English Text Summarization Using Statistical, Linguistics, Heuristics and Machine Learning Techniques

Dr. Ahmed Tariq Sadiq
Enas Tariq Khuder

Abstract

This paper presents a good multiple techniques for English text summarization. The proposed system use statistical, heuristics, linguistics and machine learning techniques to summarize the text. Statistically, the proposed system uses the classical measures in the text summarization such as words frequency, cue frequency. Heuristically, the proposed system uses the title's words, position of words…etc. Linguistically, the proposed system uses the natural languages processing tools such as part-of-speech, NP-chunk and n-grams. As a machine learning technique the proposed system uses association rules extraction to find the relational words in different documents. These four techniques executed on 20 different documents to summarize (20-40)% of original document, the proposed system have 96% as an acceptable ratio compare with reference human summary.

تلخيص النصوص الانكليزية باستخدام تقنيات احصائية, لغوية, تنقيبية, وتعلم الماكنة

د. أحمد طارق صادق
ايناس طارق خضير
قسم علوم الحاسبات
الجامعة التكنولوجية

هذا البحث يقدم تقنيات متعددة جيدة لتلخيص النصوص الانكليزية. النظام المقترح يستخدم تقنيات احصائية, لغوية, تنقيبية, وتعلم الماكنة. احصائيا النظام المقترح يستخدم مقاييس تقليدية لتلخيص النص مثل تكرار الكلمة وتكرار النموذج. تنقيبيا يستخدم النظام كلمات العنوان وتاثير موقع الكلمة.. الخ. لغويا يستخدم النظام ادوات معالجة اللغات الطبيعية مثل مقاطع الكلام, المقاطع الاسمية وعدد من المقاطع الحرفية. تم استخدام تقنية استخلاص القواعد العلائقية كتقنية لتعلم الماكنة وذلك لايجاد الكلمات ذات العلاقة ببعضها في عدة نصوص. هذه التقنيات الاربعة تم تجربتها على 20 نص مختلف وبنسبة تلخيص تراوحت بين (20-40)% من النص الاصلي, وبعد مقارنة النص الملخص من قبل النظام المقترح بالنصوص الملخصة من قبل الخبراء حصلنا على نسبة تشابه قدرها 96%.

1. Introduction

Automatic text processing is a research field that is currently extremely active; one important task in this field is automatic text summarization, which consists of reducing the size of a text while preserving its information content [1,2]. Summarization can be defined as the selection of a subset of the document sentences and which is reprehensive of its content. This is typically done by ranking the document sentences and selecting those with higher score and with a minimum overlap [1,3], in general there are two types of automatic text summarization which they are extract and subtract [1, 4].

Text Mining is also a much more complex task (than data mining) as it involves dealing with text data that are inherently unstructured and fuzzy. Text mining is a multidisciplinary field, involving information retrieval, text analysis, information extraction, clustering, categorization, visualization, database technology, machine learning, and data mining [5, 6].

Text summarization is a computer program that summarizes a text. The summarizer removes redundant information the input text and produces a shorter non-redundant output text. The output is an extract from the original text. Text summarization has largely sentence extraction techniques. These approaches have used a battery of indicators such as cue phrases, term frequency, and sentence position to choose sentences to extract and form into a summary [7, 8]. Automatic Text summarization is extremely useful in combination with a search engine on the Web. By presenting summarizes of retrieved documents to the user, Automatic Text summarization can also be used to summarize a text before it is read by an automatic speech synthesizer, thus reducing the time needed to absorb the essential parts of document. In particular, automatic Text summarization can be used to prepare information for use in small mobile devices, which may need considerable reduction of content [7, 9]. In this paper, a hybrid automatic text summarization system will be presented depends on several techniques which are statistical, linguistics (Natural Languages Processing (NLP)), heuristics and machine learning (association rules extraction) techniques.

2. Text Summarization Techniques

Most algorithms for text summarization considered are to take a document as input and automatically generate a summarized document. Below some techniques for keyword extraction:

a. Summarization procedure based on the application of ML algorithms, which employs a set of features extracted directly from the original text. A ML approach can be envisaged if we have a collection of documents and their corresponding reference extractive summaries. A summarizer can be obtained by the application of a classical ML algorithm in the collection of documents and its summaries. In this case the sentences of each document are modeled as vectors of features extracted from the text. The summarization task can be seen as a two-class classification problem, where a sentence is labeled as "correct" if it belongs to the extractive reference summary, or as "incorrect" otherwise. The summarizer is expected to "learn" the patterns which lead to the summaries, by identifying relevant feature values which are most correlated with the classes "correct" or "incorrect". When a new document is given to the system, the "learned" patterns used to classify each sentence of that document into either a "correct" or "incorrect" sentence, producing an extractive summary [11, 12].

    1- **Association Rules :** Association rule mining is one of the most popular techniques in data mining. The problem of mining association rules is to discover all association rules that have support and confidence greater than the user-specified minimum support and minimum confidence [7, 13]. The association task for data mining is the job of finding which attributes "go together." Most prevalent in the

business world, where it is kwon as affinity analysis or market basket analysis, the task of association seeks to uncover rules for quantifying the relationship between two or more attributes. Examples of association tasks in business and research include [6, 8]:

- Investigating the proportion of subscribers to a company's cell phone plan that respond positively to an offer a service upgrade.
- Examining the proportion of children whose parents read to them who are themselves good readers.
- Predicting degradation in telecommunications networks.
- Finding out which items in a supermarket are purchased together and which items are never purchased together.
- Determining the proportion of cases in which a new drug will exhibit dangerous side effects.

2- **Artificial Neural Networks** : Methods based on similarity measures do have intrinsic limitations: they rely on simple predefined features and measures, they are developed for generic documents, their adaptation to a specific corpus or to different document genres has to be manually settled. Machine learning allows to better exploit the corpus characteristics and to improve context since document types may vary considerably. In [22] a technique, which takes into account the coherence of the whole set of relevant sentences for the summaries and allows to significantly increasing the quality of extracted sentences. Firstly, define features in order to train our system for sentence classification. A sentence is considered as a sequence of term, each of them being characterized by a set of features. The sentence representation will then be the corresponding sequence of these features. The proposed system in [22] used four values for characterizing each w of sentence s: tf(w,s), tf(w,q), (1-(log(df(w)+1)llog(n+1)) and Sim(q,s)-computed as in (2) – the similarity between q and s. The first three variables are frequency statistics which give the importance of a term for characterizing respectively the sentence, the query and the document. The last one gives the importance of the sentence containing w for the summary and is used in place of the term importance since it is difficult to provide a meaningful measure for isolated terms [25]. A first labeling of the sentences as relevant or irrelevant is provided by the baseline system. By tuning a threshold over the similarity measures of sentences for a given document, sentences having higher similarity measures than this threshold were set to be relevant. Then use self- supervised learning to train a classifier upon the sentence labels provided by the previous classifier and repeat the process until no change occurs in the labels. As a classifier, two linear classifiers have been used, a one layer with a sigmoid activation function [23] and a Support Vectore Machine (SVM) [24], to compute P(R0/s), the posterior probability of relevance for the query given a sentence, using these training sets. At last, uses the same word representation as in the case of self-supervised learning. The system [22] has labeled 10% of the sentences in the training set using the news-wire summaries as the correct set of sentences. Then train our classifiers in a first step using these labels. Training proceeds after that in the same way as for the self supervised case: this first classifier is used to label all the sentences from the training set, these labels are used for the next step using unlabeled data and so on until converge.

b. Summarization procedure is based on the application of **NLP** is the process used by a computer to understand and produce a language that is understood by humans. In this way people can communicate with machines as communicating with other humans. The NLP has some methods such as [14]:

1- Principles of P-O-S Tagging

There are two main approaches to part-of-speech tagging: rule-based and probabilistic. The tagger presented in this document belongs to the purely probabilistic ones. That means that for disambiguating tags within a text it only uses probabilities, and no rule-based mechanism. The first step in any tagging process is to look up the token to be tagged in a dictionary. If the token cannot be found, the tagger has to have some fallback mechanism, such a morphological component or some heuristic methods. This is where the two approaches differ: while the rule-based approach tries to apply some linguistic knowledge, a probabilistic tagger determines which of the possible sequences is more probable, using a language model that is based on the frequencies of transitions between different tags [15, 16].

2- N-Gram

An N-gram is an N-character (or N-word) slice of a longer string although in the literature the term can include the notion of any co-occurring set of characters in a string. Typically, one slices the string into a set of overlapping N-grams. We also append blanks to the beginning and ending of the string in order to help with matching beginning-of-word and ending-of-word situations. Thus, the word "TEXT" would be composed of the following N-grams:

bi-grams: _T, TE, EX, XT, T_

tri-grams: _TE, TEX, EXT, XT_, T_ _

quad-grams: _TEX, TEXT, EXT_, XT_ _, T_ _ _

In general, a string of length k, padded with blanks, will have k+1 bi-grams, k+1 quad-grams, and so on. N-gram-based matching has had some success in dealing with noisy ASCII input in other problem domains, such as in interpreting postal addresses, in text retrieval, and in a wide variety of other natural language processing applications. The key benefit that N-gram-based matching provides is derived from its very nature: since every string is decomposed into small parts, any errors that are present tend to affect only a limited number of those parts, leaving the remainder intact [14].

3- NP-Chunk

By chunk we mean breaking the text up into small pieces. When manually assigned keyword is inspected, the vast majority turn out to be noun or noun phrases. In our experiments a partial parser was used to select all NP-chunks from text. This approach is one of linguistic approaches that is used in the proposed system which depend on the following grammar to extract each phrase matching it

< det > , < noun >

< det > , < noun > , < noun >

After the parser is complete the extraction of noun phrase is stored in the database, the system start a new step that filter the candidate keyphrase after stemming them and start to measure the features that will be mentioned later to get the final weight of the current phrase [14].

c. Summarization procedure is based on the statistical information obtained from a document corpus drawn as an important item for constructing summarization procedure. Such systems consist of two parts: the training part and the summarization part. There are some methods in this model such as [10]. The features are based on the frequency of some elements in the text; which give different kinds of information about relevance of sentences to the summary. These features are sufficiently relevant to the single document summarization task [1, 17]. There is some methods in this model as below.

1- **Cue Word Probability:** The most frequent words or phrase in the manually tagged summary sentences are defined as cue words. An occurrence of a cue word in a sentence assumed to indicate that the sentence is likely to form a good summary since it is found in many summary sentences selected by human judges. Phrases like "the purpose of this article" and "suggests the procedure for" signal that a sentence containing them is them bearing ones [7, 18].
2- **Position Feature:** Sentences in a document are distinguished according to whether they are in the initial (for sentences within the first five), final (for sentences within the first five), or middle part. Since we consider only introduction and conclusion sections of a document for summarization. There are six different position values that can be assigned to individual sentences. Based on our observation, sentences in the final part of an introduction section or in the first part of conclusion section are more likely to be included in a summary than those in other parts of the sections.
3- **Theme Words:** Content-bearing words or key words have played an important role in information retrieval since they can represent the content of a document reasonably well, since it is intuitively appealing to consider only those sentences with strong keywords. The system used this feature as evidence for a good summary sentence. The more important keywords are included in a sentence, the higher score it gets.
4- Resemblance to the Title: This feature is concerned with how similar a sentence is to the title of the source document, since a document title is usually a good representation of what is in the document. It is in a sense a summary [19].

Some of those features are listed below:

- **N-Gram frequency:** This feature calculates for every n-gram a specific value based on its frequency. An n-gram is a sequence of n contiguous characters including blanks but excluding punctuation marks [1, 20].
- **Word Frequency (WF):** Open class words (content words) which are frequent in the text are more important than the less frequent, equation (2-1) shows the WF of word (w) [1], [2].

$$WF(w) = \sum_{i=1}^{w \epsilon d} occure\ of\ (w) \dots \dots (1)$$

Where WF (w) is the number of the times a word (w) appears in a document d.

d. Summarization procedure is based on the application of Heuristic It is based on sentence length or position and some other features as follows [1, 2, 21]:

1- **Position Score:** The assumption is that certain genres put important sentences in fixed positions. For example, newspaper articles have the most important terms in the first four paragraphs. Line position is less important in reports than the newspaper text. In newspaper text, the most important part of the text is the first line followed by other lines

in a descending order. The equation (2) is used for the calculation of the position score for newspaper texts.

Position Score = (1\line number)*10 …….. (2)

2- **Title:** Words that appears in the title are important and get high score; the following equation is used for the calculation of the title frequency for word (w).

TF (w) = 1 if w occurs in the title of the document, otherwise 0 …….. (3)

Where TF (w) is the number of times w occurs in the title

3- **Indicative Phrases:** Sentences containing key phrases like "this report …".

4- **Proper word:** Proper names, such as the names of persons and places, are often central in news reports and sentences containing them are scored higher.

5- **Quotation:** Sentences containing quotations might be important for certain questions from the user.

3. The Proposed System

Each text summarization technique have advantages and disadvantages, therefore, in our proposed system we attempt to collect the most important techniques to obtain the high advantages and low disadvantages as soon as possible. The proposed system has several step techniques, the following points illustrate the headlines of these techniques:

a. Preprocessing.
b. Words Processing (Stemming, Dictionary, Drop Stop Words).
c. Heuristic Techniques (Title, Position, Indicative Phrase…etc).
d. Natural Language Processing (NP-Chunk, PoS …etc).
e. Statistical Techniques (Word Frequency, Chunk Frequency…etc).
f. Learning Technique (Association Rules Extraction).

Figure (1) illustrates the block diagram of the proposed system.

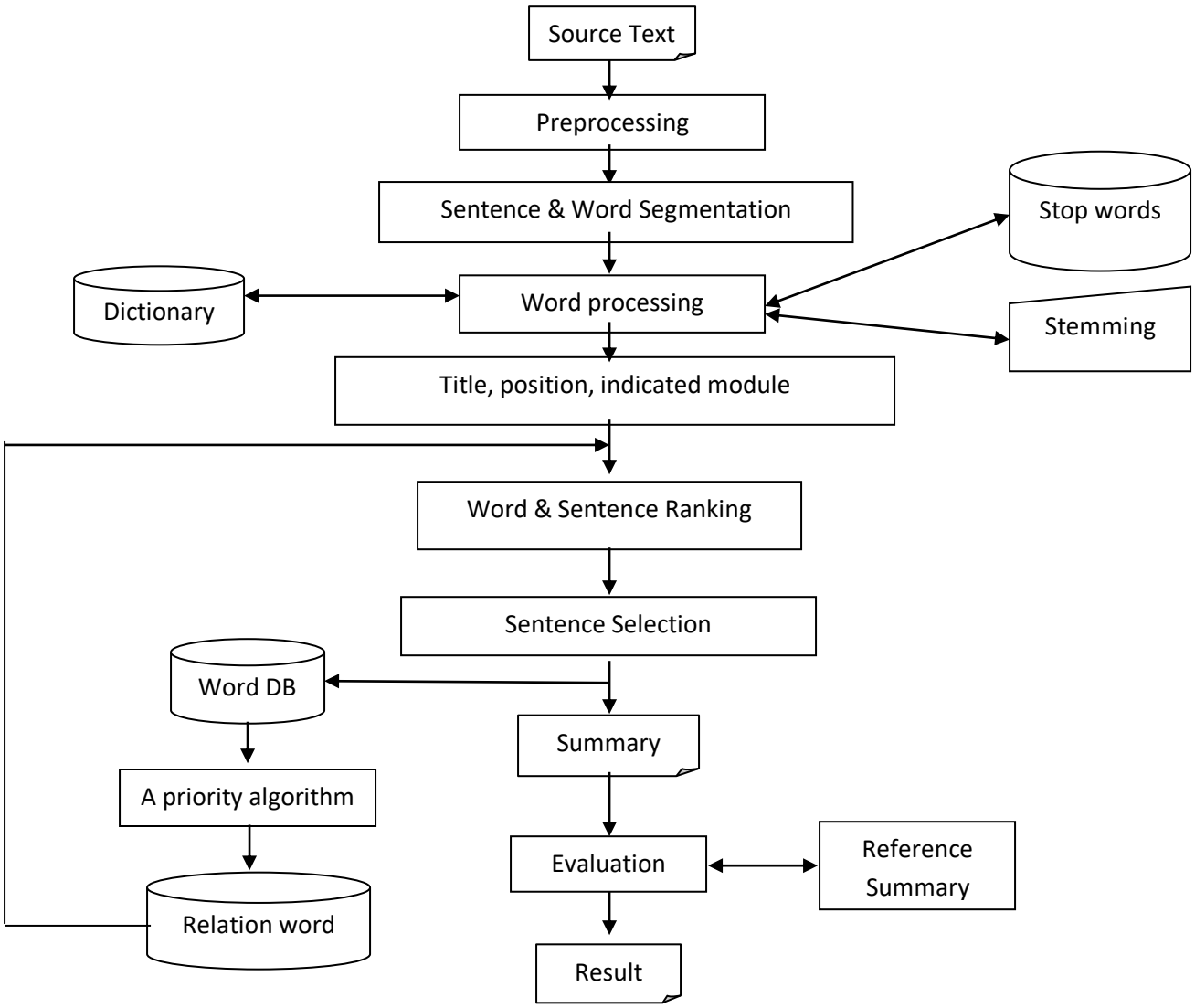

Fig (1) Block diagram for proposed system

3.1 Preprocessing operations

The source text enters to the system some preprocesses must be performed on it; as shown in the following algorithms.

1. Abbreviation

As the task of the tokenizer is to detect sentence, and word boundaries in a written text and to provide a uniform segmentation before the processing of the text takes place. In English text, word boundaries can be delimited by space, punctuation, digits, new lines, and some special characters, and sentence boundary can be delimited by dots, equation mark, and surprising mark . Once more, any stop appearing on an abbreviation is ambiguous with a full-stop and can thus mark the end of a sentence. So any of the character abbreviation patterns are ambiguous if followed by a dot; as might also indicate the end of the sentence. In these cases, the system should specify or identify the abbreviations before sentences and words identifications.

2. Ignore data in parenthesis

The data between parentheses don't give important themes as it is give or display the same themes of the text or phrase before it. The proposed system ignores the text that appears between parentheses.

3. Ignore numbers

The whole numbers in the English language is expressed by using a dot. When the tokenizer segments the sentence and words of the text, so the dot in whole numbers is ambiguous as it is may mark the end of a sentence, and the numbers don't give any theme to the sentence, and as it make ambiguity the process of the tokenizing. So the proposed system ignores the number if it appears in the text before tokenizing process taken place.

3.2 Sentence & Word Segmentation

The sentences must be identified; the characters that mark the end of the sentence must be specified as it hard for the software program to specify them. The proposed system used four special characters ('.','?','!', and new line) as a sentence boundary, it reads the text if the new character is like any one of them means we have new sentence, so the proposed system identified it as a sentence and gives it a unique ID for the purpose of the final accessing.

The word segmentation is the important step in the proposed system, after the sentence segmentation the summarizer will segment and specify each words to be prepare for operations. In English text, word boundaries can be delimited by space, punctuation, digits, new lines, and some special characters. The system must specifying the word boundary at first, depends on them will segment the text to the words. The summarizer get each sentence, segment them to the words, each word will be checked if all it's characters is capital letters it will be ignored as it means it is the abbreviation or it's not the meaning words in the English language dictionary.

3.3 Word processing

1- Ignore stop words

Stopword referred to counter the obvious fact that many of the words contained in a document do not contribute particularly to the description of the documents content; they are very frequent and non-relevant words. For instance, words like "the", "is" and "and" contribute very little to this description. Therefore it is common to remove these so-called stopwords prior to the construction of the document descriptions, leaving only the content bearing words in the text during processing.

2- Stemming

Removing suffixes by automatic means is an operation which is especially useful in the field of information retrieval. Terms with a common stem will usually have similar meanings, for example:

CONNECT – CONNECTED – CONNECTING - CONNECTION

Frequently, the performance of an IR system will be improved if term groups such as this are conflated into a single term. This may be done by removal of the various suffixes -ED, -ING, -ION to leave the single term **CONNECT**. The proposed system uses a porter algorithm, and this section shows how the system performs stemming. It is removes about 60 different suffixes, which involves a multi-step process that successively removes short suffixes, rather than removing in a single step the longest suffix.

3.4 Some Heuristic Module

In this step the system concerned with the title and position of token. Its means that certain token positions tend to carry important topics to yield good summary sentences. The proposed system takes in account three approaches of this module:

- the first approach deals with the words in the title.

-the second approach is concerned with partitioned original English texts into paragraphs and found the number of paragraphs that the token occurrence in it. The system points to this way by NP (Number of Paragraph).

-the third approach is concerned with the first paragraph in the text because of in most texts the first paragraph may be containing the key meaning in it. So the proposed system matching the token with first paragraph, and if it found it compute the number of first paragraph's sentences was the token occurrence in it. The system points to this way by FP (First Paragraph).

3.5 Word and Sentence Ranking

In this section the score or weight of all words and sentences in the document by three features are finding, to use them for sentences weighting, as follows

Feature one: WF/ID of each word w

Statistically find the weight of words by counts the appearance of them in the whole document.

$$\mathbf{WF(w)= \sum_i^{w \epsilon d} wi \quad \ldots\ldots\ldots\ldots\ldots (4)}$$

w is the word, **d** is document

For each word (**w**) in the document count all appearance of it in the document.

Feature two: Similarity with the title of document of each word w, Heuristically for each word in document check if it is appears in the document title or not.

TF(w)= 1 if w appears in the title, otherwise 0

Feature three: WS of each word w**,** Statistically for word (w) find the weight of it by counts the number of sentences it will appears in it.

$$\mathbf{S(w)=\sum_i^{s \epsilon d} w \epsilon s_i \ldots\ldots\ldots\ldots\ldots\ldots\ldots (5)}$$

S(w) is the number of sentences in which **w** occurred, **s** is sentence, **d** is document.

For each word (w) in the document count the number of sentences that it is appears in it.

The sentences need to be scored, by using weights of its words. It depends on the methods that are to be used in the system, which uses the specified features far each using methods. The proposed system uses two methods which are statistics and heuristics, for each of them used the three specified features as follows:

1- Statistics: Using simple statistics, to determine the summary sentence of a text by using the words weight; and three statistics features:

Feature One: Weight Sentence WS

$$\mathbf{WS(s)= \sum_i^{w \epsilon s} w_i \ldots\ldots\ldots\ldots\ldots\ldots (6)}$$

For each sentence (**s**) in the document count all its words.

Feature Two: Sentence Frequency SF of each sentence S

$$\mathbf{SF(s)=\sum_i^{w \epsilon S} WF(w) \ldots\ldots\ldots\ldots\ldots\ldots (7)}$$

Where **S** is the sentence, **WF(w)** is the frequency of word **w** that accurse in the sentence **S.**

Feature Three: TF/ISF the sentence score is the summation of its words scores or weights. The score of each word w is given by the following formula:

$$\mathbf{TF/ISF=F(W)*(\log n/ S(W)) \ldots\ldots\ldots\ldots\ldots\ldots (8)}$$

Where **F(w)** is the frequency of **w** in the sentence, **n** is the number of words in the sentence and **S(w)** is the number of sentences in which **w** occurred.

2- Heuristics: Using heuristic features, to determine the summary sentence of a text, the system used also three features which are:

Feature first: Similarity with the title, specify if the words in the sentence occurs in the document's title or not. Give higher degree to sentence's that have more words similarity with words in the document's title.

$$\mathbf{TF(s)= \sum_i^{w \epsilon s} TF(wi) \ldots\ldots\ldots\ldots\ldots\ldots (9)}$$

TF(s) is the number of time w occurs in the title

Feature two: Position of the sentences, the assumption is that certain genres put important sentences in fixed positions; the proposed system assumes that the important sentences are those that are in the first three sentences in document.

PF(s)= 1 if s is one of the first three sentences, otherwise 0

Feature third: Length of the sentences, the score assigned to a sentence reflects the length of the sentence as to prevent long sentence from getting higher score, and prevent ignoring the small sentences. The proposed system assigns score of sentence depending on the length of it, to normalized length by the longest sentence in the document, as appears in this formula:

Word-count = number of words in the text.
Worde-count = nnumber of wordes in the text.
Average sentence length (ASL) = Word-count / Worde-count
W_{sl} = (ASL * Sentence Score)/ (number of words in the current sentence

3.6 Sentence selection

The proposed system uses statistics and heuristic methods for finding summary, after finding score of each sentence by the specified methods and features the combination function is required to ranking sentences with different weights for giving them the final sentence score, which uses simple combination function for this purpose. The sentence score is calculated then, according to the following formula:

Sentence score = Σ CjPj (Coefficient, Parameter, j = 1...n, n = nr of parameters)…….(10)

The summarization ratio and final sentence scores are used to select the summary sentences from the candidate sentences.

3.7 Extract Relational Words Using Association Rules Mining

After the system summarized several texts it can builds a database of relational words to used it in computing the weight of token, adding information about word relations could significantly increase summarize quality. This system used for integrating text summarization and data mining, text summarization benefits DM by extracting structured data from textual documents, which can then be mined using traditional method. The predictive relationship between different words discovered by data mining and association rules can provide additional clues about what words should be extracted from a document and added it to the lists of relational words.

The process of building a relational words database is as follows the system gets high weight words and start data mining steps for finding frequent item set in the database of the words that are extracted previously from different text documents. The system stores all extracted words from each document in a database that will be mined latter. Through the growth of the database by learning its become more and more accurate and reliable. After the building of relational words the system computes the frequency of each token in this database RW (i), this frequency used in compute the weight of each token in the input text to generate summarize depends on relational words that increase summarize quality.

3.7 Evaluation and Results

Evaluation issues of summarization systems have been the object of several attempts; the proposed system considers classical measures that they are the Precision, Recall, RC, and RR ratio:

1- With a summary S and a text T:

CR = (length S) / (length T) …….(11)

RR = (info in S) / (info in T)……....(12)

A good summary is one in which CR is small (tending to zero) while RR is large (tending to unity). Letters number used as a metric for measuring length

2- Reference summary: The proposed system gets reference summary from the user, evaluates between it and the system summary using precision, recall measures for finding summary performance**.**

Precision= number of sentences between system summary and reference summary/ total number of sentences in system summary

Recall= number of sentences between system summary and reference summary/ total number of sentences in reference Summary

A good summary is one in which precision and recall is large (tending to unity). The evaluation is performed on twenty documents with different sizes; evaluates each of them by statistical, linguistics, heuristics and machine learning methods with three different summarization ratios (%20, %30, and %40 of the sentence number in the source document) to get the RR, and CR ratios.

3.8 Summary Generation

The final step in the proposed system is the generation of a summary. This step extracts the sentences which included the high weight tokens, combines it and deletes the repeated sentences then generate text. In the tow modules this step then returns the top-ranked (25-40) % of sentences in the original text as its final result.

4. Experimental Results

The experimental results have obtain from twenty documents using four text summarization techniques (statistical, linguistics, heuristic and machine learning) with different mixture of these techniques with three sizes of text summarization ratio for 20 different documents with size from 1 KB to 30 KB. Table (1) shows the acceptable ratio of output text compare with reference text.

Table (1) : Acceptable Ratio for 20 Different Documents

Text Summarization Techniques	**Acceptable Ratio with Three Different Size Ratio**		
	20%	**30%**	**40%**
Statistical & NLP	80%	83%	91%
Statistical, NLP & Heuristics	81%	85%	95%
Statistical, NLP & ML	81%	84%	94%
Statistical, NLP, Heuristics & ML	82%	88%	96%

5. Conclusions

The proposed system produces a hybrid model for text summarization that differs from the previous studies, it uses statistics, linguistic, heuristics and machine learning methods, each one plays a role in solving a particular sub-problem, by combining these approaches we reach an integrated system of text summarization. Several conclusions are drawn after the proposed system implemented they are as follows:

1. Statistical and linguistic methods are very important methods, therefore any text summarization tool must be used these two techniques.
2. Using of the heuristics techniques added important features to summarized text, therefore the acceptable ratio increase with this features.
3. Applying the association rules application through building a database of relational words using A priori algorithm improves system accuracy by finding the weight of each token depending on the relation between it and all other tokens in the text. Therefore, the proposed system doing two tasks at same time the first task is to build the database of relational words by learned it from the several input texts, and the second task is to generate a summary text.

4. The extracted text from the different 20 original text saved the meaning of the original text with non-redundant features. The sentences of output text of the proposed system are (20-40) % from the sentences of the original text. From the implementation and the proposed system experiments a good results obtained about 96% for text summarization compared with text summarized manually by human with used all techniques (statistical, linguistics, heuristics and machine learning).

References

1. Ah-Hwee Tan, "**Text Mining: the State of The Art and The Challenges**" ,In Processings, PAKDD'99 Workshop on Knowledge discovery from Advance Databases(KDAD'99), Beijing, pp. 71-76,April (1999).
2. Anette Hulth, Karlgren, J., Jonsson, A., Bostrom, H. & Asker, L. "**Automatic Keyword Extraction Using Domain Knowledge**", In Proceedings of the Second International Conference on Computational Linguistics and Intelligent Text Processing, Springer (2002).
3. Bishp C., "**Neural Networks for Pattern Recognition**", Oxford University Press, 1995.
4. Burges C., "**A Tutorial on Support Vector Machines for Pattern Recognition**", Bell Lab. Press, 1998.
5. Daniel T. Larose, "**Discovering Knowledge in Data an Introduction to Data Mining**", Canada (2005).
6. David Hand, Heikki Mannila, Padhraic Smyth, "Principles of Data Mining", Mit Pr, USA, (2001).
7. Hercules Dalianis, Martin Hassel, Jagen Wedekind, Dorte Haltrup, Koenraad de Smedt, and Till Christoper Lech, "**From SweSum to ScandSum-Automatic text Summrization for the Scaninavian Languages**", In Holmboe, H. (ed.) Nordisk Sprogteknologi, Norway, 2002.
   http://ccl. Pku.edu.cn/doubtfire/NLP/Lexical_Analysis/W
8. Huaizhong KOU and Georges Gardarin, **"Keyword Extraction Document Similarity &Categorization"**, University of Versailles, France (2002).
9. Hussein Keitan, "Design and Implementation of embedded association Rules Miner", Ph.D. thesis, University of technology, (2002).
10. Ian H. Witten, "**Adaptive Text Mining: Inferring Structure from Sequences"**, J. Of Discrete Algorithms, Vo1. 0 No. 0, pp. 1-23, Hermes Science Publications (2004).
11. Kanar Shukr Muhamad, "**English Text Summarization Using Hybrid Methods**",M.Sc. Thesis.Unifersity of Salahaddin -Hawler, 2008.
12. Kanus D. & et al, "**Highlighting Relevant Passages for Users of the Interactive SPIDER Retrieval System**", TREC-4 Proceeding, 1994.
13. Martine Hassel, "Resource Lean and Portable Automatic Text Summarization", ph.D Thesis, Stockholm, Sweden, 2007.
14. Massih-Reza Amini & Patrick Gallinari, "**Self-Supervised Learning for Automatic Text Summarization by Text-span Extraction**", 23rd BSC European Annual Colloquium on Information Retrieval, 2001.
15. Massih-Reza Amini, "Interactive Learning for Text Summarization", proceedings of PKDD'2000/MLTIA'2000 Workshop on Machine Learning and Textual Information Acess,pp. 44-52, Lyon, france,2000.
16. Massih-Reza Amini, Nicolas Usunier, and Patrick Gallinari,"**Automatic Text Summarization Based on Word-Clusters and Ranking Algorithms"**, ECIR, LNCS 340, Springer- Verlag Berlin Heidelberg, USA, 2005.
17. Nima Mazdak, "**A Persian text summarizer**", Master Thesis, Department of Linguistic, Stockholm University, Sweden, January 2004.

18. Noor Amjed Hassan Alzuhairy, "**Multi Method for Text Summarization Based on Learning Technique**". M.Sc. Thesis. University of Technology, 2008.
19. Oliver Mason, "Principles of P-O-S Tagging" , (1998), Ord Segmentation Tagging/QTAG/Principles%20of%20 P-O-S%20Tagging.htm
20. Rafiq Abdul-Rahman, "**Automatic Keywords Extraction Using Combined Methods**", Ph.D. thesis. University of Technology, (2006).
21. Raymond J. Monney and Un Young Nahm, "**Text Mining with Informatiom Extraction**", Department of Computer Sciences, University of Texas, (2007).
22. Rene Schneider, "**n-gram of Seeds: A Hybrid System for Corpus-Based Text Summrization**", In Proceedings of LREC (Third International Conference on Language Resources and Evaluation), Las Plams de Gran Canaria, Spain, 19-31, 2002.
23. Salton G. and McGill M., "**Introduction to Modern Information Retrival**", New Yourk McGraw-Hill Book C, New Yourk, (1993).
24. Sholom M. Weiss and others, "**Text Mining Predictive Methods for Analyzing Unstructured Information**", Sipringer (2005).
25. Sung H. Myaeng and Dong H.Jang, "**Development and Evaluation of a Statistically-Based Document Summarization System**", Department of of Computer Science Chungnam National University, Korea, (1998).