Hybrid Proposed Method Using Statistical Linguistic Features (SLF) And Neural Network In Arabic Texts Summarization

Iman Qays Abduljalee *Amal Hameed Khaleel* *suhad muhajer kareem*

Basra university \ Science collage \Computer science dept.

emankais@yahoo.com

Abstract

With the counting growth of electronic information on World Wide Web, it has become necessary important to provide mechanisms to find and present a shorter version would suffice, so that automatic text summarization technique which plays an important role to help users to determine whether it has to do with information they need or not.

In this paper we present a short historical overview and advancement of automatic text summarization and the most relevant approaches currently used in this area. We proposed a new technique in automatic summarization area for Arabic news articles using a neural network by selecting important sentences from the original text and put it in the summary. A Multi-layer Perceptron neural network (MLP) is trained to learn the relevant characteristics of sentences that should be included in the summary of the article. The neural network is then modified to generalize and combine the relevant characteristics apparent in summary sentences . Finally, system is evaluated by comparing the final summary of the system with the summary produced by expert in Arabic language, we measure the performance of the system by computing the precision, recall and F-Measure and we obtain good result that we display in the conclusion.

Keywords: Text summarization, Multi-layer Perceptron Neural Network (MLPNN), F-Measure.

I. Introduction

Text summarization has become an important and timely tool for assisting and interpreting text information in today’s fast-growing information age. It is very difficult for human beings to manually summarize large documents of text. There is an abundance of text material available on the internet(V. Gupta *et al*.2010). Text summarization is the process to produce a condensed representation of the content of its input for human consumption (S.D. Afantenos *et al*.2005).

The input to a summarization system can be one or more text documents, When only one document is the input, it is called single document text summarizeation and when the input is a cluster of related text documents, it is called multidocument summaryzation. We can also categorize the text summarization based on the type of users the summary is intended for query focused summaries are tailored to the requirements of a particular user or group users and generic summaries are aimed at a broad readership community (K. Sarkar 2009). Text Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method consists of selecting important sentences, paragraphs etc. from the original document and concatenating them into shorter form. The importance of sentences is decided based on statistical and linguistic features of sentences. An Abstractive summarization attempts to develop an understanding of the main concepts in a document and then express those concepts in clear natural language. It uses linguistic methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document. This paper focuses on extractive text summarization methods. Extractive summaries are formulated by extracting key text segments (sentences or passages) from the text, based on statistical analysis of individual or mixed surface level features such as word/phrase frequency, location or cue words to locate the sentences to be extracted. The "most important" content is treated as the "most frequent" or the "most favorably positioned" content. Such an approach thus avoids any efforts on deep text understanding. They are concept-ually simple, easy to implement (V. Gupta *et al.*2010, J. Lin 2009, K. Jezek and J. Stteinberger 2007).

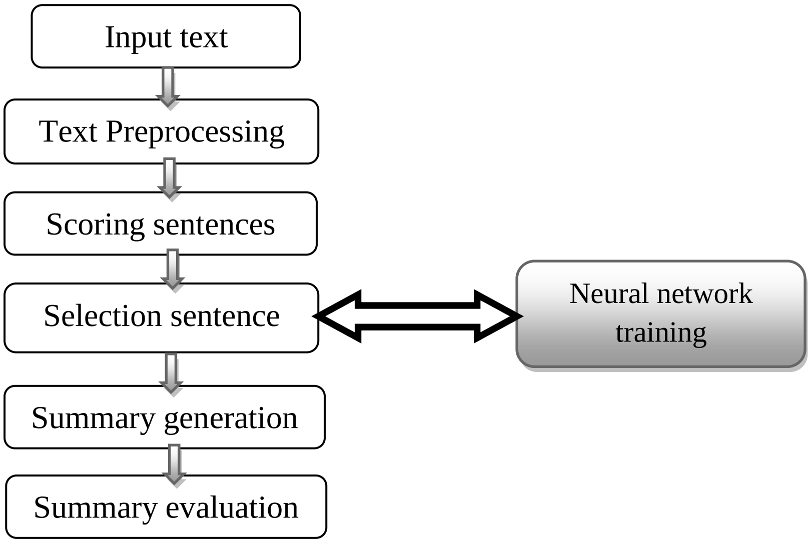
II. Related Works

Most early work on single summarization system focused on frequency word proposed by luhn in 1958(H. P. Luhn 1958), his research based on high frequency to find important parts from original document to generate summary. In 1969, where edmunson used cue words and the similarity to the title as two features to extract salient parts from text. In 2001 lin & gon used latent semantic analysis to extract the meaning from the words and the sentences by using model of singular vector decomposition (SVD) (J. Steinberger 2001). A new method used in text summarization using graph-based method proposed by Radav & Mihalca in 2004 to present the text in graph architecture where nodes are the sentence and the edges are the similarity between sentences (R. Mihalcea 2004).

After that a numbers of learning techniques are proposed and used in text summarization such as: kupiek in 1995 used Bayesian classifier to compute the probability of each sentence in the text to appeared in the summary(J. Kupiek *et al*. 1995), Khosrow kaikhah in 2004 used neural networks to train and learn the relevant features of sentences that should be included in the summary of text documents (Khosrow 2004),in 2011 Ben King *et al.* proposed extractive text summarization using neural network(Ben King *et al.* 2011), and numbers of researches are still proposed in text summarization.

III. The Proposed Method Architecture

In this papers, we proposed a new method to generate summary for article Arabic texts by using neural network where the method consist of numbers of stages to generate summary and we shown in the following diagram figure (1):-



Figure(1) the proposed method architecture

The model consist of the following stages:-

1- Text Preprocessing stage:- this stage consist of :

1-1 tokenization:- process to determines the boundaries of each word and sentence in the original text.

1-2 stemming:-

1-3 we obtain the stem of each word by removing suffixes from this word such as (يدرس → درس ، يذهبون → ذهب ،….est.)

1-4 part of speech (word tagging):- is the process of assigning POS to each word in the sentence to give the class of this word such as (noun, verb, adjective,…ect). This process is depend on the lexical stored in the system.

1-5 stop words removal:- is the process to filter the text from the most frequent important words by using list of stop word stored in the system such as (التي ، ....، الذي، ماذا، الجدير بالذكر ect). The advantage of removing this word to increasing the performance of computing scoring of the sentence.

2- Sentences Scoring stage:-is the process to compute the weight of each sentence based on number of features we are called as SLF(Statistical Linguistic Features) features of texts, each sentence is presented as vector [F_1,F_2,F_3,F_4, F_5, F_6, F_7], where:-

***F_1* (frequency word feature) :-** *compute number of the occurrences of words in full original text.*

$$F1= \sum_{i=1}^{n} frequency(w)$$

***F_2* (length sentence feature):-** *compute number of words within its sentence without stop of word.*

$$F2=\sum_{i=1}^{n} no.of\ words(s)$$

***F_3* (position sentence feature):-** *the location sentence in the text.*

$$F3= \frac{1}{position(s)}$$

***F_4* (no. Verbs feature):-** *compute the number of verbs in the sentence of input text based on*

POS.

F2=$\sum_{i=1}^{n} no.of\ verbs(s)$

***F**$_5$ **(no. proper names feature):-** compute the number of proper names in the sentence of input text based on POS.* *F2*=$\sum_{i=1}^{n} no.of\ proper\ names(s)$

***F**$_6$ **(no. digits feature):-** compute the number of digits in the sentence of input text.* *F2*=$\sum_{i=1}^{n} no.of\ digits(s)$

***F**$_7$ **(similarity to title feature):-** compute the number of common words between the word of sentence and the word of title of input text. F7==*Σ SIM(Si , Sj) i>=j

3- Sentences Selection stage:- we used neural network method that is performed in two different phases namely, training phase and testing phase. The first phase of the process involves training the neural networks to learn the types of sentences that should be included in the summary. This is accomplished by training the network with sentences where each sentence is identified as to whether it should be included in the summary or not. This is done by a human reader. We use a three-layered feedforward MLP neural network. Our neural network consists of seven input-layer neurons, five hidden-layer neurons, and one output-layer neuron (as shown in figure 2). We uses the sigmoid activation function for both the weights of the hidden layer and the output layer.

<u>Algorithm</u>

i. Create an architecture consists of three layers: input, hidden, and output layers, for input layer, assign as net input to each unit (x_i, i=1,…,7) its corresponding features in the input vector. The output for each unit is its net input.

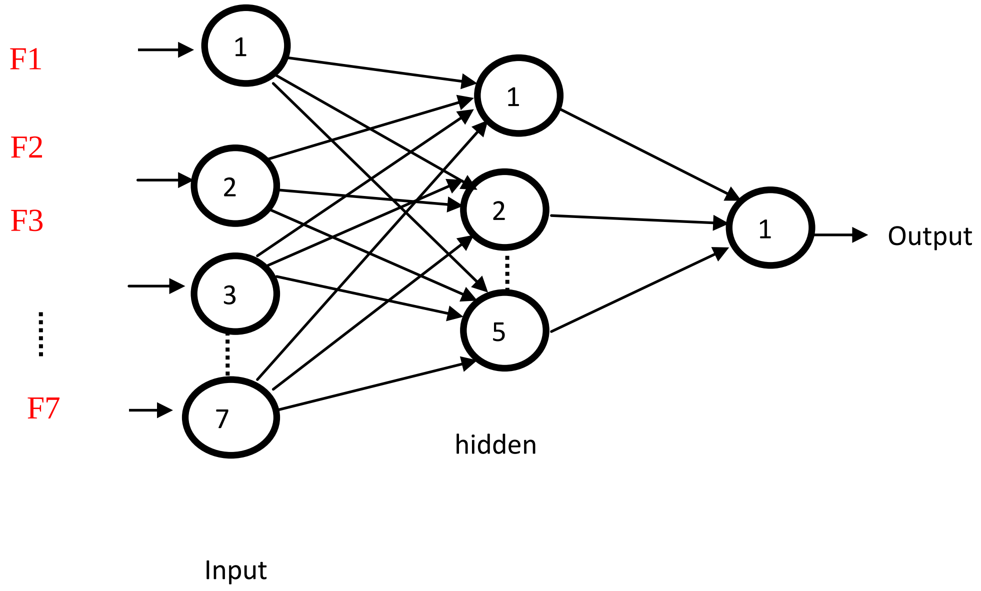
ii. Initialize the weights and bias to random values .

iii. Initialize the network parameters. (Learning rate =0.3, Momentum rate =0.7)

iv. Train the network with initialized parameters, and with sigmoid activation function.

v. Calculate the error using MSE method

vi. Repeat the process until the maximum epochs are reached or the desired output is identified or the minimum gradient is reached. (Maximum epochs = 10000, Total error = 10^{-6})



Figure(2) :-Structure of Multi-layer Perceptron Neural Network

4- Summary generation stage:- In this step, after the scores of sentences were selected by neural network to appear in the summary, we use the following steps to generate the summary:-

*Summary length = Summarization Rate *(Document Length) /100* ….(1)

- Sort this scores of selected sentences by neural network in decreasing order and select n top of this score, the value of n depend on summary length which computed from the following formula:-

In this paper, we used summarization rate is 45%.

- Sentences may not be ordered in the correct order so we should ranked it in order as the order appeared in the original texts and we put title of the original text as the title of the summary. The advantage of this step to make summary more readable and coherent.

5- Summary evaluation stage:- summaries can be evaluated using intrinsic or extrinsic measures, while intrinsic methods attempt to measure summary quality using human evaluation therefore, extrinsic methods measure the same through a task-based measure such the information retrieval-oriented task(D. Mallett, J. Elding and Mario 2004).

In this work we used ROUGE (Recall-Oriented Understudy Gisting Evaluation) measure is one of most intrinsic methods to evaluate summary which written by human(expert human) with automatic summary to measure performance of the summary. This measure produce numerical value by use precision(P),recall(R) and F-Measure(F)(R. M. ALGuliev and R. M. ALGullyev 2007):-

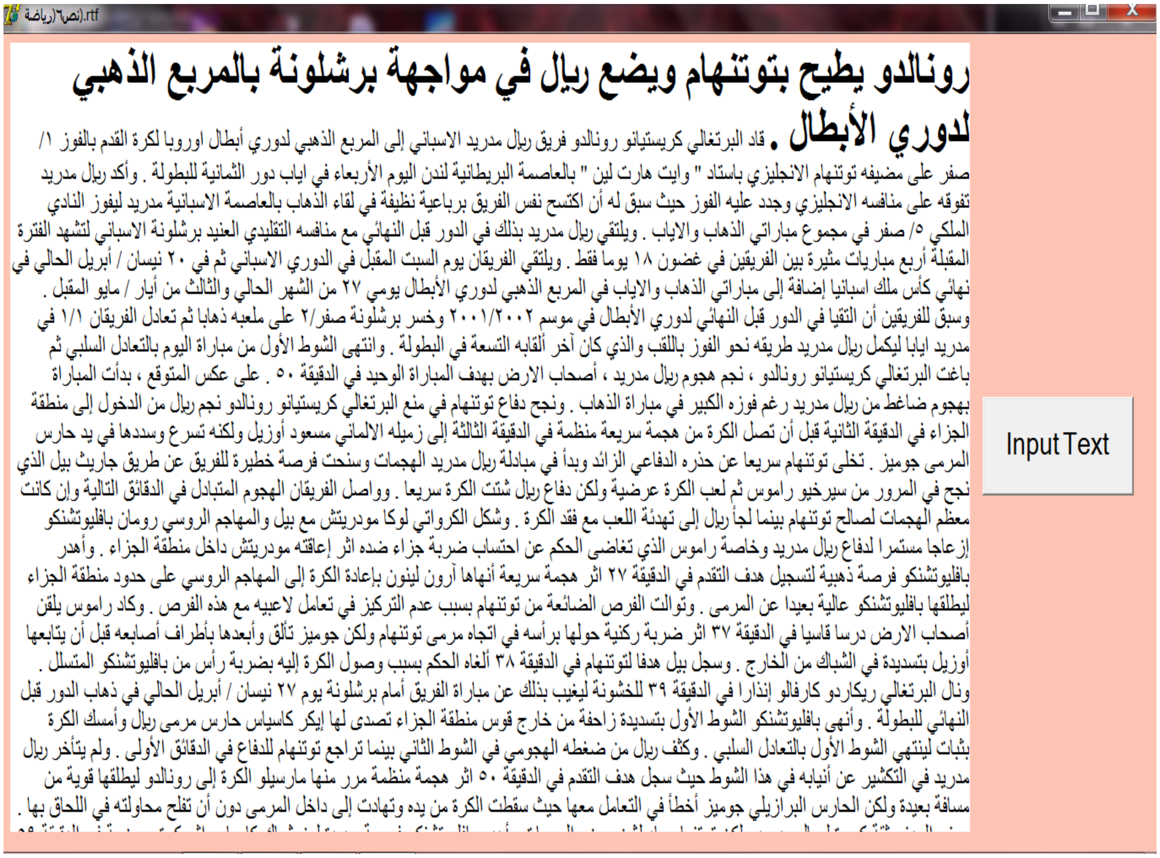
$$P= \frac{Sman \cap Sauto}{Sauto} \qquad \ldots(2)$$

$$R= \frac{Sman \cap Sauto}{Sman} \qquad \ldots.(3)$$

$$F\text{-}Measure = \frac{2\times P\times R}{P+R} \qquad \ldots(4)$$

Where :- ***Sman*** is the manual summary and ***Sauto*** is the automatically-generated summary.

Examples :-the input text to summarize it.



رونالدو يطيح بتوتنهام ويضع ريال في مواجهة برشلونة بالمربع الذهبي لدوري الأبطال . قاد البرتغالي كريستيانو رونالدو فريق ريال مدريد الاسباني إلى المربع الذهبي لدوري أبطال اوروبا لكرة القدم بالفوز ١/
صفر على مضيفه توتنهام الانجليزي باستاد " وايت هارت لين " بالعاصمة البريطانية لندن اليوم الأربعاء في اياب دور الثمانية للبطولة . وأكد ريال مدريد
تفوقه على منافسه الانجليزي وجدد عليه الفوز حيث سبق له أن اكتسح نفس الفريق برباعية نظيفة في لقاء الذهاب بالعاصمة الاسبانية مدريد ليفوز النادي
الملكي ٥/ صفر في مجموع مباراتي الذهاب والاياب . ويلتقي ريال مدريد بذلك في الدور قبل النهائي مع منافسه التقليدي العنيد برشلونة الاسباني لتشهد الفترة
المقبلة أربع مباريات مثيرة بين الفريقين في غضون ١٨ يوما فقط . ويلتقي الفريقان يوم السبت المقبل في الدوري الاسباني ثم في ٢٠ نيسان / أبريل الحالي في
نهائي كأس ملك اسبانيا إضافة إلى مباراتي الذهاب والاياب في المربع الذهبي لدوري الأبطال يومي ٢٧ من الشهر الحالي والثالث من أيار / مايو المقبل .
وسبق للفريقين أن التقيا في الدور قبل النهائي لدوري الأبطال في موسم ٢٠٠١/٢٠٠٢ وخسر برشلونة صفر/٢ على ملعبه ذهابا ثم تعادل الفريقان ١/١ في
مدريد ايابا ليكمل ريال مدريد طريقه نحو الفوز باللقب والذي كان آخر ألقابه التسعة في البطولة . وانتهى الشوط الأول من مباراة اليوم بالتعادل السلبي ثم
باغت البرتغالي كريستيانو رونالدو ، نجم هجوم ريال مدريد ، أصحاب الارض بهدف المباراة الوحيد في الدقيقة ٥٠ . على عكس المتوقع ، بدأت المباراة
بهجوم ضاغط من ريال مدريد رغم فوزه الكبير في مباراة الذهاب . ونجح دفاع توتنهام في منع البرتغالي كريستيانو رونالدو نجم ريال من الدخول إلى منطقة
الجزاء في الدقيقة الثانية قبل أن تصل الكرة من هجمة سريعة منظمة في الدقيقة الثالثة إلى زميله الالماني مسعود أوزيل ولكنه تسرع وسددها في يد حارس
المرمى جوميز . تخلى توتنهام سريعا عن حذره الدفاعي الزائد وبدأ في مبادلة ريال مدريد الهجمات وسنحت فرصة خطيرة للفريق عن طريق جاريث بيل الذي
نجح في المرور من سيرخيو راموس ثم لعب الكرة عرضية ولكن دفاع ريال شتت الكرة سريعا . وواصل الفريقان الهجوم المتبادل في الدقائق التالية وإن كانت
معظم الهجمات لصالح توتنهام بينما لجأ ريال إلى تهدئة اللعب مع فقد الكرة . وشكل الكرواتي لوكا مودريتش مع بيل والمهاجم الروسي رومان بافليوتشنكو
إزعاجا مستمرا لدفاع ريال مدريد وخاصة راموس الذي تغاضى الحكم عن احتساب ضربة جزاء ضده اثر إعاقته مودريتش داخل منطقة الجزاء . وأهدر
بافليوتشنكو فرصة ذهبية لتسجيل هدف التقدم في الدقيقة ٢٧ اثر هجمة سريعة أنهاها آرون لينون بإعادة الكرة إلى المهاجم الروسي على حدود منطقة الجزاء
ليطلقها بافليوتشنكو عالية بعيدا عن المرمى . وتوالت الفرص الضائعة من توتنهام بسبب عدم التركيز في تعامل لاعبيه مع هذه الفرص . وكاد راموس يلقن
أصحاب الارض درسا قاسيا في الدقيقة ٣٧ اثر ضربة ركنية حولها برأسه في اتجاه مرمى توتنهام ولكن جوميز تألق وأبعدها بأطراف أصابعه قبل أن يتابعها
أوزيل بتسديدة في الشباك من الخارج . وسجل بيل هدفا لتوتنهام في الدقيقة ٣٨ ألغاه الحكم بسبب وصول الكرة إليه بضربة رأس من بافليوتشنكو المتسلل .
ونال البرتغالي ريكاردو كارفالو إنذارا في الدقيقة ٣٩ للخشونة ليغيب بذلك عن مباراة الفريق أمام برشلونة يوم ٢٧ نيسان / أبريل الحالي في ذهاب الدور قبل
النهائي للبطولة . وأنهى بافليوتشنكو الشوط الأول بتسديدة زاحفة من خارج قوس منطقة الجزاء تصدى لها إيكر كاسياس حارس مرمى ريال وأمسك الكرة
بثبات لينتهي الشوط الأول بالتعادل السلبي . وكثف ريال من ضغطه الهجومي في الشوط الثاني بينما تراجع توتنهام للدفاع في الدقائق الأولى . ولم يتأخر ريال
مدريد في التكشير عن أنيابه في هذا الشوط حيث سجل هدف التقدم في الدقيقة ٥٠ اثر هجمة منظمة مرر منها مارسيلو الكرة إلى رونالدو ليطلقها قوية من
مسافة بعيدة ولكن الحارس البرازيلي جوميز أخطأ في التعامل معها حيث سقطت الكرة من يده وتهادت إلى داخل المرمى دون أن تفلح محاولته في اللحاق بها .

Figure (3) : the input Arabic text

IV. Experiment Result

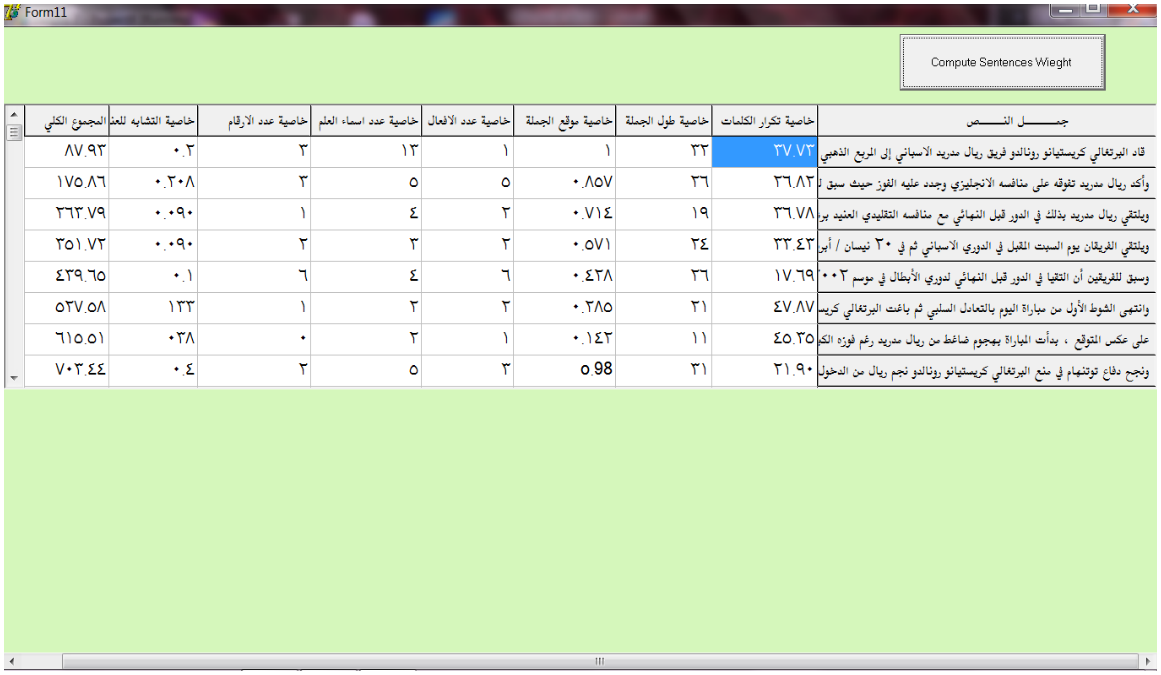
Where each document is converted into a list of sentences. Each sentence is represented as a vector $[f_1,f_2,\ldots,f_7]$, composed of 7 features.

The selection of features plays an important role in determining the type of sentences that will be selected as part of the summary and, therefore, would influence the performance of the neural network. The proposed system is implemented in Delphi 7 . In this section we presented analysis and evaluation result.

a. Analysis results

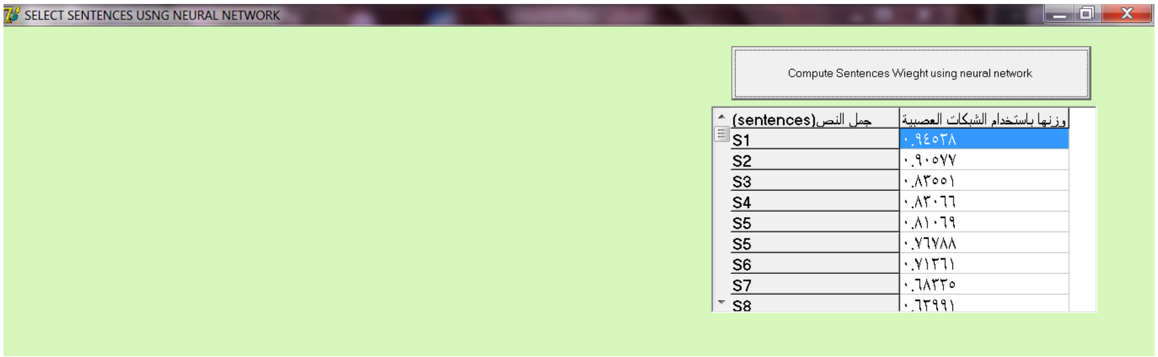
We used 150 news articles from the Internet (i.e. Wikipedia) articles with various topics such as technology, sports, and world news to train the network. Each article consists of 20 to 100 sentences. Every sentence is labeled as either a summary sentence or an unimportant sentence by a human reader. We then used the same 50 news articles as a test set for the modified network. The accuracy of the modified network ranged from 91% to 100% with an average accuracy of 94.3% when compared to the summaries of the human reader, the modified network included a sentence that was not selected by the human reader. That is, the network was able to select all sentences that were labeled as summary sentence in most of the articles.

Figure 4 show the features scores for the input text which it interred to system from (figure3) while figure 5 show feature score that choose from features scores for the text document using combines of statistical and linguistic features to compute final scores of each sentence in input text, this final scores presented to neural network to learn it and select sentences which select to generate final summary which show in figure 6.



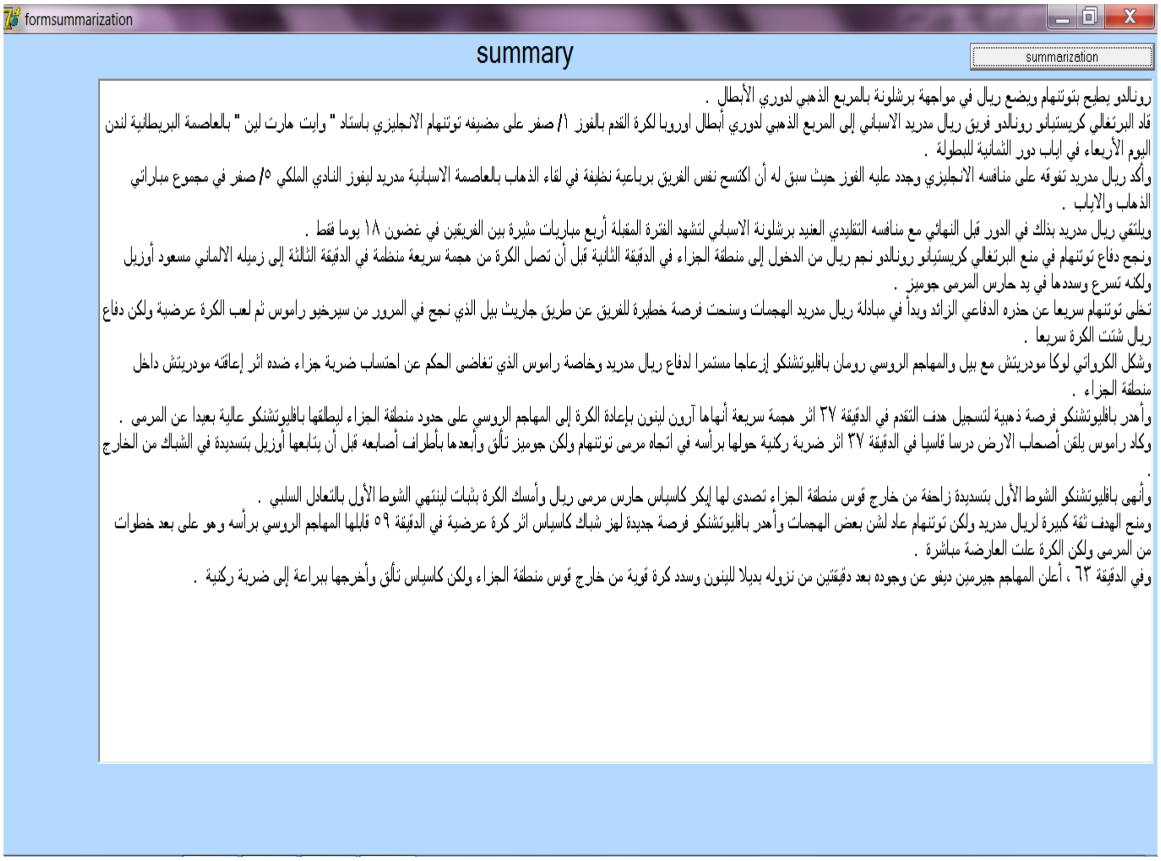
المجموع الكلي	خاصية التشابه للعنا	خاصية عدد الارقام	خاصية عدد اسماء العلم	خاصية عدد الافعال	خاصية موقع الجملة	خاصية طول الجملة	خاصية تكرار الكلمات	جمـــــــل النـــــــص
٨٧.٩٣	٠.٢	٣	١٣	١	١	٣٣	٣٧.٧٣	قاد البرتغالي كريستيانو رونالدو فريق ريال مدريد الاسباني إلى المربع الذهبي
١٧٥.٨٦	٠.٢٠٨	٣	٥	٥	٠.٨٥٧	٣٦	٣٦.٨٢	وأكد ريال مدريد تفوقه على منافسه الانجليزي وجدد عليه الفوز حيث سبق ل
٣٦٣.٧٩	٠.٠٩٠	١	٤	٢	٠.٧١٤	١٩	٣٦.٧٨	ويلتقي ريال مدريد بذلك في الدور قبل النهائي مع منافسه التقليدي العنيد بر
٣٥١.٧٢	٠.٠٩٠	٢	٣	٢	٠.٥٧١	٢٤	٣٣.٤٣	ويلتقي الفريقان يوم السبت المقبل في الدوري الاسباني ثم في ٢٠ نيسان / أبر
٤٣٩.٦٥	٠.١	٦	٤	٦	٠.٤٢٨	٣٦	١٧.٦٩	وسبق للفريقين أن التقيا في الدور قبل النهائي لدوري الأبطال في موسم ٢٠٠٢
٥٢٧.٥٨	١٣٣	١	٢	٢	٠.٢٨٥	٢١	٤٧.٨٧	وانتهى الشوط الأول من مباراة اليوم بالتعادل السلبي ثم باغت البرتغالي كريس
٦١٥.٥١	٠٣٨	٠	٢	١	٠.١٤٢	١١	٤٥.٣٥	على عكس المتوقع ، بدأت المباراة بهجوم ضاغط من ريال مدريد رغم فوزه الك
٧٠٣.٤٤	٠.٤	٢	٥	٣	0.98	٣١	٢١.٩٠	ونجح دفاع توتنهام في منع البرتغالي كريستيانو رونالدو نجم ريال من الدخول

Figure (4) : Features scores for the sentences of input Arabic text using combines of statistical and linguistic (SAL) features.


SELECT SENTENCES USNG NEURAL NETWORK
Compute Sentences Wieght using neural network

(sentences)جمل النص	وزنها باستخدام الشبكات العصبية
S1	٠.٩٤٥٢٨
S2	٠.٩٠٥٧٧
S3	٠.٨٣٥٥١
S4	٠.٨٣٠٦٦
S5	٠.٨١٠٦٩
S5	٠.٧٦٧٨٨
S6	٠.٧١٢٦١
S7	٠.٦٨٣٣٥
S8	٠.٦٣٩٩١

Figure (5) :Feature score that choose from features scores using MLP neural network for the input Arabic text



رونالدو يطيح بتوتنهام ويضع ريال في مواجهة برشلونة بالمربع الذهبي لدوري الأبطال .
قاد البرتغالي كريستيانو رونالدو فريق ريال مدريد الاسباني إلى المربع الذهبي لدوري أبطال اوروبا لكرة القدم بالفوز ١/ صفر على مضيفه توتنهام الانجليزي باستاد " وايت هارت لين " بالعاصمة البريطانية لندن
اليوم الأربعاء في اياب دور الثمانية للبطولة .
وأكد ريال مدريد تفوقه على منافسه الانجليزي وجدد عليه الفوز حيث سبق له أن اكتسح نفس الفريق برباعية نظيفة في لقاء الذهاب بالعاصمة الاسبانية مدريد ليفوز النادي الملكي ٥/ صفر في مجموع مباراتي
الذهاب والاياب .
ويلتقي ريال مدريد بذلك في الدور قبل النهائي مع منافسه التقليدي العنيد برشلونة الاسباني لتشهد الفترة المقبلة أربع مباريات مثيرة بين الفريقين في غضون ١٨ يوما فقط .
ونجح دفاع توتنهام في منع البرتغالي كريستيانو رونالدو نجم ريال من الدخول إلى منطقة الجزاء في الدقيقة الثانية قبل أن تصل الكرة من هجمة سريعة منظمة في الدقيقة الثالثة إلى زميله الالماني مسعود أوزيل
ولكنه تسرع وسددها في يد حارس المرمى جوميز .
تخلى توتنهام سريعا عن حذره الدفاعي الزائد وبدأ في مبادلة ريال مدريد الهجمات وسنحت فرصة خطيرة للفريق عن طريق جاريث بيل الذي نجح في المرور من سيرخيو راموس ثم لعب الكرة عرضية ولكن دفاع
ريال شتت الكرة سريعا .
وشكل الكرواتي لوكا مودريتش مع بيل والمهاجم الروسي رومان بافليوتشنكو إزعاجا مستمرا لدفاع ريال مدريد وخاصة راموس الذي تغاضى الحكم عن احتساب ضربة جزاء ضده اثر إعاقته مودريتش داخل
منطقة الجزاء .
وأهدر بافليوتشنكو فرصة ذهبية لتسجيل هدف التقدم في الدقيقة ٣٧ اثر هجمة سريعة أنهاها آرون لينون بإعادة الكرة إلى المهاجم الروسي على حدود منطقة الجزاء ليطلقها بافليوتشنكو عالية بعيدا عن المرمى .
وكاد راموس يلقن أصحاب الارض درسا قاسيا في الدقيقة ٣٧ اثر ضربة ركنية حولها برأسه في اتجاه مرمى توتنهام ولكن جوميز تألق وأبعدها بأطراف أصابعه قبل أن يتابعها أوزيل بتسديدة في الشباك من الخارج
.
وأنهى بافليوتشنكو الشوط الأول بتسديدة زاحفة من خارج قوس منطقة الجزاء تصدى لها إيكر كاسياس حارس مرمى ريال وأمسك الكرة بثبات لينتهي الشوط الأول بالتعادل السلبي .
ومنح الهدف ثقة كبيرة لريال مدريد ولكن توتنهام عاد لشن بعض الهجمات وأهدر بافليوتشنكو فرصة جديدة لهز شباك كاسياس اثر كرة عرضية في الدقيقة ٥٩ قابلها المهاجم الروسي برأسه وهو على بعد خطوات
من المرمى ولكن الكرة علت العارضة مباشرة .
وفي الدقيقة ٦٣ ، أعلن المهاجم جيرمين ديفو عن وجوده بعد دقيقتين من نزوله بديلا للينون وسدد كرة قوية من خارج قوس منطقة الجزاء ولكن كاسياس تألق وأخرجها ببراعة إلى ضربة ركنية .

Figure (6):- summary of input Arabic text

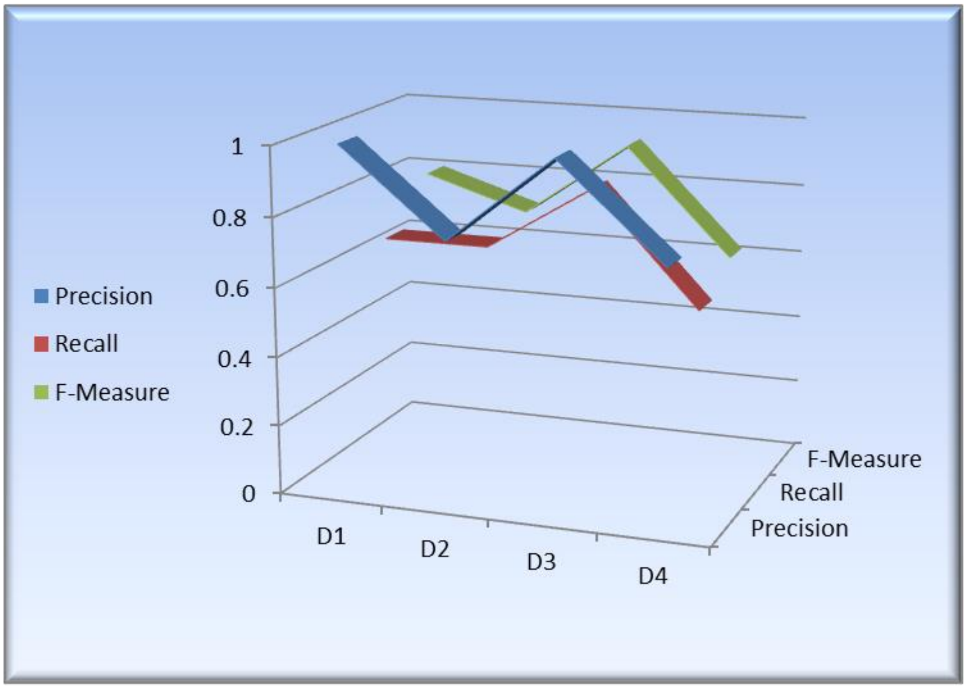
b. Evaluation results

The performance of the proposed approach is evaluated using precision, recall and F-measure from formulas (2,3,4). Precision evaluates the proportion of correct of the sentences in the summary whereas recall is used to evaluate the proportion of relevant sentences included in summary. For precision, the higher the values, the better the system is in omitting irrelevant sentences and always precision values is higher than recall values. After computes precision and recall, we compute F-measure.

The table 3 show samples of the values for the evaluation measures(precision, recall and F-Measure) with summarization rate=45. In figure 7 we shown the diagram of precision, recall and F-measure values. In the table4 ,we display two results of other researches

Table 3: the values for the evaluation measures for four document

Documents	**Precision**	**Recall**	**F-Measure**
D_1	1	0.666	0.8
D_2	0.75	0.666	0.705
D_3	1	0.857	0.923
D_4	0.727	0.533	0.615



Figure(7) :-diagram the values of precision, recall and F-measure

Table 4: the results of numbers of evaluation measures for other research

Research no.	authors	year	Performance of method
1	Khosrow Kaikhah	2004	0.91
2	Ben King et.al	2011	0.459

V. Conclusions

Our text summarization method performed well on the test article, with accuracy of 94%. The selection of features as well as the selection of summary sentences by the human reader from the training paragraphs plays an important role in the performance of the network. The network is trained according to the style of the human reader and to which sentences the human reader deems to be important in a article. Individual readers can train the neural network according to their own style. In addition, the selected features can be modified to reflect the reader’s needs and requirements. each property an important role in the selection of sentences where there is a relation between the properties of text, for example, when combining two or more properties with each other, the results were different appear. Where we found when combining property length with property similarity to the title the result was not good, as to when combining the Length property and position, they were give better results than its the previous, so we tried hard to find all the properties that, when combined we get excellent results for the bottom and we found that these features when combined get a summary of a strong and coherent as it appears in the results of the evaluation. In the future we will try to develop the integrated system by combined the proposed neural network with genetic algorithms because the role of GAs in the best choice for the sentences that will appear in the final summary.

References

Ben King et al.(2011**)** "Experiments in Automatic Text Summarization Using Deep Neural Networks", In Proceeding of the Machine Learning, pp.1-12.

D. Mallett, J. Elding and Mario, (2004) "Information -Content Based Sentence Extraction for Text Summarization", In Proceeding of the IEEE, Canada, Vol.2(2), pp.1-5.

H. P. Luhn,(1958) "The Automatic Creation of Literature Abstracts", In Proceeding of IBM Journal of Research Development, Vol.2(2), pp.159-165.

J. Kupiek, J. Pedersen and F.Chen, .(1995) "Trainable Document Summa-rizer", In Proceeding of Research and Development in Information Retrieval, pp.68-73.

J. Lin,(2009) "Summarization", In Proceeding of Springer-Verlag, pp.1-7.

J. Steinberger, (2001)"Using Latent Semantic Analysis in Text Summarization and Summary Evaluation", In proceeding of the seventh International Conference ISIM, pp.1-8.

K. Jezek and J. Stteinberger, (2008)"Automatic Text Summarization (the State of Art 2007 and New Challenges)", In Proceeding of Znalosti , pp.1-12.

Khosrow Kaikhah (2004), "Automatic Text Summarization with Neural Networks", In Proceeding on IEEE International Conference on Intelligent Systems, pp.40-44.

K. Sarkar, (2009)"Using Domain Knowledge for Text Summarization In Medical Domain", In Proceeding of ACEEE, International Journal of Recent Trends in Engineering, Vol.1, no.1, pp.200-205.

R. Mihalcea,(2004) "Language Independent Extractive Summarization ", In Proceeding of the ACL Interactive Poster and Development Sessions, pp.49-52.

R. M. ALGuliev, R. M. ALGullyev (2007)"Experimental Investigating the F-Measure as Similarity Measure for Automatic Text Summarization", In

Proceeding of the Computational Mathematical, Vol.6(2) , pp.278-287, 2007.

S.D. Afantenos, V. Karkaletsis and P.Stamatopoulos,(2005) "Summarization from Medical Documents: A Survey “, In Proceeding of *Artificial Intelligence in Medicine*, vol. 33, pp. 157-177.

VI. Gupta and G. Lehal ,(2010) "A Survey of Text Summarization Extractive Techniques “, In Proceeding of journal of emerging technologies in web intelligence, vol. 2, no. 3, pp. 258-268.

خوارزمية هجينة مقترحة باستخدام الصفات الاحصائية اللغوية SLF والشبكات العصبية لتلخيص النصوص العربية

إيمان قيس عبد الجليل أمل حميد خليل سهاد مهجر كريم

جامعة البصرة \كلية العلوم \قسم علوم الحاسبات

emankais@yahoo.com

المستخلص

مع الزيادة المستمرة لنمو المعلومات الالكترونية المتوفرة على الشبكة العنكبوتية، أصبح من الضروري والمهم للحصول على آلية لتمثيل وتقديم نص بأسلوب مختصر ومفيد، لذلك ظهرت تقنية التلخيص الآلي للنصوص والتي تلعب دورا مهما في مساعدة المستخدم لتحديد ما يحتاجه من تلك النصوص.

في هذا البحث سنقدم نبذة تاريخية مختصرة عن تلخيص النصوص واغلب الطرق ذات الصلة المستخدمة حاليا في هذا المجال. حيث تم اقتراح تقنية جديدة في مجال التلخيص الآلي للمقالات الإخبارية العربية باستخدام مفهوم الشبكات العصبية. حيث تم اختيار الجمل المهمة من النص الأصلي ووضعها في الخلاصة ، ومن خلال تدريب شبكة Multi-layer Perceptron تم استخراج الخصائص ذات العلاقة من الجمل التي ينبغي أن تدرج في الخلاصة، وبعد ذلك تم تعديل الشبكة العصبية (MLP) بالتعميم والجمع بين الخصائص ذات الصلة لتظهر في جمل الخلاصة. وأخيرا، النظام تم تقييمه بمقارنة الخلاصة النهائية للنظام مع الخلاصة التي كتبها الخبير اللغوي في مجال اللغة العربية، وقمنا بقياس انجازيه النظام وذلك بحساب مقاييس الدقة و الاسترجاع ومقياس F-Measure. أثناء التقييم حصلنا على نتائج جيدة موضحة في الاستنتاج.