Enhanced Feature-Based Automatic Text Summarization SystemUsingSupervised Technique

Madhi A. Ali[*], Ali A. Al-Dahoud[*], Bilal H.Hawashin[*]

*Faculty of Science and Information technology,Al-Zaytoonah University of Jordan

ABSTRACT

In this work, we propose an efficient text summarization methodby ranking sentences according to their scores that use a combination of existing and improved sentence features. Many works in the literature proposed improvements to text summarization but this field still needs more improvement. For this purpose, we propose improvements to Sentence position, Sentence length, and Key wordsentence features. Afterwards, we find the optimal combination between these features and some existing features such as Term frequency, Sentence centrality, Title similarity, and Upper case of word. By usingmachine learning techniques, mainly SVM, Naive Bayes and Decision Tree classifiersour paper evaluates two feature groups: a combination of seven features without any improvements,and the same seven features after making some improvements onSentence position, Sentence length, and Key word sentence features to enhance the performance of text summarization system.Experimental results showed that making enhancements on some features improved the accuracy.

Indexing terms/Keywords

Text Summarization; Classifiers; Sentence Features; Machine Learning

Academic Discipline and Sub-Disciplines

Information science, Data mining

SUBJECT CLASSIFICATION

Computer science

TYPE (METHOD/APPROACH)

Text summarization

INTRODUCTION

Nowadays, the size of videos, audio, images, anddocuments in Internet is increasing quickly In addition to the increase in number of Web users. The volumes of topics and information that are available today in the World Wide Web (WWW) become too huge.According to[23] the goal of automatic summarization is "to take an information source, extract content from it, and present the most important content to the user in a condensed form and in a manner sensitive to the user's or application's needs". Obviously, the objective of text summarization is to present the maximum significant information in a smaller text while keeping its main content. [17] indicated that summarized text can be defined as a text that is produced from one or more texts that contains a significant portion of the information in the original text(s), and that is no longer than half of the original text(s).

The large amount of topics and information available today in the internet has become tremendous .The problem facing us now is to find the relevant documents because there is no time to read everything. In addition to the huge availability of documents in the internet, it is very difficult for human beings to manually summarize large number of texts. Therefore, the need has become urgent to get the summaries of this information in less time and effort. To solve this problem, many Summarization technologies are used to find the relevant documents. Despite the development in this field, the performance needs more enhancements in some scenarios. Besides, the summary is not always relevant to user needs and backgrounds.

In our study, we propose a generic text summarization method that creates summaries of English texts by ranking and extracting valuable sentences from the original texts. This method usessome improvements to three feathers: Sentence position, Sentence length, and Key word features. Afterwards we find the optimal combination between these features and some existing features: Term frequency, Sentence centrality, Title similarity, and Upper case of word. By using machine learning techniques with SVM, Naive Bayes and Decision Tree classifiers, our paper evaluate two features groups: the first one consists a combination of the previously mentioned features without any improvements, the second group consists of the same seven features after making some improvements on Sentence position, Sentence length, and Key word sentence features to enhance the performance of text summarization system. Recall, precision, and f-measure evaluations are used to evaluate the performance. The study uses two datasets:The first contains 100 newspapers article,and the second collection contains 100 articles in various domains created by Joint Research Centre (JRC).

The contribution of this work is as follows:

-Study different techniques, methods and approaches of text summarization.

- Try to optimize some sentence features for extraction approach by using machine learning method.

The rest of the paper is as follows. Section two presentsrelated works, section three presents the proposed system,section four present experimental results and discussion, finally section five presents conclusion and future works.

RELATED WORK

Since 1950s, several well-known Text summarization algorithms have been developed and improved. Earliest instances of research on summarizing scientific documents proposed paradigms for extracting salient sentences from text using features like: Word and phrase frequency [22], Position in the text [6], syntactic analysis for machine indexing and abstracting five different word frequency and distribution [7], four features involve Key Words and sentence position [10][25].The 1980s enjoyed an explosion of a variety of different approaches based on artificial intelligence such as scientific researches of [21][9][13] and [30].In same time, Study of [19] proposed an automatic summarization method combining conventional sentence extraction and trainable classifier based on Support Vector Machine (SVM). The evaluation results showed that system achieved the best result among all the other systems with regard to contents, and closer to the human constructed summaries .but the system needs to improve readability of its summary output.

On the other hand [18] proposed two approaches: a trainable summarizer called a modified corpus-based approach (MCBA) and Latent Semantic Analysis (LSA) based Text Relationship Map (T.R.M) approach (LSA + T.R.M) to address text summarization. The Performance evaluation (F -measure) compared between CBA and MCBA when considering different heuristic functions showed: the result of CBA about 0.304, 0.412, and 0.468when CR is 10%, 20%, and30% respectively. In contrast to, the F -measure of MCBA about 0.302, 0.413and 0.483when CR is 10%20% and30% respectively.Furthermore, [15] proposed a text summarization method based on Naive Bayes algorithm. The system experimented with 320 Vietnamese texts by built a Vietnamese text corpus for summary purposes. The problem of this study that is word segmentation tools is not high accuracy in single syllable languages as Chinese, Japanese, Vietnamese, and Thai. Therefore, the study tried to enhance quality of text summary by reduce time for computing in single syllable language and enhance accuracy by chosen an approach based on supervised learning method using Naive Bayes. Study used three key features for calculating weight of sentences as follow: Information Significant, Amount of information in a sentence and Position of sentence.

In the research [4] they are proposed a text summarization system that is used a combination of Genetic Algorithms (GA) and Genetic Programming (GP) to find the optimal rule-based system and functions of fuzzy systems. Therefore, this research used six nonstructural features such as: number of title words in the sentence, first and last sentence in the paragraph, and the number of words in the sentence. Data set consist 3 news articles with various topicsMoreover; paper [4]proposeda system for generating summary by sentence extraction. It is included 10 textsby used Naive Bayes classifier with 10%, 20%, and 30% Compression rates.The result showed the best result is given in the F-measure score with 30% of compression rate.According to [11]An automatic text summarization was proposed to make a comparison between the results of classifying full documents for feature selection such as term frequency with summarized same document. Sakhr summarizer is used in this system for Arabic texts to find the important sentences that is most relevant to the topic of text. At the same time, term frequency for feature selection and same documents pass a text summarizer are classified using SVM classifiers, by using Weakaito Environment for Knowledge Acquisition (WEKA) where SVM is already implemented in Java.In below, Table.1 shows some text summarization systems comparison includes: name, year, and other features. The details of Automatic Text Summarization we explained in next section.

Table1: Comparative between various summarization systems

System [Ref] Year	Source Inputs	Domain	Summary Output	Features
ADAM [**25**]1975	Single document	Chemistry	Abstract	-Cue phrases and term frequencies - sentence selection and rejection
SUMMARIST [**17**] 1997	Single document	News	Extract	-stages for summarization are divided in: interpretation and topic identification. - it is a multi-lingual system
MultiGen [**5**]1999	Multi document	News	Abstracts	-it identifies and synthesizes similar elements across related text from a set of multiple documents -it is based on information fusion and reformulation.
NTT [**16**] 2002	Single document	Generic	Extracts	-it employs the(SVM) machine learning Technique. -it also uses the following features: position, length, weight, similarity with the headline.

CLASSY **[8]**2005	Multi document	Query	Extracts	- it is a query-based system - it is based on Hidden Markov Model algorithm
NetSum [**33**] 2007	Single document	News	Extracts	-it is based on machine learning techniques . -it uses a neuronal network algorithm to enhance sentence features.
NGD [**27**] 2009	Multi- document	Generic	Extracts	- set of sentences are clustered into non overlapping groups of clusters. - Word stemming(Porter's) was used
NGD [**28**] 2012	Multi- document	Generic	Extracts	- Approach is by optimizing two objectives: Content coverage & Redundancy. - applies extraction method.
POS [**34**] 2013	Multi- document	Generic	Extracts & Abstracts	- Hybrid association rule mining method to identify implicitly. - Candidate basic rules are reasonable and very common; furthermore, the frequency method can achieve the best performance.
ExB [**32**] 2015	Single & multi-document	Generic	Extraction	- used graph theory and system supports at least the 38 languages -apply different domains and tasks.

THE PROPOSED SYSTEM

In this section, we proposed a method that combines both modified and existing sentence features to improve the text summarization accuracy. Thesystem in figure 1 is divided into three phases:A is preprocessingphase, B isFeature Extraction and Mechanism learning phase, and C is post processing phase.

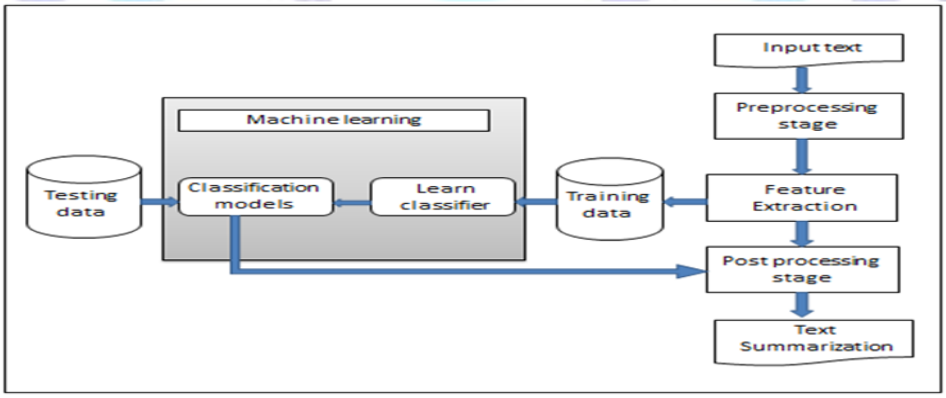

Figure1: System architecture

According figure 1, our proposed system is composed of three stages:

Preprocessing

In this stage, our system breaks the text document into sentences, sentences are further broken into words and after that stop words are removed. Preprocessing phase involves four steps:

Segmentation:The output of sentence segmentation phase is collection of sentences.

Tokenization: it is breaking down the sentences into words.

Stop words removal: they are meaningless and do not have any importance into the sentences.

Root Word Identification: it is identifying the words towards their root.

Feature extraction and machine learning

First step on Feature Extraction stage in our system is convert the sentences to vector space model which is a one common approach of representing sentences.Vector space model as shown in table 2 is a mathematical matrix that describes the frequency of terms that occur in a collection of documents. In a vector space approach, rows correspond to sentences while columns correspond to the weight of feature terms. Each sentence containsa set of features, while the label rows define the sentence is an important or not. Features represent properties of the sentences.

Table2: Vector space model

Sentence no	**Features**							**Label**
	F1	**F2**	**F3**	**F4**	**F5**	**F6**	**F7**	**Yes=1/No=0**
S1	X 11	X12	X13	X14	X15	X16	X17	**Y/N**
S2	X 21	X22	X23	X24	X25	X26	X27	**Y/N**
S3	X31	X32	X33	X34	X35	X36	X37	**Y/N**
…….	…..	…	……	……	……	……	……	**…..**
S_i	Xi n	Xi n	Xi n	Xi n	Xi n	Xi n	Xi n	**Y/N**

After convert the sentences to term matrix formula by used the space vector model, Our system converts the text type of training and testing data set to ARFF file form. After that, the training and testing data set for original and improve features selection input the machine learning on WEKA algorithms to evaluate the system with Naive Bayes, SVM and Decision Tree classifiers to get the performance from recall, precision and F-measure metrics as equation (13),(14) and(15). Thereafter, we get the results of original and improve features among classifiers with compression rate the summary 20% and 40% from original text**.**

Improving features

In our proposed system, we improved Sentence position, Sentence length, and Key word featuresafter that,classification model used to train and test the summarizer to extract important sentence. Vector space model as shown in table 2 used to perform this document representationin the training and testing matrix.

Improved Sentence Position (ISP)

According to most of the previous studies that claimed the sentences located at the beginning of the document are more important than those at the end of document. So we tried to modify the method of calculating the weights of sentences that fall in the second half of the document. Therefore, to improve the weights based on sentence position, our study used the same original equation (10) for the sentences in the first half of document and modified the equation for sentences in the second half of document only by doing the following step: according to the algorithm which is explained in below, in case (R) reaches a threshold, thereupon the system well reduces the(P) value by 1 each time; where R is equal to the total number of sentences in the document and P is countdown counter.(In our system we used a half or less than half value of the total number of sentences in the document as threshold because it has achieved the best results through experience).

When R becomes equal to 5, thereupon (R) reaches a threshold which is half or less than half value of the total number of sentences in the document, after that the system starts reduce the (P) value by 1 each time with continuing to decrease R value. Theweight of sentences in the document is calculated as follows:

The algorithm of scenario is:

```
Let R=no of sentences of text
Let countdown counter (Pi) = R
For (R; R> ½ Pi; R--)
}
Score f (Si) =(R / Pi);
{
For (R= ½ Pi; R> 1; R-- Pi--)
    }
Score f (Si) =(R / Pi);
    }
```

For example the first sentence =10/10=1, second sentence= 9/10=0.9… sixth sentence = 5 / (10-1) = 5/9 = 0.55 instead of 0.5 in the original equation, and the tenth sentence= 1 / (6-1) = 1/5 = 0.2 instead of 0.1 in the original equation.

Improved Sentence Length (ISL)

The main reason for the improvement of this feature is that most previous researches ignored long sentences, some of which may be important. So we tried to make a distinction between the long sentences and very long sentences by proposing a new mechanism for calculating sentence weight.To get the improvement of sentence length feature, our study follows these steps:

First step, we extracted the threshold which is the summation of words in the document divided by the number of sentences in this document.In this case, we apply the following formula:

$$\boldsymbol{Threshold} = \frac{\textbf{sum of words in the document}}{\textbf{number of sentences in document}} \dots\dots\dots\dots\dots\dots..(\mathbf{1})$$

Second step, we calculate the weight of sentence length for each sentence in a document by observing the threshold value which is two cases:

The first case, if the threshold value that was extracted in the first step is greater than or equal to the number of words in a sentence in this case we apply the following formula:

$$\boldsymbol{Scor\ f\ (Si)} = \frac{\textbf{number of words in a sentence}}{\textbf{threshold}} \dots\dots\dots\dots\dots..(\mathbf{2})$$

For example, if the number of words in a sentence equal to 15 and the threshold is equal to 20. Thus, the weight of the sentence length is 0.75.

In the second case, if the threshold value is smaller than the number of words in the sentence, we will have another two situations:

First situation if sentence is long and the number of words in a sentence greater than the threshold value and less than twice the threshold value .In this case, we apply the following formula:

$$\boldsymbol{Score\ f\ (Si)} = \mathbf{1} - \frac{(\textbf{number of words in a sentence} - \textbf{ threshold})}{\textbf{Threshold}} \dots\dots\dots\dots\dots\dots.(\mathbf{3})$$

For example, if the number of words in a sentence equal to 35 and the threshold is equal to 20 So that, the weight of the length of a sentence be 0.25.

Secondsituation if the number of words in a sentence is greater than the threshold value and at the same time greater than twice the threshold value .In this case, the value of the weight of sentence length is equal to zero because the sentence is very long.

Improved Key Words (IKW)

Depending on the researches and previous studies, we found that key words number have not been determined based on the size of the text, but have been adopted on other mechanisms such as the percentage of the number of words in the document or by reliance on the word frequency.Our system classified the number of key words according to the number of words in the text: if the number of items in the document is less than 500 then, the key words number is 4. When the words are more than 500 and less than 1000 then the key words number is 6. Otherwise the key words number is 8.our system used these optimal values threshold depending on the experiments results.

Findingthe best combination

Our system evaluated the second group of featureswithout any improving to get the best combination among them, this group contains 7 features.

Term Frequency

Depending on studies [3][1].It is the number of occurrences of the term in that document as equations (4) and (5).

$$\mathbf{tf_i} = \frac{\mathbf{n_i}}{\sum_{\cdot\mathbf{k}} \mathbf{n_k}} \dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots\dots(\mathbf{4})$$

Where (n $_i$) isthe number of occurrences of the considered term in document, and ($_k n_k$) in the denominator is a number of occurrences of all terms in document. After the term frequencies of all terms in the sentence are found, the term frequency score of a sentence S is calculated as follows:

$$\textbf{Score f}\ (\mathbf{S_i}) = \frac{\sum_{\mathbf{i=1}}^{\mathbf{m}} \mathbf{tf_i}}{\mathbf{m}} \dots\dots\dots\dots\dots..(\mathbf{5})$$

Where m = the number of terms in the sentence s.

Sentence Centrality

Sentence centrality is the similarity between a sentence and other sentences in the document. The centrality of a sentence can be evaluated as the degree of vocabulary imbrication between the sentence and other sentences [20][12][3].In order to get the centrality degree of a sentence S, the following formula is used.

$$\boldsymbol{Score\ f\ (S_i) = \frac{|words\ in\ \ S\ \cap words\ in\ other\ sentences\,|}{|words\ in\ \ S\ \cup\ words\ in\ other\ sentences|} \ldots\ldots\ldots..(6)}$$

Title Similarity

Title contains the set of words that give important clues about text concept. Sentence propinquity to the title is the vocabulary overlap between this sentence and the text title [12][1].If the sentence has a large degree of intersection with the title words, then the score of this feature is higher. Hence, we can formalize Title similarity score of a sentence S as equation (7).

$$\boldsymbol{Score\ f\ (S_i) = \frac{|words\ in\ \ S\ \cap words\ in\ the\ title\,|}{|words\ in\ \ S\ \cup\ words\ in\ the\ title|} \ldots\ldots\ldots(7)}$$

Upper-Cased Words

This feature assigns higher scores to words that contain one or more upper case letters [26]. It can be a proper name or important word as equations suggests.

$$\boldsymbol{TCW(S_i\,) = \frac{\mathrm{NCW}(s)}{\mathrm{NW}(s)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(8)}$$

Where TCW(s) =Total of first letter capital words in sentence. NCW(s) =Number of first capital words in sentence .

NW(s) =Number of words in sentence.After that, we get the sentence feature score as:

$$\boldsymbol{Score\ f(S_i) = \frac{\mathrm{TCW}(S_i)}{\mathrm{MAX}\big(\mathrm{TCW}(S_i)\big)} \ldots\ldots\ldots..\ldots\ldots\ldots(9)}$$

Sentence Position

Sentence position is the sentence location in the document. Therefore, each sentence has a different importance. According to study [35], the sentences in the first positions of the textshould be embedded in the summary as given in (10). The general formula of the sentence position is:

$$\boldsymbol{Score\ f(S_i) = \frac{\mathrm{R} - p_j}{R} \ldots\ldots\ldots\ldots\ldots\ldots..(10)}$$

Where, p_j is the position value (p_j is equal to j counter, whereas the position score of the first sentence is 0). R is the total number of sentences in the corresponding document.

Sentence Length

This feature is used to penalize sentences that are too short; hence these sentences are not expected to be in the summary [20][12][3].Too short or too long [1], these sentences are not considered as an optimal selection. Finally, it penalizes sentences that are shorter than a certain length [17].Therefore, the sentence length is the ratio of the number of words in the sentence divided by the amount of words occurring in the largest sentence of the text as given by (11).

$$\boldsymbol{Score\ f\ (s_i) = \frac{W(s_i)}{\mathrm{LongSen}} \ldots\ldots\ldots\ldots\ldots\ldots(11)}$$

Where W ($\boldsymbol{s_i}$) is Total number of words in sentencess_i. LongSen is Total no of words in the longest sentence.

Key Words

Keywords are words that appear with unusual high frequency in a text document. According to [14], the Keywords are the 10% words with the max frequency system. On this basis, the first step is select 10% term frequency in a document treated as key words. (i.e.) we are ranking the terms in document according to their iteration and choice top 10%tem frequencies. Second step is calculating the feature score of each sentence according to number of key words in it. Key words feature weights in each sentence are calculated according the equation (12).

$$\boldsymbol{Score\ f(s_i) = \frac{\mathrm{key\ words}(s_i)}{Sen\ len(s_i)} \ldots\ldots\ldots\ldots\ldots\ldots\ldots(12)}$$

Where Key word ($\boldsymbol{s_i}$) is total no of keywords in sentence ($\boldsymbol{s_i}$). Senlen ($\boldsymbol{s_i}$) is no of terms in the sentence ($\boldsymbol{s_i}$).

Post processing

In this stage the selected sentences are sorted according to their order in the original text to have a good-form summary. The text summarization size depends on compression rate which is the ratio between summary length and the original text length. It is very important parameter for each summary, it allows determining how much information needs from the source text and usually the good summary is from 5% to 30% [23].in our system which is 20% or 40%.

EXPERIMENTAL RESULTS and DISCUSSION

In this chapter, we describe our Data Set, evaluation techniques, and we present the evaluation results of our summarization system then discuss the results.

Data set

Our paper used two data sets In order to test the performance of our summarization system. The first data set is a collection of 100 newspaper articles, and their summaries are created by four evaluatorsIndependent human annotators, who are journalists and graduate students to choose important sentences from those newspaper articles in order to create their summaries. 2nd data set is a collection of 100 articles created by The European Commission-Joint Research Centre (JRC) and their summaries are created by the authors of article.

Evaluation metrics

The study can evaluate the text summarization system performance in different approaches. In our research, we have selected intrinsic evaluation method. Intrinsic evaluation judges the accuracy of a machine learning generated summary based on the conformity between manual summary and the generated summary by using training and testing data sets. The recall (13),precision (14), and f-measure (15) metrics are used to judge the covering human summaries and machine learning summaries. Recall is the fraction of the number of correctly selected sentences divided by the number of all sentences in the human generated summary. Precision is the fraction of the number of correctly selected sentences divided by the number of all sentences in the machine generated summary. F-measure, precision and recall, provides a method for combining precision and recall scores into a single value.These evaluation metrics are given in the following formulas.

$$\boldsymbol{Recall} = \frac{|\boldsymbol{S} \cap \boldsymbol{T}|}{|\boldsymbol{S}|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (\mathbf{13})$$

Where: S is machine generated summary, T is a manual summary.

$$\boldsymbol{Precision} = \frac{|\boldsymbol{S} \cap \boldsymbol{T}|}{|\boldsymbol{T}|} \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots . (\mathbf{14})$$

$$\boldsymbol{f - measure} = \frac{\mathbf{2}(\boldsymbol{Precision} * \boldsymbol{Recall})}{(\boldsymbol{Precision} + \boldsymbol{Recall})} \ldots\ldots\ldots\ldots\ldots . (\mathbf{15})$$

Our study uses Environment for Knowledge Acquisition (WEKA) Created by researchers at the University of Waikato in New Zealand .The experiments were conducted using SVM, Naive Bayes and Decision Stump classifiers that are already implemented in Java language.WEKA divided the two data sets into training and testing corpus. Our study will used SVM, Naive Bayes and Decision tree classifiers to evaluate the system performance with 20% and 40% compression rates. The machine learning step evaluates two sets of features The first one contains a combination of seven features which are Term frequency(TF), Sentence centrality(SC), Title similarity(TS), Upper case of word(UC), Sentence position(SP), Sentence length(SL), and Key word(KW); While the second group involves same seven features with improvements in three features that are: Improved Sentence position(ISP), Improved Sentence length(ISL), and Improved Key word(IKW) to enhance the performance of text summarization system.

Classifiers types

We used three classifiers:SVM, Naive Base, and Decision Tree.

SVM classifier

In 1963 thelinear classifiers of original Support vector machines (SVM) or support vector networks algorithm was discovered by Vladimir Vapnik and Alexey Chervonenkis,in 1992, Bernhard, Isabelle, and Vladimir Vapnik suggested a way to create nonlinear classifiers.SVMs are supervised learning models used for classification and text summarization to recognize patterns, extracts document summary, and analyze data with associated learning algorithms. It uses support vector points to find a border among the classes.it is either linear which is a data point is viewed as a list number of p-dimensional vector to separate points with a (p-1)-dimensional hyper plane or nonlinear using what is called the kernel trick, by mapping their inputs into high-dimensional feature spaces. In WEKA environment our system used a liner classifier called smo which is mean SVM.

Naive Bayes classifier

Naive Bayes was discovered in the early 1960s.Itis fast, space efficient, and a simple technique for constructing classifiers: modelsshowed as vectors of feature valuesafter assigningclass labels which are drawn from some finite set to problem instances. Theclassifier detectsthe probability of the previously unseen instancebelonging to which class, thereafter simply chose the most likely class.

Decision Tree classifier

Decision Tree Classifier is a simple and widely used classification technique. The decision tree induction algorithm works by recursively selecting the best attribute to split the data and expanding the leaf nodes of the tree until the stopping criterion is met.This Classifierposes a set of accurately crafted questions about the features of experiment record. The terminal node is assigned a class label Yes or No.

Experiments results

In this section our system evaluates the corpus by using three different classifiers which include recall, precision, and f-measure evaluation methods. The figures 2 and 3 represents the final experiments results, our study explain the results in the next section.

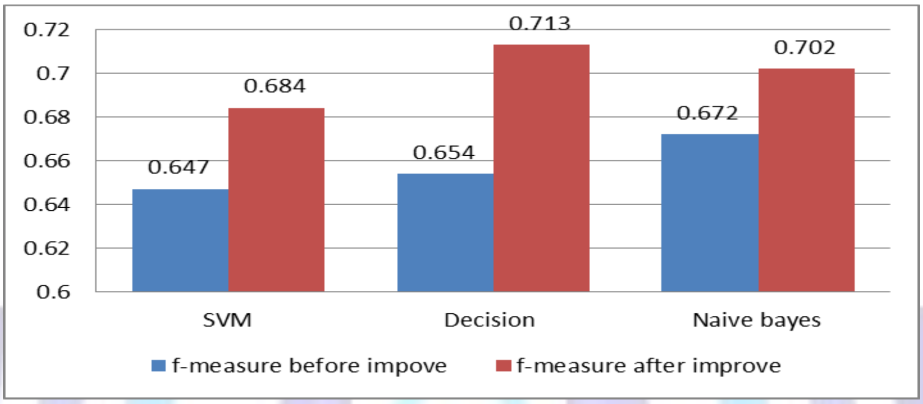

Figure2. F-measure results of SVM, Decision and Naïve Bayes for 40% CR

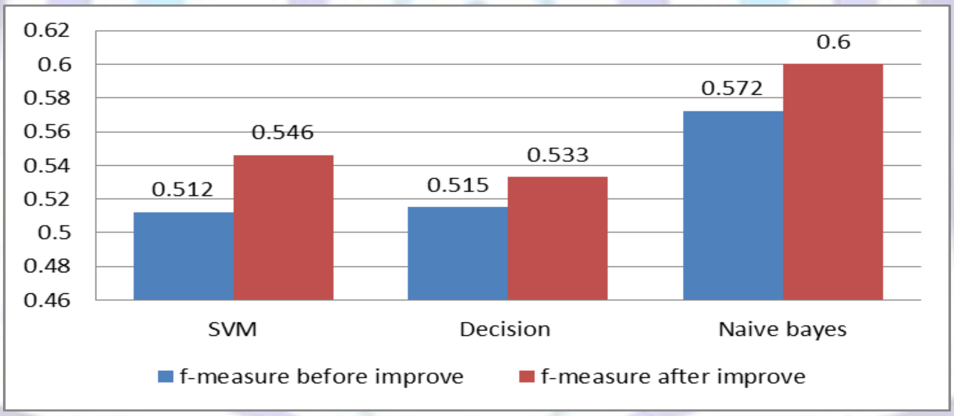

Figure3. F-measure results of SVM, Decision and Naïve Bayes for 20% CR

SVM classifier results

By using the SVM metric we get the recall, precision and f-measure grades for first feature group. Then we compare the result with second feature set that involves the three improved features.in the summarization process stage our first goal is evaluate the effectiveness of the seven original features combination group which is (case 1) that considered as 40% and 20% CR. In 40% CR our results of recall, precision and f-measure metrics produced 0.738, 0.598 and 0.647 successively; while the results of the effectiveness with 20% compression rates are 0.574, 0.493 and 0.512 respectively.

Similarly, the second measure is to evaluate the effectiveness of the three improved features with the four remaining features which is (case 2) to finding the weights of features with 40% and 20% compression rates. The results showed that the best recall, precision and f-measure results are produced when we used the three improved features (IKW, ISP, and ISL) together with (TF, SC, TS, and UC). The results were 0.773, 0.624, and 0.684 respectively with 40% CR; Instead of 0.610, 0.527and 0.546 successively with 20% CR.

Naive Bayes classifier results

the results of evaluate the effectiveness of the seven original features combination group which is (case 1) that considered as 40% and 20% CR. In 40% CR our results of recall, precision and f-measure metrics showed 0.671, 0.680 and 0.672 respectively. On the other hand the results of the metrics with 20% CR were 0.580, 0.575and 0.572 successively.

On the other hand, the second measure is to evaluate the effectiveness of the three improved features with the four remaining features which are (case 2) to finding the optimal weights. The results showed that the best score of recall, precision and f-measure are produced when we used the three improved features (IKW, ISP, and ISL) together with (TF, SC, TS, and UC). The results were 0.705, 0.706, and 0.702 successively with 40% CR; In contrast to 0.608, 0.596 and 0.600 respectively with 20% CR.

Decision Tree classifier results

The third classifier is Decision Tree; our study get the recall, precision and f-measure values for two feature groups in same in the same way as SVM and Naive Bayes. The results of evaluate the effectiveness of the seven original features combination group which is (case 1). In 40% CR our results of recall, precision and f-measure metrics produce of 0.703, 0.621, and 0.654 respectively. In same time the results of the effectiveness with 20% compression rates are 0.547, 0.526, and 0.535respectively.

On the other hand, the second measure is to evaluate the effectiveness of the three improved features with the four remaining features which are (case 2) to finding the optimal weights of features. With40% CR the results showed that the best score of recall, precision and f-measure are produced when we used the three improved features (IKW, ISP, and ISL) together with (TF, SC, TS, and UC) the results were 0.757, 0.685, and 0.713 respectively .On the other hand with 20% CR, the top of result through combined three improved features (IKW, ISP, and ISL) together with (TF, SC, TS, and UC) was in the recall, precision, and f-measure which are 0.571,0.545 , and 0.553 respectively.

Discussion

The Recall, Precision, and f-measure results ,which are displayed in Table(4) , Table(5),and Table (6) indicate that the effectiveness when we combine of the three improved features together with the four remaining features in both40% and20% CR gave us the best results. The reason of the good results is that intrinsic methodis a good approach to extract the important sentences because their technical ability to match the correctly sentences.The second reason is the correctly select of the important sentences that are chosen by the annotators in the data set that helps our system to generate a right summary.The drawback in our system that is needs some development through the use of more features andtests the system in other evaluation methods, such as Pyramid or ROUGE methods.When we look at the big picture, we can say that the improved sentence position (ISP) and improved sentences length (ISL) features are the most important features to form a good summary when combined with other features.

CONCLUSION

Our proposed system tries to get best results over combining among a new features and improving some it.We can draw the following conclusions from experiments results: (1) the values of precision, recall, and f-measure for all the three classifiers with compression rate of 40% are significantly higher than compression rate of 20% results. (2) The best results were obtained by our trainable summarizer with Decision Tree classifier for 40% compression rates; while the better result for 20% compression rates was Naive Bayes classifier.(3) the best improved feature was (ISP) and Somewhat lesser extent (IS L) and (IKW).

We planning in the future work to increasing the features and adding some new scenarios to improve the features.Besides these, we plan to apply these document features to Arabic text summarization.

REFERENCES

[1] Abuobieda A. , Salim N.,Albaham A.,Osman A., and Kumar Y. J 2012.Text summarization features selection method using pseudo genetic-based model.In International conference on information retrieval knowledge management. (2002), 193–197.

[2]Ahmad Najibullah 2015. Indonesian Text Summarization based on Naïve Bayes Method. Proceeding of The International Seminar and Conference 2015, 1(1), 67-78.

[3] Arman B. Kiani, Akbarzadeh T., and Moeinzadeh M. 2002 .Intelligent Extractive Text Summarization Using Fuzzy Inference Systems. IEEE Conference on Intelligent Engineering, 149-153.

[4] ArmanKiani and M.R. Akbarzadeh2006.Automatic Text Summarization Using: Hybrid Fuzzy GA-GP. In Proceedings of 2006 IEEE International Conference on Fuzzy Systems, Canada, 977-983.

[5] Barzilay R, McKeown K, and Elhadad M 1999.Information fusion in the context of multi-document summarization.Proceedingof ACL.

[6] Baxendale, P. 1958. Machine-made index for technical literature - an experiment. IBM Journal of Research Development, 2(4), 354–361.

[7]Climenson W. D., Hardwick N. H.,and Jacobson S. N. 1961.Automatic syntax analysis in machine indexing and abstracting. American Documentation , 178-183.

[8] Conroy J. M, Schlesinger J. D, and Gold stein J 2005.CLASSY Query-Based Multi Document Summarization.In the Document Understanding Workshop (presented at the HLT/EMNLP Annual Meeting), Vancouver, B.C., Canada.

[9] DeJong G.F. 1982.An Overview of the FRUMP System.

[10] Edmundson, H. P. 1969. New methods in automatic extracting. Journal of the ACM, 16(2):264–285.

[11] Eman Al-Thwaib 2014.Text Summarization as Feature Selection for Arabic Text Classification.World of Computer Science and Information Technology Journal, 4(7), 101-104.

[12] Fattah Mohamed Abdeland Ren Fuji 2009.Ga, mr, ffnn, pnn and gmm based models for automatic text summarization.Computer Speech and Language, 23(1), 126–144.

[13] Fum D., Guida G., and Tasso, C. 1985. Evaluating Importance: A step towards text summarization.

[14]Gurmeet Singh 2014. A Novel Features Based Automated Gurmukhi Text Summarization System, master thesis. Computer Science and Engineering Department Thapar University, Patiala, India.

[15] Ha Nguyen Thi Thu 2014. An Optimization Text Summarization Method Based on Naive Bayes and Topic Word for Single Syllable Language.Applied Mathematical Sciences, 8(3), 99 – 115.

[16] Hirao T, Sasaki Y, Isozaki H, et Al 2002).NTT's Text Summarization system for DUC-2002. In Workshop on Text Summarization (In conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC2002 Meeting on Text Summarization), Philadelphia.

[17]Hovy, E. and Lin, C.Y. 1997.Automatic text summarization in SUMMARIST. In proceedings of the ACL'97/EACL'97 workshop on intelligent scalable text summarization, Madrid, Spain , 18-24.

[18] Jen-Yuan Yeh, Hao-RenKe, Wei-Pang Yang, and I-HengMeng 2005. Text summarization using a trainable summarizer and latent semantic analysis.Information Processing and Management 41(1),75-95.

[19]Kai ISHIKAWA, ANDO Shin, DOI Akitoshi, and OKUMURA 2002.Trainable Automatic Text Summarization Using Segmentation of Sentence. Proceedings of the Third NTCIR Workshop in Japan .

[20] Kupiec J. , Pedersen J. O., and Chen, F.1995.A trainable document summarizer. In Proceedings of the 18th ACM-SIGIR Conference Association of Computing Machinery , 68-73.

[21] Lehnert W. 1981.In introduction to plot units.

[22] Luhn H. P. 1958.The automatic creation of literature abstracts. IBM Journal of Research and Development, 2(2), 159–165.

[23] Mani, 2001. Summarization evaluation: An overview.In Proceedings of the North American chapter of the association for computational linguistics (NAACL) workshop on automatic summarization.

[24] MucahidKutlu, CelalCıgır, and IlyasCicekli2010 .Generic TextSummarizationfor Turkish . the Computer Journal, 53(8), 1315-1323.

[25] Pollock J. J., and Zamora A. 1975. Automatic Abstracting Research at Chemical Abstracts Service. Journal of Chemical Information and Computer Science, 226-232.

[26] Prasad, Rajesh Shardanand, Uplavikar, NitishMilind, Wakhare, SanketShantilalsa, Jain, Vishal, Yedke and TejasAvinash2012. Feature based text summarization. International Journal of Advances in Computing and Information Researches.

[27] Ramiz M. Aliguliyev 2009.A new sentence similarity measure and sentence based extractive technique for automatic text summarization. Institute of Information Technology of National Academy of Sciences of Azerbaijan,Baku, Azerbaijan, Expert Systems with Applications , vol 36(4), pp. 7764–7772.

[28] Rasim M. Alguliev, Ramiz M. Aliguliyev, and Nijat R. Isazade2012.DESAMC+DocSum: Differential evolution with self-adaptive mutation and crossover parameters for multi-document summarization. Institute of Information Technology, Azerbaijan National Academy of Sciences, Baku, Azerbaijan, v 36 ,pp.21–38.

[29] Rath G. J., Resnick A., and Savage, T. R. 1961.The Formation of Abstracts by the Selection of Sentences. American Documentation, 139-141

[30] Reimer U., and Hahn U. 1988. Text condensation as knowledge base abstraction.

[31] Rene A Garcia and YuliaLedeneva 2009. Word sequence models for single text summarization. Second International Conferences on Advances in Computer-Human Interactions IEEE computer society, 44-48.

[32] Stefan Thomas, Christian Beutenmuller, Xose de la Puente Robert Remus, and Stefan Bordag 2015.ExB Text Summarizer.Prague, Czech Republic, Proceedings of the SIGDIAL 2015 Conference, pp. 260–269.

[33] Svore K, Vanderwende L, and Burges C 2007.Enhancing Single-Document Summarization by Combining RankNet and Third-Party Sources. In Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language learning (EMNLP-CoNLL), pp. 448–457.

[34] Wei Wang, HuaXu2013.Implicit feature identification via hybrid association rule mining . Tsinghua University, Beijing, China, Expert Systems with Applications v 40 ,pp. 3518–3531.

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

Author' biography with Photo

Madhi A. Ali he received Higher Diploma in Information Technology from Iraqi Commission for Computers & Informatics in 2010.Madhi A. completed his B.Sc. in computer engineering and Information Technology from technology University, Iraq in 2003. Currently he is a Master student in the department of Computer Science, Faculty of Science and Information Technology, Al-Zaytoonah University of Jordan.

Prof. Dr. Ali A. Al-Dahoud he is Dean of the Faculty of Science and Information Technology at Al-Zaytoonah University. He received his Ph.D. in engineering sciences, National Technical University of Ukraine 1996; his B. SC. engineer of organization (computer), Belgrade, Yugoslavia, 1986. Dr. Ali A. Al-Dahoud is Full Professor, Senior IEEE, MIAENG, SMASDF. His current research interests include Distributed Systems (Communication), Algorithms, Data mining and E-Systems. He has various publications in referred journals and conference proceedings.

Dr. Bilal Hawashin is currently an Assistant Professor in the Department of ComputerInformation Systems at Alzaytoonah University of Jordan. He received his Ph.D inComputer Science, College of Engineering from Wayne State University in 2011. Alsohe worked in the Department of Computer Information Systems at Jordan Universityof Science and Technology from 2003-2007. His current research interests includeSimilarity Join, Text Mining, Information Retrieval, and Database Cleansing. He hasvarious publications in referred journals and conference proceedings. Dr. Hawashinreceived his B.S. in Computer Science from The University of Jordan in 2002, and hisM.S. in Computer Science from New York Institute of Technology in 2003.