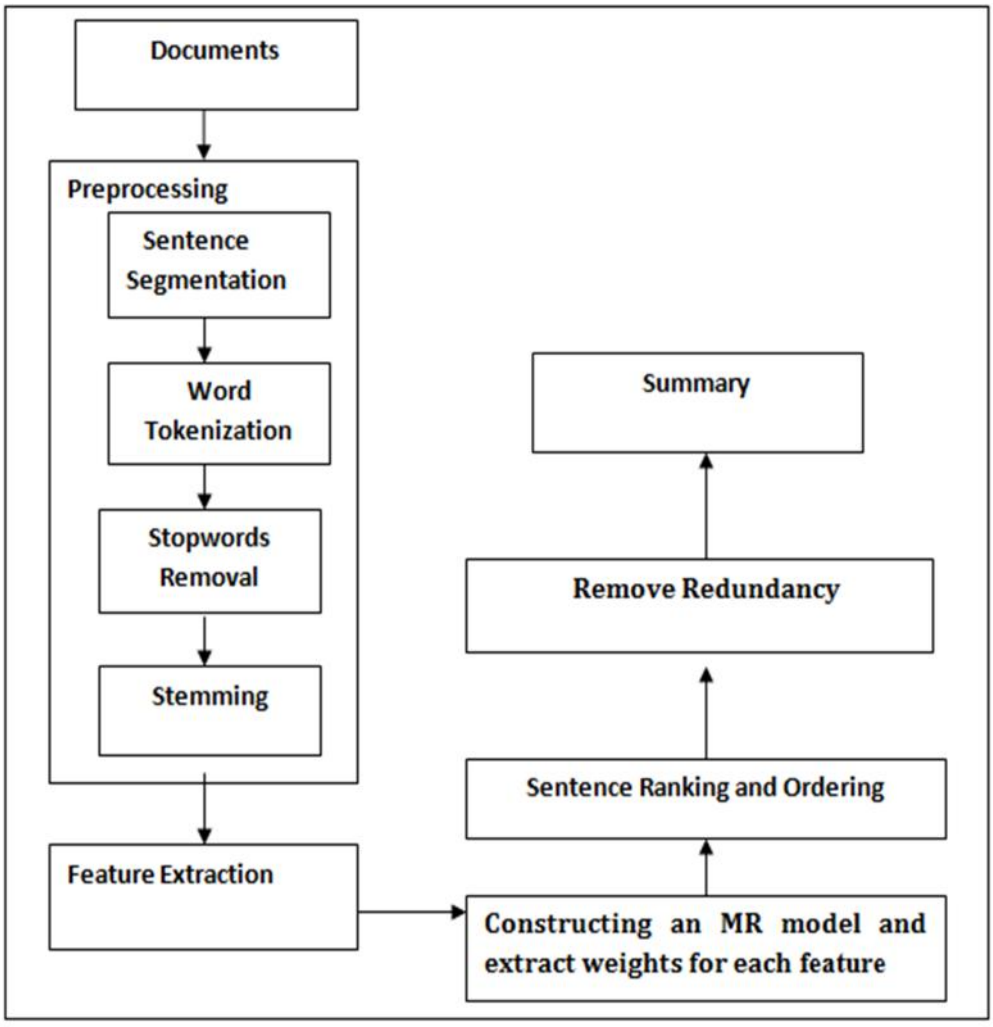

Figure (1) Block Diagram for Proposed MDS

3.1 Preprocessing

The preprocessing is the first stage in the proposed MDS. This stage consists of four steps for preparing the data, these steps are[14]:-

A- **Segmentation** : Sentence segmentation is an important approach in text processing it can be done by separating sentences according to the dot between sentences.

B- **Tokenization**: Is the process of splitting sentence into words

C- **Stop Words Removal**: Words which don't give the necessary information for identifying significant meaning of the document content and appear frequently are removed. There are a variety of methods used for specifying of such stop words list. . Presently, a number of English stop word list is usually used to help text summarization process

D- **Stemming**: is the process of producing root of the word, in This paper word stemming is performed by removing suffixes proposed by Porter's stemming algorithm [15].

3.2 Features Extraction

It's an important part of ATS, which include compute of features score for every sentence. These features include sentence position, sentence length, numerical data, Thematic word, title word, proper noun and centroid value[14].

A-: Sentence Postion (SP): Where the higher score will give to the first sentence, and the score decreases according to the sentence position in the document. This feature can be computed according to equation (1).

$$F1 = \frac{N - P + 1}{N} \quad (1)$$

Where N represents the number of sentences in the document
P current position of the sentence

B- Sentence length (SL): This feature is computed by dividing the sentence length by the length of longest sentence in the document as in equation (2).

$$F2 = \frac{Sentence\ length}{Longest\ sentence\ length} \quad (2)$$

C- Numerical data (ND): The appearance of this feature in the document has important information to be included in the summary. This feature calculated by dividing the number of numerical data in the sentence by the sentence length. This feature is calculated as in equation (3).

$$F3 = \frac{Number\ of\ numerical\ data\ in\ the\ sentence}{Sentence\ length} \quad (3)$$

D- Thematic Words (TW): is the term that appears most frequently in the document. This feature can be calculated by computing the repetition of all terms in the document, then top (n) terms with the highest repetition is selected, in this research, we used top (5). This feature is calculated by dividing the number of thematic words in the sentence by the maximum thematic words in the document as explained in equation (4).

$$F4 = \frac{NO.Of\ Thematic\ words\ in\ the\ sentence}{MAx\ NO.\ of\ Thematic\ in\ the\ document} \quad (4)$$

E-Title Feature (TF): This feature is important when summarizing the document. The score was calculated as in equation (5).

$$F5 = \frac{No.\ of\ title\ word\ in\ the\ sentence}{Title\ length} \quad (5)$$

F-Proper Noun (PN): The sentence is important if it includes the maximum number of proper nouns [16]. This feature is calculated as in equation (6)

$$F6 = \frac{No.of\ proper\ noun\ in\ the\ sentence}{Sentence\ length} \quad (6)$$

G-Centroid value (CV): Is a feature used to specify salient sentences in the multiple documents [17]. This feature can be calculated as follows

$$F7 = \sum_{i=1}^{n} C_{wi} \quad (7)$$

$$C_{wi} = TF * IDF \quad (8)$$

$$IDF = \log\left[\frac{total\ NO.\ documents}{No.of\ documents\ containing\ the\ given\ word}\right] \quad (9)$$

Where

C_w is the centroid value of the words.

TF is the term frequency which represents the frequency of a given term in the document.

IDF is the inverse term frequency computed by division of the total number of documents and the number of documents including the given term.

3.3 Multiple Linear Regression

MLR is a statistical method for formulating the relationship between the independent variables and a dependent variable. Where there are two or more independent variables, but only one dependent variable [18]. MLR can be formulated as in equation (1)

Y= W0+W1X1+W2X2+---+XnWn (10)

Where

[Y] is the output (dependent variable).

[W] the weights for each extracted features.

[X] the extracted features (independent variables).

The regression model can represent in matrix form as follows.

$$\begin{bmatrix} Y1 \\ Y2 \\ . \\ . \\ Yn \end{bmatrix} = \begin{bmatrix} X01 & X02 & X03 & X04 & X05 & X06 & X07 \\ X11 & X12 & X13 & X14 & X15 & X16 & X17 \\ . & . & . & .. & .. & .. & .. \\ .. & ... & .. & ... & .. & ... & .. \\ Xn1 & Xn2 & Xn3 & Xn4 & Xn5 & Xn6 & Xn7 \end{bmatrix} \begin{bmatrix} W1 \\ W2 \\ . \\ . \\ W7 \end{bmatrix}$$

Where n is the number of sentences from the collected document data set. There are two modes in our proposed MDS. Training mode and testing mode.

A- Training mode: Where there are 70 documents from the TAC-2011 dataset used for training our model. The seven extracted features (X1, X2,…, X7) that described in section 3.2 used as input to the model. Y can be computed by using cosine similarity between all

sentences from selected trained documents and each sentence from the manually summarized documents. As in equation (11)

$$\cos(Si, Sj) = \frac{(Si.Sj)}{||Si||Sj||} \quad (11)$$

Where Si sentence from trained documents

Sj sentence from manually summarized documents.

Si compared with all sentences from the manually summarized sentences and higher score value assign to Yi. The score of the equation (11) ranges from (0) to (1), zero score where there is no any matching between Si and Sj whereas one where Si identical to Sj. Thus we have values of (Xi) and (Y) for equation (10) , our goal is to estimate the values of (Wi) which represent the weights of the selected features [19]. W can be calculated as described below.

$W=(X.X^t)^{-1}.X^tY$ (12)

B- Testing mode: where there are 100 English documents from TAC-2011 used as input in this mode.

3.4 Sentences Ranking

In this stage the seven extracted features and the weights of each feature that extracted from the equation (12) used directly to compute the score of every sentence in the documents collection. The score can be calculated as in equation (13)

$$Score(S) = \sum_{i=1}^{7} Fi(S) * Wi \quad (13)$$

All sentences are ranked in descending order depending on their scores. Sentences with higher scores are selected to be included in the document summary, depending on document summary size, we specify the number of sentences to be chosen.

3.5 Remove Redundancy

One of the main problems of any MDS is the redundancy. Since there are many documents with same topics some sentences may be repeated in more than one document therefore remove redundancy is a very important part on any ATS. The equation used to compute redundancy between two sentences as in [20].

$$R = \frac{2 \square Ms}{M1 + M2} \quad (14)$$

Where R number of redundancy between two sentences
Ms number of similar words between two sentences
M1 number of words in sentence one
M2 number of words in sentence two.
If the redundancy between selected sentence to be included in the summary and one of the sentences that already exist in the summary greater than threshold this sentence will be ignored.
The following algorithm illustrates how we remove redundancy in our MDS model.

input 1- set of ranked sentences in desending order called score_sent

2- Max summary size called Max_size

output generated summary called summary

Step1: let summary ={}

Step2 : from Score_sent select Si with highest score

Step3 : compare Si with all Sentences in the summary according to Eq.(11)

If (cos(si,sj) >threshold) delete Si go to step2

Else Put Si in the summary

Step4: if summary size<max_size goto step2

Else end

4. Dataset and Evaluation Metrics

The dataset used in our proposed method is the TAC-2011 which consist of seven languages (English, Arabic, Greek, Czech, French, Hindi, Hebrew). There are10 topics, each of 10 documents for each language [21]. Our proposed method deal with English language only.

ROUGE [22] is used to evaluate the proposed system the outputs of rouge package are three numbers which represent, precision (P), Recall (R) and F−score. They formulated as follows.

$$P = \frac{Sentence\ \epsilon human\ summary \cap Sentence\ \epsilon\ system\ summary}{Sentence\ \epsilon\ system\ summary} \quad (15)$$

$$R = \frac{Sentence\epsilon\ human\ summary \cap Sentence\ \epsilon\ system\ summary}{Sentence\ \epsilon human\ summary} \quad (16)$$

$$F = \frac{(1+\beta^2)R * P}{R + \beta^2 P} \quad (17)$$

Where

$$\beta = \frac{P}{R}$$

5. Experimental Results

There are two main purposes in the proposed method for MDS. The first purpose is to compute the weight of each selected features which indicates the importance of these features. Figure (2) shows the weights of each feature in our proposed method. From the results we can see the order of effective features weights as follows CV, TF,SP, PN ,TW,SL and ND.

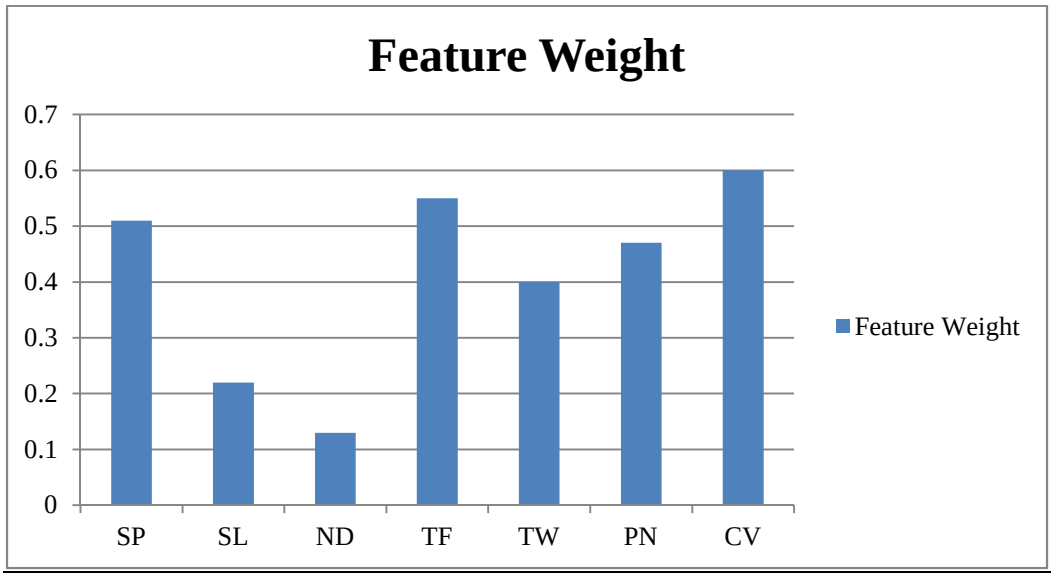

Figure (2) Features weights

The second purpose of our proposed method is to evaluate these computed weights to calculate the score of each sentence to select the highest sentences to be included in the final summary. The results compared with the peer summary (summary that generated by the system) of TAC-2011 data. Figure (3) shows the results using ROUGE-1, the proposed method gives better results compared with the peer summary results.

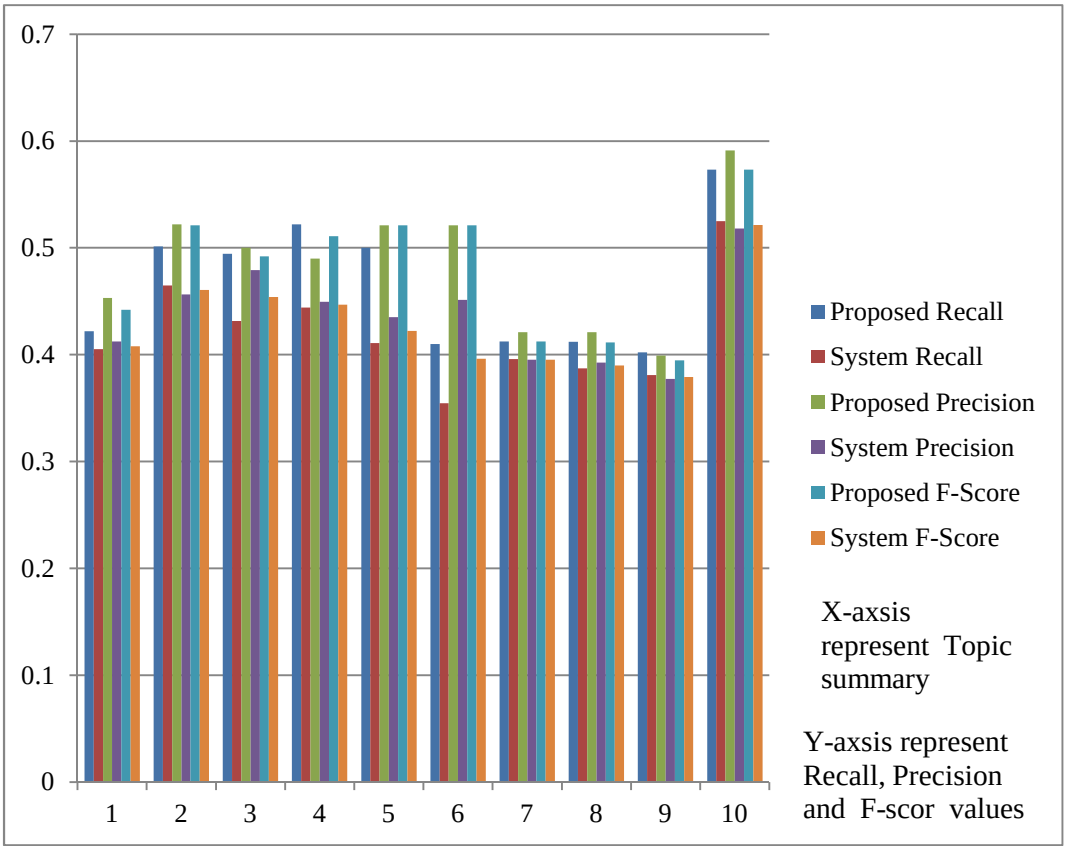

Figure (3) Evaluation measures

6. Conclusions

In our proposed method for MDS we introduced a method for assigning a weight to the selected features depending on a mathematical model MLR, which consider an effective way to estimate the weights between dependent and independent variables. We used TAC-2011 which consist of 100 documents to test our proposed method. We saw that CV, TF, SP, PN and TW are more important than SL and Nd. The results show the effect of the weights to improve final MDS compared with the peer summary of the system. Where the score of the sentences increased and effect on the results.

References

[1] R. M. Alguliev, R. M. Aliguliyev, and N. R. Isazade, “Expert Systems with Applications Multiple documents summarization based on evolutionary optimization algorithm,” *Expert Syst. Appl.*, vol. 40, no. 5, pp. 1675–1689, 2013.

[2]R. Kumar and D. Chandrakal" A survey on text summarization using optimization algorithm,” ELK Asia Pacific Journals vol. 2, no. 1, 2016.

[3] L. Huang, Y. He, F. Wei, and W. Li, “Modeling Document Summarization as Multi-objective Optimization,” pp. 2–6, 2010.

[4] W. Song, L. Cheon, S. Cheol, and X. Feng, “Expert Systems with Applications Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9112–9121, 2011.

[5] S. A. Babar and P. D. Patil, “Improving Performance of Text Summarization,” *Procedia Comput. Sci.*, vol. 46, no. Icict 2014, pp. 354–363, 2015.

[6] W. Song, L. Cheon, S. Cheol, and X. Feng, “Expert Systems with Applications Fuzzy evolutionary optimization modeling and its applications to unsupervised categorization and extractive summarization,” *Expert Syst. Appl.*, vol. 38, no. 8, pp. 9112–9121, 2011.

[7] R. Ferreira, L. De Souza, R. Dueire, G. De Frana, F. Freitas, G. D. C. Cavalcanti, R. Lima, S. J. Simske, and L. Favaro, “Expert Systems with Applications Assessing sentence scoring techniques for extractive text summarization,” *Expert Syst. Appl.*, no. May, 2013.

[8] S. S. Megala and A.Kavitha “Feature Extraction Based Legal Document Summarization,”ijarcsms, , pp. 346–352, 2014.

[9] W. Luo, F. Zhuang, Q. He, and Z. Shi, “ Exploiting relevance , coverage , and novelty for query-focused multi-document summarization,” *Knowledge-Based Syst.*, vol. 46, pp. 33–42, 2013.

[10] M. S. Binwahlan, N. Salim, and L. Suanmali, “Swarm Based Features Selection for Text Summarization,” IJCSNS , vol. 9, no. 1, pp. 175–179, 2009.

[11] A. Abuobieda, N. Salim, A. T. Albaham, A. H. Osman, and Y. J. Kumar, “Text summarization features selection method using pseudo genetic-based model,” *Proc. - 2012 Int. Conf. Inf. Retr. Knowl. Manag. CAMP’12*, pp. 193–197, 2012.

[12] R. A. Ghalehtaki, “A combinational method of fuzzy , particle swarm optimization and cellular learning automata for text summarization,”,IEEE conference, vol. 15, no. 1, 2014

[13] H. H. Saleh1, N. J. Kadhim" Extractive Multi-Document Text Summarization Using Multi-Objective Evolutionary Algorithm Based Model", *Iraqi Journal of Science, Vol. 57, No.1C, pp: 728-741*,2016.
[14] ANSAMMA JOHN, “Multi-Document Summarization System: Using Fuzzy Logic and Genetic Algorithm,” Int. J. Adv. Res. Eng. Technol., vol. 7, no. 1, pp. 30 – 40 , 2016.
[15] porter stemming algorithm:
http://www.tartarus.org/martin/PorterStemmer

[16] C. N. Satoshi, S. Satoshi, M. Murata, K. Uchimoto, M. Utiyama, and H. Isahara, “Sentence Extraction System Assembling Multiple Evidence,” Proc. 2nd NTCIR Work., pp. 319–324, 2001.
[17] A. John and D. M. Wilscy, “Random Forest Classifier Based Multi-Document Summarization System,” IEEE Recent Adv. Intell. Comput. Syst. RANDOM, pp. 31–36, 2013.

[18] S. CHATTERJEE and A.HADI "Regression analysis by example", A JOHN WILEY & SONS, INC., PUBLICATION fourth edition , 2006,

[19] Z. Timothy, "*Multiple Regression and Beyond: An Introduction to Multiple Regression and Structural Equation Modeling*", by Routledge, Second edition published 2015.
[20] D. R. Radev, H. Jing, M. Sty, and D. Tam, “Centroid-based summarization of multiple documents,” Information Processing and Management , vol. 40, pp. 919–938, 2004.
[21]] G. Giannakopoulos, M. El-Haj, B. Favre, M. Litvak, J. Steinberger, and V. Varma, “TAC 2011 MultiLing Pilot Overview,” no. November, 2011.

[22] C. Y. Lin, “Rouge: A package for automatic evaluation of summaries,” Proc. Work. text Summ. branches out (WAS 2004), no. 1, pp. 25–26,2004.

تلخيص النصوص المتعدده بالاعتماد على الانحدار الخطي المتعدد

أ.م .د سهاد مال الله * م.زهير حسين علي**

المستخلص

بالنظر للكميات الكبيره الموجوده من المعلومات في الانترنت ادى ذلك الى الحاجه الضروريه لتلخيص المعلومات. أن عمليه تلخيص المعلومات تتضمن أستخراج الجمل المهمه من النصوص مع المحافظه على الافكار الرئيسيه للنصوص الملخصه. في هذا البحث تم بناء نموذج رياضي اعتمادا على **الانحدار الخطي** المتعدد وذلك بتقديرالأوزان للخواص المتعدده المستخرجه من كل جمله من جمل النصوص. تم أستخدام قاعدة البيانات (TAC-2011) لتجربة النموذج الرياضي المقترح. أختبارات النتائج باستخدام **برنامج** ROUGE **أثبتت كفاءة الظلم المقترح**

*الجامعة التكنولوجية
**الجامعة المستنصرية