Journal of Engineering and Computer Innovations Vol. 3(1), pp. 11-25, February 2012
Available online at http://www.academicjournals.org/JECI
DOI: 10.5897/JECI11.053
ISSN 2141-6508 ©2012 Academic Journals

Full Length Research Paper

New approaches to automatic headline generation for Arabic documents

Fahad Alotaiby[1]*, Salah Foda[1] and Ibrahim Alkharashi[2]

[1]Department of Electrical Engineering, College of Engineering, King Saud University, Riyadh, Saudi Arabia.
[2]Computer Research Institute, King Abdulaziz City for Science and Technology, Riyadh, Saudi Arabia.

Accepted 23 December, 2011

A headline is considered a condensed summary of a document. The necessity for automatic headline generation has been on the rise due to the need to handle a huge number of documents, which is a tedious and time-consuming process. Instead of reading every document, the headline can be used to decide which ones contain important and relevant information. There are two major approaches to automatic headline generation. The first is linguistic, in which the knowledge about the structure of the language itself is considered. The second approach is statistical and it comprises all quantitative approaches to automated language processing. However, the Arabic language has a different statistical structure than the English language, and requires special treatment to generate Arabic headlines, especially when there is no dedicated technique for the Arabic language. Therefore, two new statistical methods in automatic headline generation have been developed to create representative headlines for textual documents in the Arabic language. The first is an extractive method based on character cross-correlation, and the second one is an abstractive method based on the hidden Markov model (HMM). The extractive method achieved ROUGE-L of (0.1938) and the HMM method achieved ROUGE-L of (0.2332). In addition, both techniques were assessed via human examiners who evaluated the resulting headlines.

Key words: Summarization, automatic headline generation, hidden Markov model, language model.

INTRODUCTION

Headline generation is an important field of natural language processing (NLP), which includes language analysis, understanding, and synthesis. Thus, generating a headline for a textual document requires analyzing the document, understanding the main idea of the document, and creating a headline that reflects the content of the document. Therefore, the problem of headline generation concerns complex language processing. A headline is a condensed summary of a document that accurately represents the main idea of that document. From this definition, it is obvious that headline generation is a compressed version of summarization, and thus the study of headline generation is a part of the summarization field. The increased amount of information emerging in the modern digital world has created an information overload (Yang et al., 2003). Information overload refers to the difficulty in understanding a topic and making decisions because of the presence of too much information. Therefore, the necessity of automatic headline generation has been raised due to the need to manually review huge numbers of documents, which is a tedious and time-consuming process. Instead of reading every document, the headline can be used to decide which of them contains important or relevant information. Automatic headline generation can be classified according to several dimensions, such as linguistic versus statistical or extractive versus abstractive.

*Corresponding author. E-mail: falotaiby@hotmail.com. Tel: +966 5050429828. Fax: +9661 4352633.

Abbreviations: CCC, Character cross correlation; **DUC,** document understanding conference; **EWM,** exact word matching; **HMM,** hidden markov model; **HTK,** hidden markov model toolkit; **LDC,** linguistic data consortium; **LM,** language model; **MSA,** modern standard Arabic; **NIST,** national institute of standards and technology; **NLP,** natural language processing; **ROUGE,** recall-oriented understudy for gisting evaluation.

In the extractive approaches, the most suitable text unit is extracted from the original document, and then it may be trimmed to the proper size. However, in abstractive headline generation, the original document is analyzed and proper headline words are selected and ordered to represent a consistent and readable headline. On the other hand, statistical approaches include all quantitative approaches to automatically processing the document and generating a headline (Manning and Schütze, 1999). In contrast, the linguistic approaches include the use of knowledge about the structure of the language itself to analyze the document and generate the headline (Allen, 1995). This paper presents two new methods. The first is an extractive approach that employs character cross-correlation to extract the best headlines and overcome the complex morphology of Arabic language. The second is a statistical abstractive approach that utilizes the HMM and statistical language model (LM) to automatically construct a headline for Arabic documents containing news stories. The resulting headlines are evaluated using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) (Lin, 2004a), in addition to human evaluation by a group of examiners. The next section presents related work in the area of automatic headline generation. Then, a brief introduction about the Arabic language, used datasets and headline length is presented, followed by description of the presented approaches and experimental designs. Finally, the results are presented with comments and discussions.

RELATED WORK

There are several systems that automatically generate headlines for documents in languages other than Arabic. Some of them are extractive (Songhua et al., 2010; Lloret and Palomar, 2011) and some others are abstractive (Reddy et al., 2011). For the Arabic language, there is only one system dedicated to generating very short summaries (headlines), *Lakhas* (Douzidia and Lapalme, 2004). *Lakhas* was one of the systems presented at the Document Understanding Conference in 2004 (DUC, 2004), which the American National Institute of Standards and Technology (NIST) organized. Task 3 in this conference was to generate a very short summary of a machine-translated text from Arabic into English. In contrast to the systems in the Document Understanding Conference (2004), *Lakhas* first summarizes the original Arabic document and then applies machine translation to the summary only. Therefore, the published results were not for the Arabic headline, but for the translated headline into English. On the other hand, Conroy and Leary (2001) used HMM to extract sentences from the document in English to form a summary. Only one HMM model was used with 2*s*+1 states, where *s* represents the number of sentences in the summary. In addition, three main features were utilized: position of the sentence, number of terms in the sentence, and the likelihood of the sentence terms given the document terms. In a similar way, HMM Hedge (Zajic et al., 2002) is an algorithm for selecting headline words from a document based on a standard “noisy channel” model of processing with a subsequent decoder for producing headline words from stories. In HMM Hedge, there is only one HMM. It has two types of states: headline state or gap state. The HMM is constructed with states for only the first *N* words of the story, where *N* is a constant (60) or the number of words in the first sentence.

ARABIC LANGUAGE

The Arabic language is a Semitic language spoken by more than 280 million people. Arabic was originally an oral language. For that reason, the classical Arabic writing system was originally consonantal. Each of the 28 letters in the Arabic alphabet represents a single consonant. To overcome the problem of different pronunciations of consonants in Arabic text, graphical signs, known as diacritics, were invented in the seventh century. Currently in Modern Standard Arabic (MSA), diacritics are almost always omitted from written text. As a result, this omission increases the number homographs (words with the same written form). However, Arab readers normally differentiate between homographs by the context of the script. Arabic is a morphologically complex language. An Arabic word may be constructed from a stem as well as affixes and clitics. Furthermore, some parts of the stem may be deleted or modified when appending a clitic to it according to specific orthographical rules. As a final point, different orthographic conventions exist across the Arab world (Buckwalter, 2004a). As a result of omitting diacritics, complex morphology, and different orthographical rules, two of the same stem words may be regarded as different if compared literally.

In Arabic, clitics are attached to a stem or to each other without any orthographic marks (that is, an apostrophe). A clitic is a linguistic unit that is pronounced and written like an affix, but it is grammatically independent. Linguistically speaking, if one can parse an Arabic linguistic unit attached to a stem, it should be considered a clitic. This covers most of the clitics, except the definite article {*Al* ال}. It is important to mention that the transliteration used in this work is based on the style proposed by Buckwalter (2004b). The number of clitics in Arabic is limited. However, when concatenated, clitics can generate a chain of up to four clitics before the stem (proclitics) and three clitics after the stem (enclitics) (Alotaiby et al., 2010). Clitics and affixes attached to stems make a direct comparison between words impractical. Therefore, this work employed character cross-correlation to extract the best headlines and overcome the Arabic language’s complex morphology.

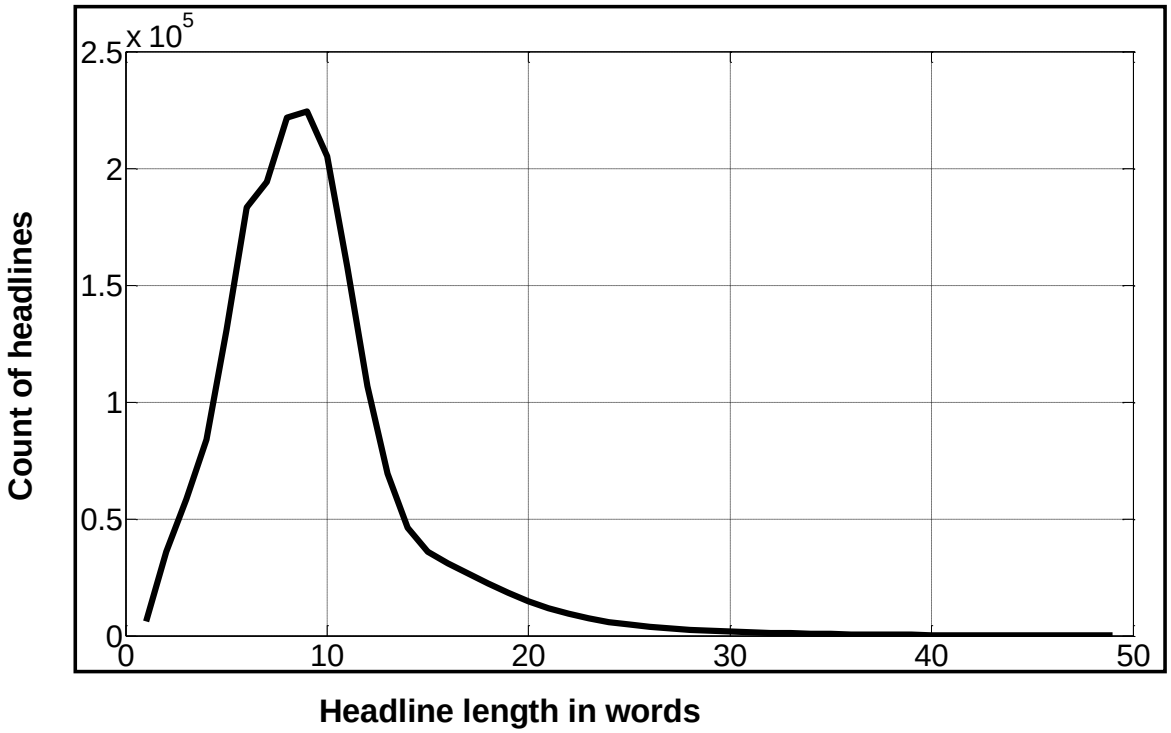

Figure 1. Original headline length distribution in the Arabic Gigaword corpus.

DATASETS

Datasets used in this work were extracted from the Arabic Gigaword (Graff, 2007). The *Arabic Gigaword* is a collection of text data extracted from newswire archives of Arabic news sources, and their titles that have been gathered over several years by the Linguistic Data Consortium (LDC) at the University of Pennsylvania. Text data in the *Arabic Gigaword* were collected from four newspapers and two press agencies. The *Arabic Gigaword* corpus contains almost 2 million documents with nearly 600 million words. However, there are some problems with this dataset, such as spelling mistakes, inconsistent use of punctuation and documents that have no headlines. Yet, simple problems in this corpus such as the presence of odd control characters and word binding were automatically corrected. A common problem in the Arabic corpus is the omission of white spaces between main tokens that end with graphically non-connecting characters, as in the following paragraph: {… *and fighting the spread of nuclear weapons pointing* …, ... ومكافحةانتشارالاسلحةالنوويةمشيرا}, which contains five connected words.

HEADLINE LENGTH

In the Document of Understanding Conference in 2004 (DUC, 2004), an evaluation of the very short summary was done on the first 75 bytes of the summary. Knowing that the average word size in Arabic is five characters (Alotaiby et al., 2009) in addition to space characters, the specified summary size in Arabic words will be roughly equivalent to 12 words (assuming each byte represents a single character). In the meantime, the average length of the original headlines in the *Arabic Gigaword* corpus was approximately 9.5 words. Figure 1 shows the headline length distribution in the words used in the corpus. In this work, a 10-word headline is considered as an appropriate length.

BASELINE HEADLINES

Since there are neither official Arabic datasets for automatic headline generation nor published results on Arabic documents as per the knowledge of the authors, it is important to find some ways to assign the resulting evaluation scores a meaning. Therefore, two techniques (besides the original headline that comes with every document under the test) are used to show the upper, lower, and baseline scores for evaluating the proposed techniques. The first baseline technique is a headline-generation system that randomly selects 10 words from the document (Rand-10). This headline represents the worst-case headline. In contrast, the headline is generated by the author of the document (Original) represents the best-case headline. These two headlines

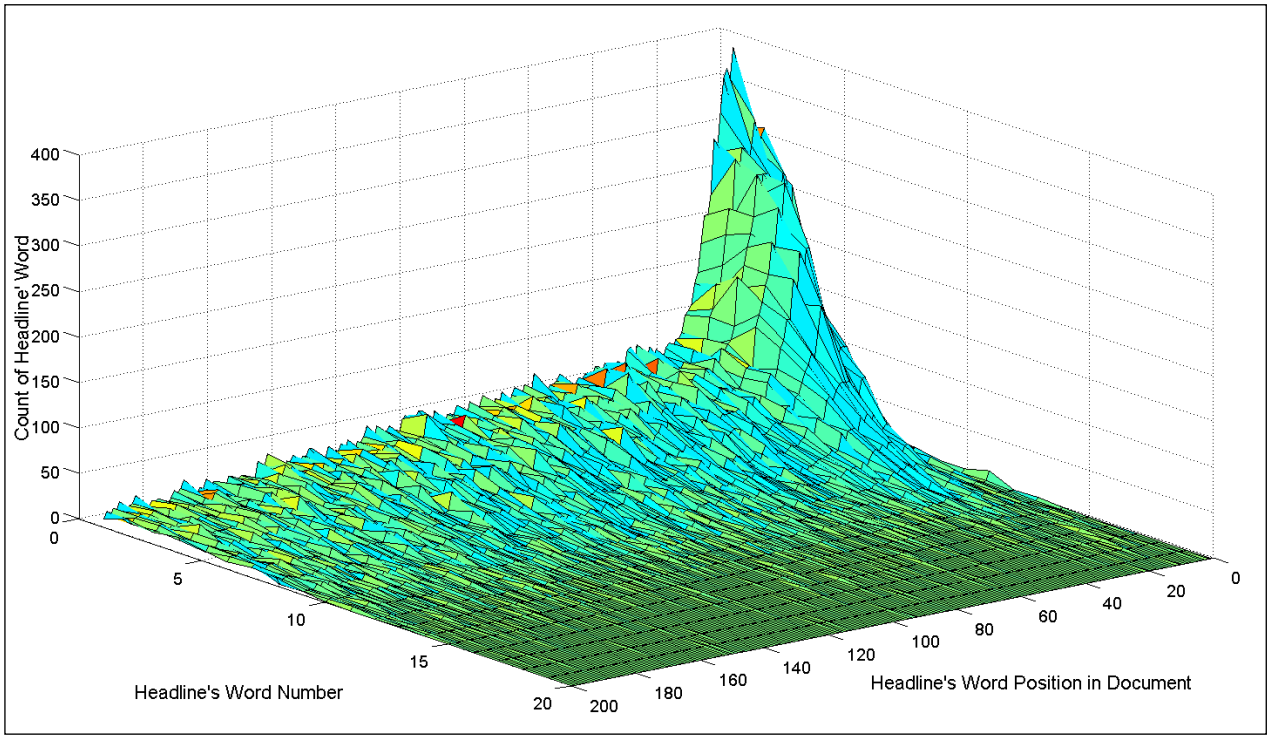

Figure 2. Distribution of headline words over the document.

are the extreme ranges of any evaluation scores. Summarization researchers have observed that the lead sentence of an English news story is often an appropriate summary of the text. Therefore, some of the headline generation systems utilize only the first sentence to generate a headline (Zajic et al., 2002). Those observations are based on English document stories. Therefore, to ensure that this is also the case in Arabic, a statistical study on headline words in Arabic Gigaword was performed to determine the distribution of headlines words among the documents words. Figure 2 shows the distribution of where the words of the headlines appear in the documents in Al-Hayat newspaper stories. It is clear that the first 60 words of the stories contain most of the headline words and that most of the headlines contain fewer than 10 words. As a result, the first 10 words of the document were selected as a baseline headline (Lead-10). It is important to mention that these 10 words are taken in the same sequence as they appear in the document; this gives headline more credit when it is evaluated, especially using automatic evaluation systems such as ROUGE.

EVALUATION TOOLS

Correctly evaluating the automatically generated headlines is an important phase. Automatic methods for evaluating machine-generated headlines are preferred to human evaluations because they are faster, more cost-effective, and can be performed repeatedly. However, they are not trivial because of various factors such as the readability of headlines and consistency of the headlines (whether the headlines indicate the main content of the news story). Hence, it is difficult for a computer program to judge. However, some automatic metrics are available for headline evaluation. BLEU (Papineni et al., 2002) and ROUGE (Lin, 2004a) are the main metrics used. The evaluation of this experiment was performed using Recall-Oriented Understudy for Gisting Evaluation (ROUGE). ROUGE is a system used for measuring the quality of a summary by comparing it to a correct summary created by human. ROUGE provides four different measures: ROUGE-n (usually n = 1, 2, 3 and 4), ROUGE-L, ROUGE-W, ROUGE-S, and ROUGE-SU. Lin (2004*b*) showed that ROUGE-1, ROUGE-L, ROUGE-SU, and ROUGE-W were very reliable measures in the short summaries category, and they will be recorded for this work.

PROPOSED APPROACHES

Two main approaches are presented with different technical variations in each of them. The first is an extractive method of automatic headline generation that utilizes the cross-correlation of letters to overcome the heavy existence of clitics and affixes in Arabic. The

Table 1. An example of headline nomination.

a	ارتبطت نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية الإسلامية، The emergence of the Arabic manuscripts in Sudan was associated with the rise of the Arabic-Islamic culture
b	ارتبطت نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية The emergence of the Arabic manuscripts in Sudan was associated with the rise of the Arabic culture
c	نشأة المخطوطات العربية في السودان ببروز معالم الثقافة العربية الإسلامية The emergence of the Arabic manuscripts in Sudan … with the rise of the Arabic-Islamic culture

second is an abstractive method in which the hidden Markov model and different statistical language models are used to build a meaningful headline that represents the corresponding document. The following subsections provide details of the proposed approaches.

Extractive automatic headline generation

The main idea of the used method is to extract the most appropriate set of consecutive words (phrase) from a document body, which should represent an adequate headline for the document. Then, those headlines are evaluated by calculating the ROUGE score against a set of three reference headlines. To do so, a list of nominated headlines was first created from the document body. After this, four different evaluation methods were applied to choose the best headline that reflects the idea of the document among the nominated list. The aim of these methods is to determine the most suitable headline that matches the document. The idea here is to choose the headline that contains the largest number of the most frequent words in the document, while ignoring stop words and giving more weight to earlier sentences in the documents.

Nominating a list of headlines

A window of a length of 10 words was passed over the paragraphs word by word to generate chunks of consecutive words (sentences) that could be used as headlines. Moving the window one word at a time may corrupt the fluency of the sentences. A simple approach to reduce this issue is to minimize the size of the paragraphs. Therefore, the document body was divided into smaller paragraphs at new-line, comma, colon and period characters. This step increased the number of nominated headlines with a proper start and end. The result is a nominated list of headlines with a length of 10 words. In the case of a paragraph containing fewer than 10 words, there will be only one nominated headline of the same length for that paragraph. Table 1 shows an example of a nominating headline list, where *a* is the selected paragraph, *b* is the first nominated headline, and *c* is the second nominated headline.

Calculating word matching score

In this step every word in the nominated headlines will be compared to all words in the document to calculate matching scores. The very basic process of making a matching score between every two words in the nominated headline and document body is to assign a score of 1 if the two words match exactly or 0 if there is even one mismatched character. This basic step is called exact-word matching (EWM). Unfortunately, the Arabic language contains clitics and is morphologically rich. This means that the same word may appear with a single clitic attached to it and yet be considered a different word in the EWM method. Therefore, the idea of using the character cross-correlation (CCC) method emerged, in which a variable score in the range of 0 to 1 is calculated depending on how many characters match each other. For example, if the word {*and he wrote it*, وكتبها } is compared with the word {*he wrote*, كتب} using the EWM method, the resulting score will be 0. However, when using the CCC method, it will be 0.667. The CCC method comes from signal cross-correlation, which measures the similarity of two waveforms. In this method, the score is calculated according to the following equation:

$$CCC_{w_i,w_j} = \frac{2\max_n c[n]}{M+N} \quad (1)$$

And

$$c[n] = \sum_{m=-(N-1)}^{M-1} w_i[m] * w_j[n+m] \quad (2)$$

Where w_i is the first word containing M characters, w_j is the second word containing N characters, and the operation *result is 1 if the two corresponding characters match each other and 0 otherwise.

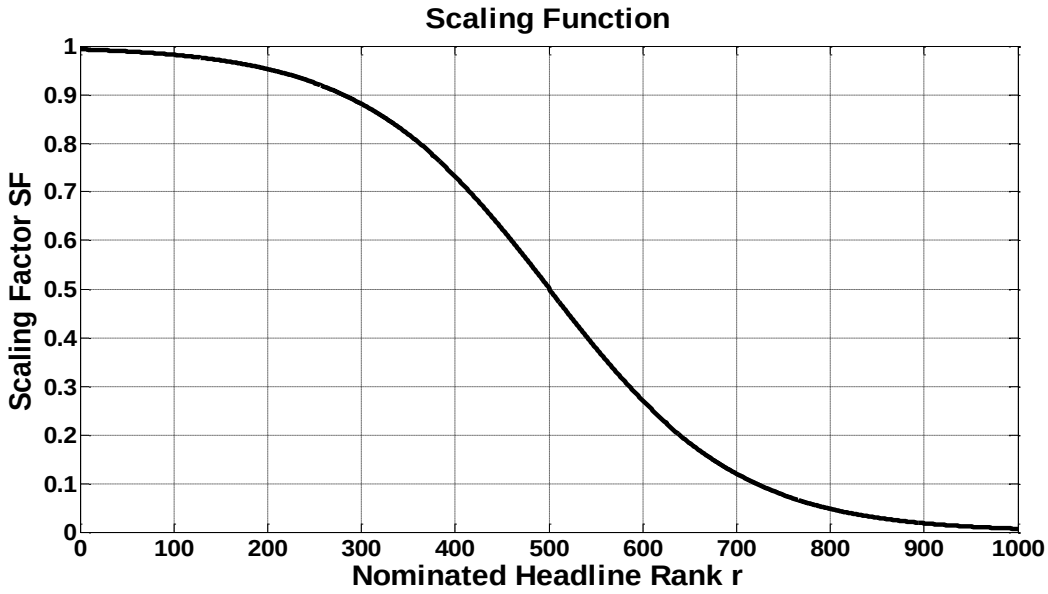

Figure 3. Scaling function of 1,000 nominated headline documents.

Calculating best headline score

After preparing the matching score of the two tables of words, they will be utilized in the selection of the best headline. Except stop words, every word in the document body (w_d) will be matched with every word in the nominated headline (w_h) using the CCC and EWM methods, and a score will be registered for every nominated sentence. A simple stop word list consisting of approximately 180 words was created for this purpose. Calculating a matching score for every sentence is also performed in two ways. The first way is the SUM method, which is defined in the following equation:

$$SUM_p = \sum_{i=1}^{L}\sum_{j=1}^{K} CCCw_d, w_j \quad (3)$$

Where SUM_p is the score using the SUM method for the nominated headline p, K is the number of unique words in the document body, and L is the number of words in the nominated headline (except stop words). In this method, a summation of the cross-correlation score of every word in the document body and every word in the headline is totaled. In a similar way, in the other method MAX_p the maximum score between every word in the document body and the nominated headline is added up. Therefore, for every word in the document, its maximum matching score will be added in either case, CCC or EWM. It can also be defined in the following equation:

$$MAX_p = \sum_{i=1}^{L} \max_j CCC_{w_d, w_j} \quad (4)$$

SUM_p and MAX_p were calculated using *EWM* and *CCC* method, resulting in four different variations of the algorithm, namely SUM-EWM, SUM-CCC, MAX-EWM, and MAX-CCC.

Weighing early nominated headlines

In the case of news articles, the early sentences usually absorb the subject of the article. To reflect that, a nonlinear multiplicative scaling factor was applied. With this scaling factor, late sentences are penalized. The suggested scaling factor is inspired by the hyperbolic tangent function (*tanh*) and described in the following equations:

$$SF = -\left(\frac{e^z - 1}{e^z + 1} - 1\right)/2 \quad (5)$$

Where

$$z = 5\left(\frac{2r}{s} - 1\right), \quad (6)$$

r is the rank of the nominated headline, and S is the total number of sentences. According to the nominating mechanism, hundreds of sentences could be nominated as possible headlines.

Figure shows the scaling function of 1,000 nominated headlines. After applying the scaling factor, the headline

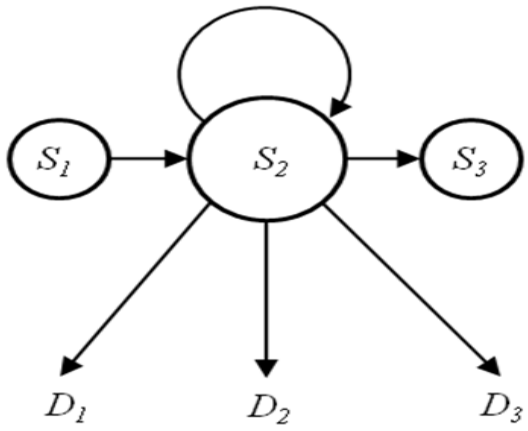

Figure 4. Single state with one entry (S1) and one exit (S3) state HMM used in the AHG system.

with the maximum score was chosen.

HMM-BASED AUTOMATIC HEADLINE GENERATION

HMM-based automatic headline generation systems use one model for the document and up to four features as observation. In the proposed approach, HMM is utilized for every word in the headline with 10 features as the observation vector. The words are also connected together through a bigram probabilistic language model built up from different resources. What follows is a more detailed description of the approach. The document consists of a sequence of words ($D=D_1, D_2 \ldots D_p$), where each document word is represented by a sequence of word observation vector D_t as follows:

$$D_t = d_1, d_2, d_3, \ldots, d_M \tag{7}$$

Where d_m is the document word features observed, so M = 10. The headline consists of a sequence of words $H = h_1, h_2, \ldots, h_n$, and the automatic headline-generation system determines the most probable word sequence H, given the observed document vector D ($argmax_H P(H|D)$).

To do this, Bayes rule is used to decompose the required probability $P(H|D)$ into two components:

$$\begin{aligned} H &= \arg\max_H P(H \mid D) \\ &= \arg\max_H \frac{P(D \mid H)P(H)}{P(D)} \\ &= \arg\max_H P(D \mid H)P(H) \end{aligned} \tag{8}$$

This equation indicates that in order to find the most likely word sequence H, the word sequence that maximizes the product of $P(H)$ and $P(D|H)$ must be found. The first term represents the a priori probability of observing H, independent of the observed document, and this probability is determined by a language model. The second term represents the probability of observing the vector sequence D, given some specified word sequence H. This probability is determined by the HMM. The process of the proposed HMM-based automatic headline generation system for determining (recognizing) headlines is as follows: a word sequence H is *assumed* and the language model computes its probability $P(H)$. For each headline word (h_i), there is a corresponding HMM model. The sequence of HMMs needed to represent the assumed headline is concatenated to form a single composite model, and the probability of that model generating the observed sequence D is calculated. This is the required probability $P(D|H)$. In principle, this process can be repeated for all possible word sequences, and the most likely sequence is selected as the recognizer output.

An important factor to make this approach successful is the assumption that the words appearing in the headline must appear in the document body and in the same sequence, but not as concatenated as in the headline. Every HMM used is a simple HMM consisting of one state besides the entry and exit states, as shown in Figure 4. Single state with one entry (S1) and one exit (S3) state HMM used in the AHG system.

Converting the above design idea into a practical system requires the solution to a number of challenging problems. First, a front-end parameterization (feature extraction) technique is needed, which can extract all the necessary information from the document words in a compact form compatible with the HMM-based statistical model. Secondly, the HMM models must accurately represent the distribution of each headline word. Furthermore, the HMM parameters must be estimated from a sufficient number of samples. Thirdly, the language model must be designed to give accurate word predictions based on the preceding history. However, regarding the HMMs, insufficient data that cover all word sequences is an ever-present problem, and the language model must be able to deal with word sequences for which no examples occur in the training data. Finally, the process outlined above for finding H by enumerating all possible word sequences is impractical. Instead, possible word sequences are explored in parallel, discarding the hypothesis as soon as they become improbable. This process is called decoding.

Feature extraction

One of the most important modules in statistical headline-generation systems is the feature extraction process, in which document words are converted into some type of

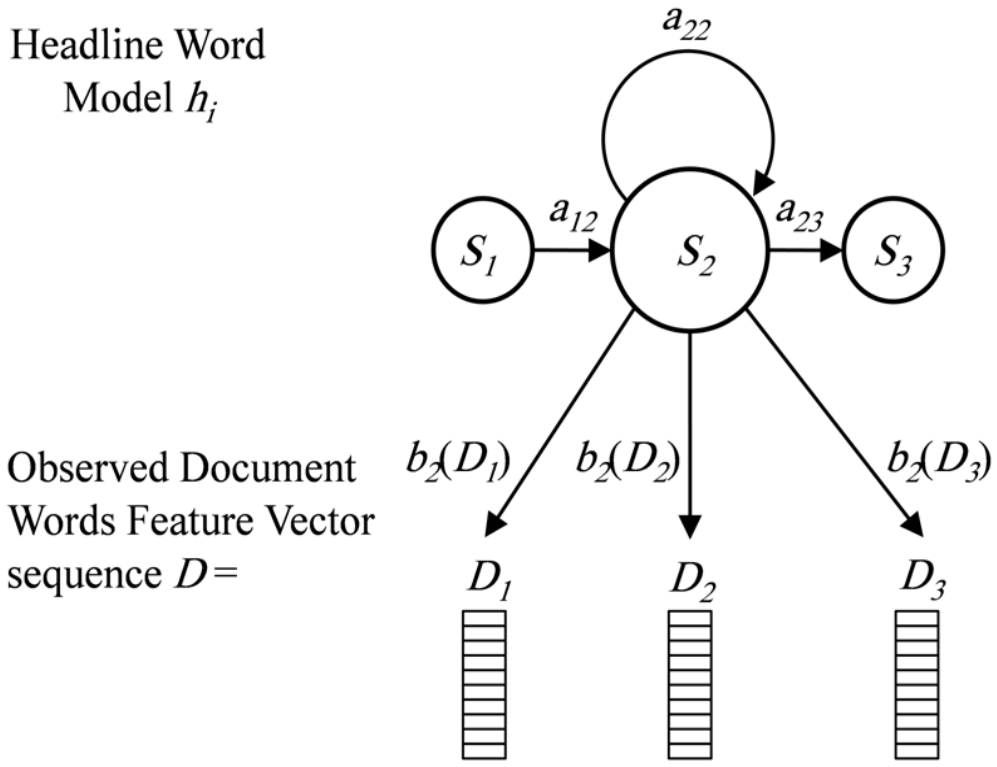

Figure 5. Headline word model used.

parametric representation for further processing. This part is important because the choice of an appropriate feature set influences the accuracy of the headline-generation process. The proposed features of the document words cover a wide range of word characteristics that have a statistical influence. The proposed features are as follows:

1. Position of the word in the current document.
2. Character length of the current word.
3. Word frequency in the current document.
4. Rank of the word in the current document.
5. Is the word a stop-word?
6. Global word frequency throughout the whole corpora.
7. Global rank of the word.
8. Paragraph number in which the word appears.
9. Global frequency of the word with the next word.
10. Global frequency of the word with the previous word.

The global features of any word are difficult to calculate. It took a long time to parameterize each document. Therefore, two methods were proposed to calculate these features. One of them is quick but less accurate, and it was used in a large text to boot up the training and alignment of features. The other one is more accurate, but it requires more time to complete, and it was used for training a smaller set of samples and for the testing phase.

Headline modeling

The modeling unit in the HMM-based automatic headline-generation system is the document words themselves. The purpose of the headline models in this system is to provide a method of calculating the likelihood of any vector sequence D given a headline word h_j. Each individual headline word is represented by a HMM. The HMM has a number of states connected by arcs. It can be regarded as a random generator of a document's word feature vectors. It consists of a main state, entry state, and exit state connected by probabilistic transitions. It changes to a new state for each new headline word, generating a new document's word feature vector according to the output distribution of that state. Therefore, the feedback transition probability models the durational variability in the document word sequence, and the output probabilities model the variability of the features of the document's words. The HMM word model used has one emitting state and a simple left-to-right topology, as illustrated in Figure 5.

The entry and exit states are provided to make it easy to join the models together. This enables words models to be joined together to form complete headlines. Each time t that a state j is entered, a document's word feature

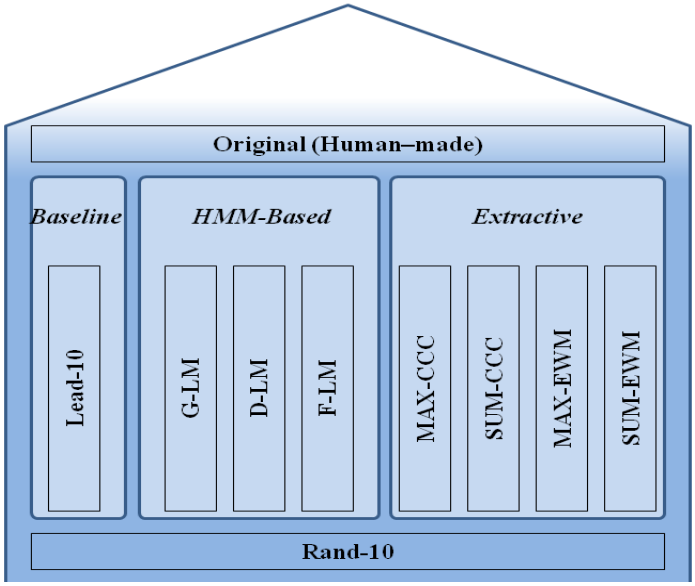

Figure 6. Proposed automatic headline generation methods.

vector D_t (observed at time t) is generated with a probability density $b_j(D_t)$. Furthermore, the transition from state i to j is also probabilistic and governed by the discrete probability a_{ij}. Figure 5 shows an example of this process in which the model moves through the state sequence X = 1, 2, 2, 2 and 4 in order to generate the sequence D_1 to D_3. The joint probability of a vector sequence D and state sequence X given some model (for example h_i, as in Figure 5) is calculated simply as the product of the transition probabilities and the output probabilities. Therefore, for the above state sequence X:

$$P(D, X \mid h_i) = a_{12}\, b_2(D_1)\, a_{22}\, b_2(D_2)\, a_{22}\, b_2(D_3) \quad (9)$$

However, the required probability $P(D|h_i)$ is easily found by summing Equation (9) over all possible state sequences. The determination of the most likely state sequence is the key to generating a headline from an unknown document's word sequence and is computed using the Viterbi algorithm.

Language model

In the proposed approach, $P(H)$ is approximated by a bigram, as shown in the following equation:

$$P(H) = \prod_{i-1}^{m} P(w_i \mid w_1 \ldots w_{i-1}) \approx \prod_{i}^{m} P(w_i \mid w_{i-1}) \quad (10)$$

The bigram language model is used to connect word pairs of the headline according to the probability of that pair. Language models are typically trained on a large corpus of text from the language so that they can obtain robust estimates of their internal parameters. On the other hand, a large and comprehensive corpus could loosely take a broad view of the language and miss important relations in the story for which the headline is generated. To check this assumption, three bigram language models are proposed in this approach. The first one is a general language model (*HMM-G-LM*), which is computed using the entire *Gigaword* corpora. The second is a document-specific language model (*HMM-D-LM*), which is computed using the document for which the headline is generated. The last one is a flat language model (*HMM-F-LM*) with equal probability between corresponding word pairs.

Complete proposed approaches and techniques

In summary, new extractive and abstractive approaches are introduced. In the extractive approach, the most appropriate sentence is extracted from the document using four different techniques: SUM-EWM, MAX-EWM, SUM-CCC and MAX-CCC. In the abstractive approach, three different techniques of HMM-based automatic headline generation are implemented, and they depend on different language models HMM-G-LM, HMM-D-LM and HMM-F-LM.

Figure 6 shows a block diagram of the proposed headlines. It is clear that *Rand-10* is the lowest-limit headline and original i the highest-limit N headline and the extractive and HMM-based techniques will compete against the baseline technique, which is Lead-10.

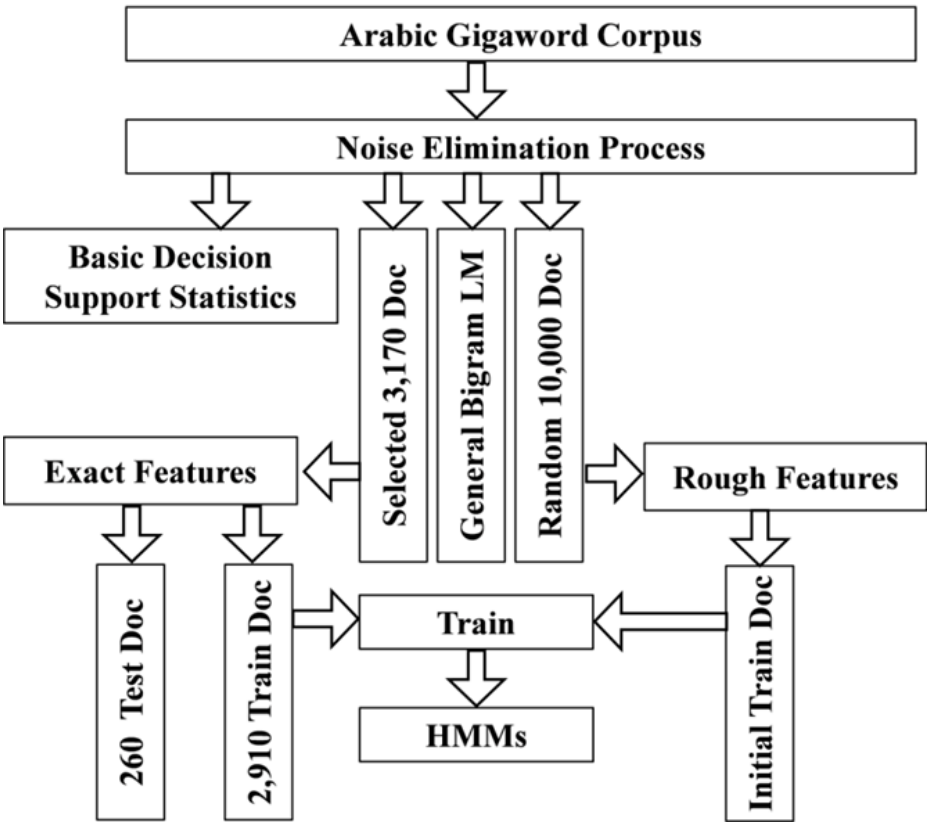

Figure 7. The dataset preparation and HMM training phase of the automatic headline-generation systems.

EXPERIMENTAL DESIGN

The experiment is divided into three main phases: the datasets preparation and training phase, the testing phase and the results evaluation phase. The application used to generate headlines is self-developed software around the Hidden Markov Model Toolkit (HTK) version 3.4 (HTK, 2009). The HTK is a free and portable toolkit for building and manipulating hidden Markov models primarily for speech recognition research. However, HTK has been widely used for other topics such as speech synthesis (Tokuda et al., 2000), character recognition (Khorsheed, 2007), and deoxyribonucleic acid (DNA) sequencing (Grundy, 1997). The main activities performed within the HTK are training, alignment and decoding. In contrast, feature extraction, feature file format conversion, Arabic text transliteration, building HMMs, evaluation and others were developed outside HTK.

Preparing and training phase

The first process in the implementation was the preparation of the *Arabic Gigaword* corpus. It included noise elimination, document investigation and document selection. In noise elimination, simple automatic corrections were applied, as described in datasets. In the document investigation, basic statistics about headlines and corresponding documents were computed to aid in the next phase. In document selection, three datasets were built from the corpus.

The first contained all the documents and was used to generate general statistical bigram and unigram language models. The second contained 10,000 documents and was used to initially train the HMMs with a less-accurate estimate of global features, because global features calculations needed greater processing time. The last dataset contained 3,170 documents with a headline size that varies from 7 to 15 words. This dataset was selected from documents that have informative headlines. Descriptive and eye-catching headlines were avoided. This dataset was divided into a training dataset of 2,910 documents and a test set of 260 documents. All features of the 3,170 documents were accurately estimated. However, two documents of the test set were discarded because of some mistakes in the assessment made by the human examiners. Since the extractive approach does not require training, the resulting 258 test documents were directly processed to generate the four extractive headlines (SUM-EWM, MAX-EWM, SUM-CCC and MAX-CCC).

Figure shows an illustrating block diagram of the dataset preparation and training phase.

Testing phase

In the testing phase, the extractive headline generation system directly generated headlines for the test documents using the four different approaches described earlier. Conversely, the already trained HMMs were utilized to generate different headlines for the test documents using three different language models. The first language model was a general bigram language model (HMM-G-LM) computed from the entire *Arabic Gigaword* corpus. The second

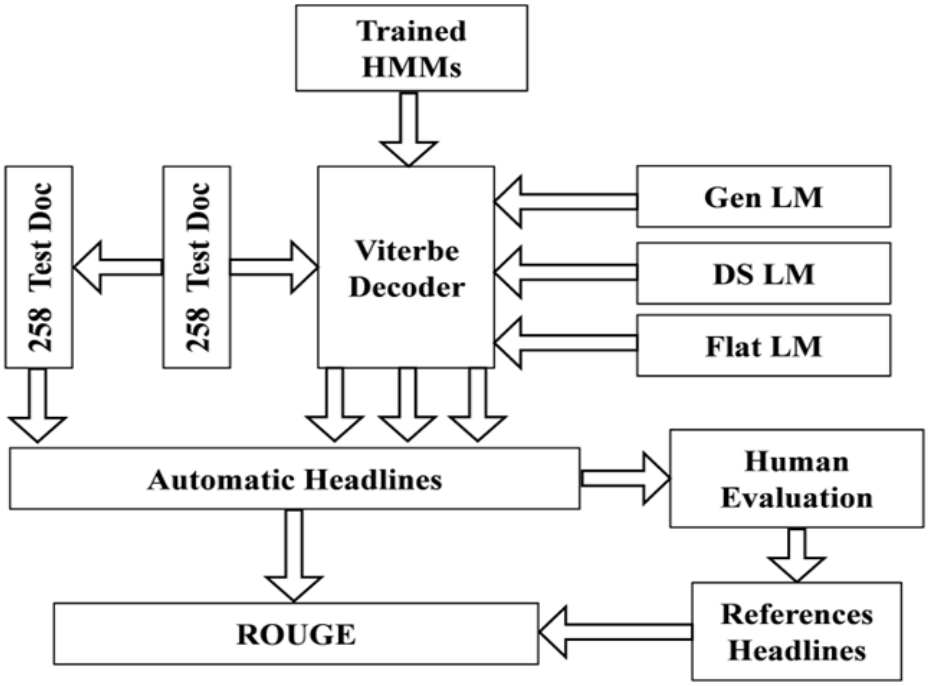

Figure 8. The testing and evaluation phases of the automatic headline-generation systems.

language model was a document-specific bigram language model (HMM-D-LM) computed from the document for which the headline is generated. Finally, the last one was a flat language model (HMM-F-LM) in which the probability of each word pair is the same. Therefore, the contribution of the language model in the *HMM-F-LM* case was almost negligible.

Systems evaluation phase

The evaluation of the resulting headlines was performed using ROUGE version 1.5.5, which generates three scores (recall, precision and F-measure) for each evaluation. Before this version, only one score was generated (recall). For consistency issues, the recall scores will be used. A stemmer is available in ROUGE. The idea of using the stemmer is to compare bare words in the reference headline and the generated headline, ignoring the morphological variation in the words. Unfortunately, the Arabic language is morphologically complex and ROUGE 1.5.5 does not support it. Therefore, the registered scores are expected to be higher if an Arabic stemming is applied in ROUGE 1.5.5. The parameters used in ROUGE 1.5.5 are as follows:

1. Confidence interval is 95%.
2. Computes skip bigram (ROGUE-S) co-occurrence with no gap length limit.
3. Maximum *n*-gram is 4.
4. Uses only the first 10 words in the automatically generated and reference headlines.
5. The rest of the parameters are the default ones.

As no reference system uses the same dataset, the automatic evaluation metric is more suitable for comparing systems rather than assigning an abstract universal score. Thus, three human examiners were hired to evaluate one set of generated headlines. They work in the field of manual document classification. Their task was to examine the readability and consistency of two generated headlines (HMM-D-LM and MAX-CCC) in addition to the original document headline, and generate three headlines (one from each examiner) so that they can be used as references in the ROUGE tool.

Figure 8 shows a brief description of the testing and evaluation phases. As described to the human examiners, the readability score represents the grammatical correctness of the headline despite its meaning, while the consistency score represents how closely the headline reflects the main content of the document, regardless of its syntax. The allowed range of scores varies from 1 to 10. The examiners were told to follow strict instructions to preserve as stable an evaluation as possible. For this purpose, a software tool was specially developed to manually evaluate the headlines. Figure shows the user interface of the manual evaluation tool. The examiner should read the document carefully, suggest an appropriate headline and then evaluate the three headlines. This process should be performed one at a time for each document, and it is advisable to evaluate a large number of headlines in every session to reduce the number of stop periods throughout the entire process.

RESULTS AND DISCUSSION

The evaluation results of the proposed automatic headline-generation systems are presented. The evaluation contains two parts. The first is the automatic evaluation using ROUGE. While the second is the manual evaluation, which was performed by a set of three examiners. The aim of the evaluation results is to compare the proposed approaches against some baseline headlines and the human examiners' evaluation.

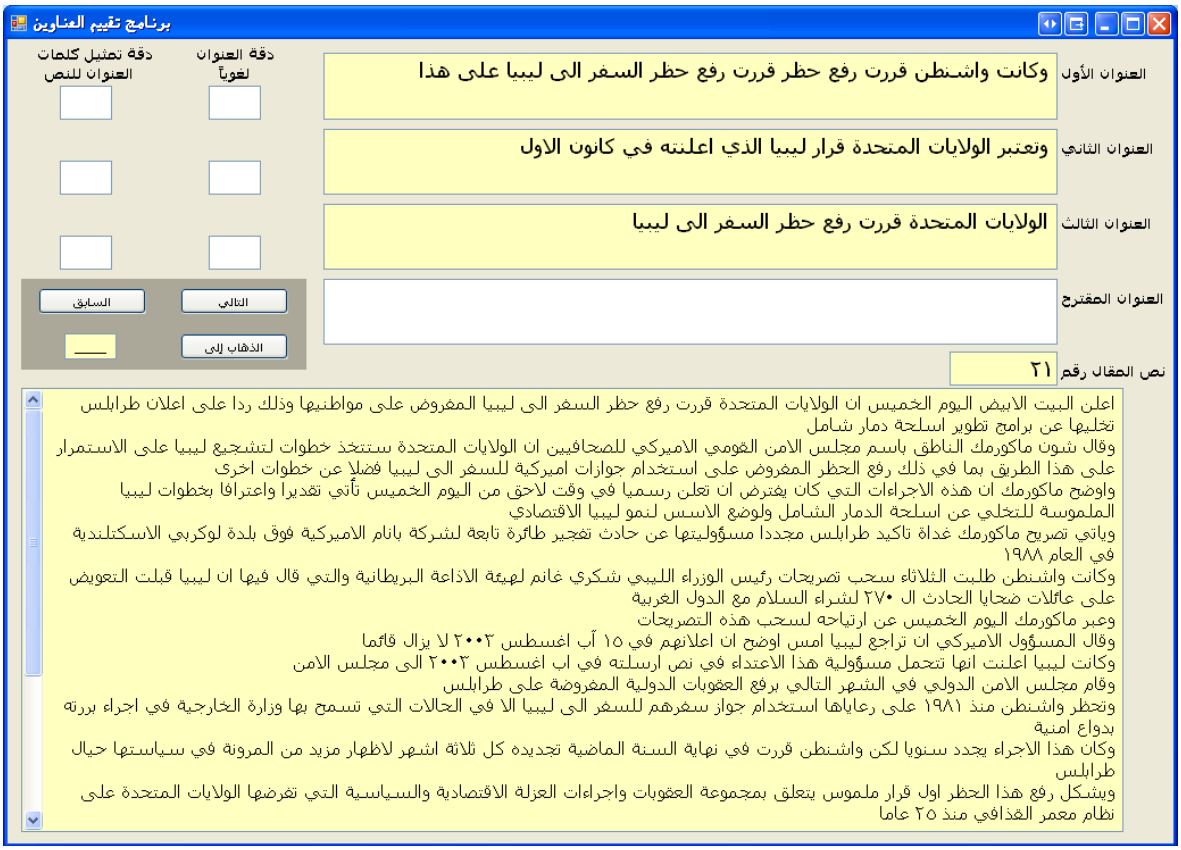

Figure 9. User interface of the manual evaluation tool.

Table 2. ROUGE scores of all headlines.

System	R-1	R-L	R-W	R-SU
Original	0.37683	0.36329	0.21867	0.22498
HMM-D-LM	0.24369	0.2332	0.14689	0.11305
MAX-CCC	0.20367	0.19384	0.12898	0.09001
SUM-CCC	0.18974	0.17944	0.11944	0.08368
Lead-10	0.18353	0.17592	0.11434	0.08761
MAX-EWM	0.18279	0.17252	0.11458	0.07360
HMM-G-LM	0.14184	0.13092	0.08106	0.0423
SUM-EWM	0.11006	0.10624	0.07247	0.04941
HMM-F-LM	0.09428	0.08772	0.05507	0.02193
Rand-10	0.08153	0.07081	0.04491	0.01521

Automatic evaluation

The aim of the evaluation results is to compare the proposed methods against each other and against the results of some baseline headlines. The reference headlines used in ROUGE were the three headlines generated by the human examiners. A total of 10 headlines were used in this evaluation, three of which are the baselines (Original, Lead-10 and Rand-10), three of which are HMM-based (HMM-G-LM, HMM-D-LM and HMM-F-LM), and four of which are the extractive methods (SUM-EWM, MAX-EWM, SUM-CCC and MAX-CCC). Lead-10 can be considered a main baseline, since it produces a meaningful headline with less effort. The other two baseline headlines are introduced to show the highest and lowest score. Although ROUGE-1, ROUGE-L, ROUGE-W-1.2 and ROUGE-SU scores are registered in this section, the ROUGE-L score will be used as a main score for comparison. Table 2 shows the ROUGE scores of the extractive, HMM - based and baselines

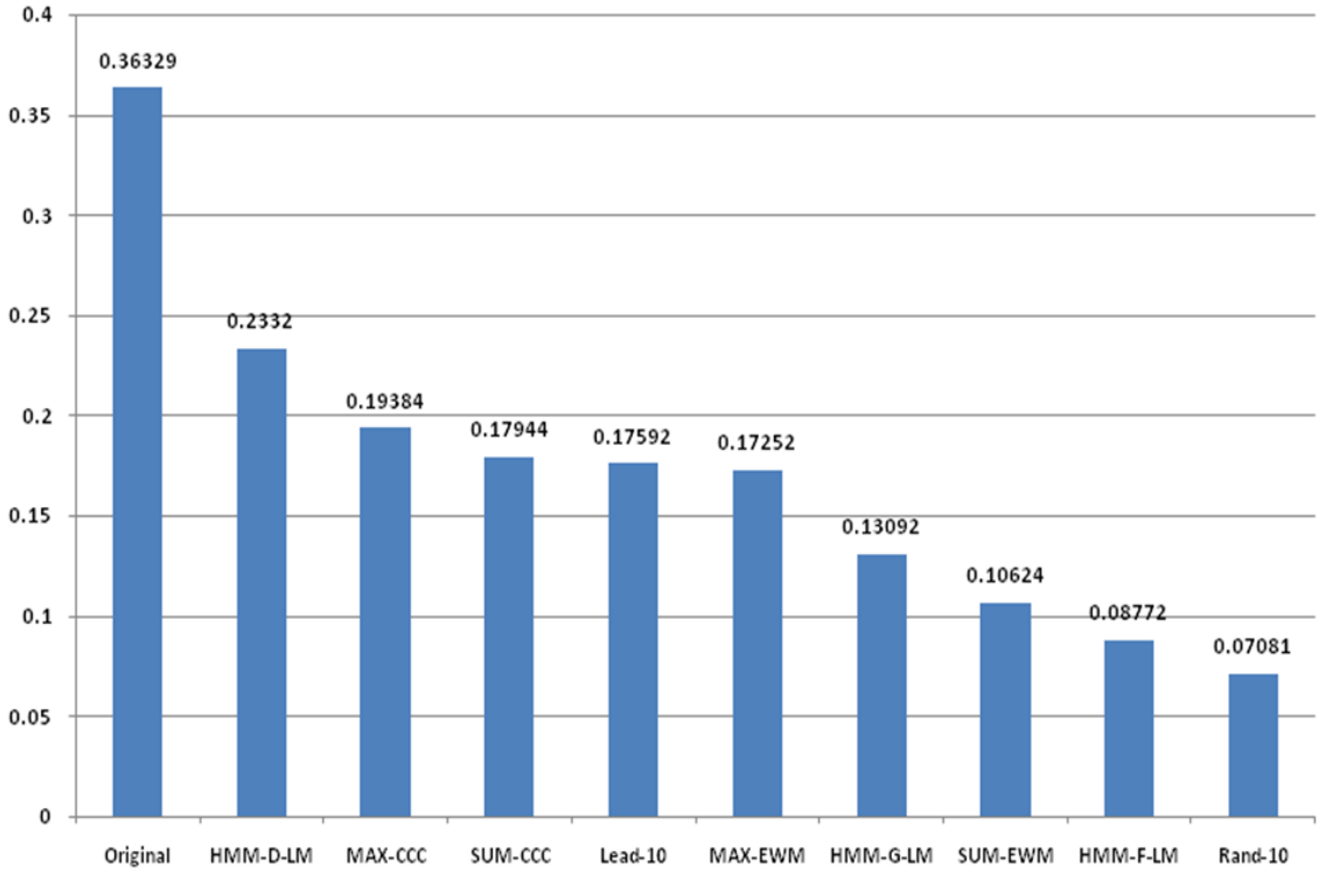

Figure 10. ROUGE-L scores of all headlines.

headlines. On the other hand, Figure 10 shows the ROUGE-L scores of all headlines. From the registered results, the MAX-CCC scores the highest result among the extractive methods. It is clear that MAX-CCC has overcome the problem of the rich existence of clitics and morphology. Character cross-correlation was a valuable procedure in choosing the best headline from the nominated sentences in Arabic documents.

The advantage of using character cross-correlation is that it can overcome the concatenation of clitics to the Arabic words. In this experiment, MAX-CCC produced ROUGE-L = 0.19384 and it outperformed the MAX-EWM, which registered ROUGE-L = 0.17252. Therefore, character cross-correlation can be an effective method for comparing words in morphologically complex languages such as Arabic. As shown in Figure ,

it is obvious that HMM-D-LM is the best automatically generated headline among the systems presented. As predicted, a general language model, one way or another, ignores important relationships in the story and may capture general relationships, but not correct ones for the specific document. In fact, the performance of the system with general language is worse than the Lead-10, which is the baseline system. As predicted, no method registered scores above the original headlines or below Rand-10.

To utilize a language model in an efficient way, a language model scale factor can be applied. The language model scale factor is the amount by which the language model probability is scaled before it is utilized in generating headlines. Unfortunately, the value of the scale factor can be found only by trial-and-error methods. To investigate the effect of the langue model contribution in the HMM-D-LM method, the LM scale factor was varied from 0 (no effect of the language model) to 14, with a step of 1. At each scale factor value, the system is rerun and results are recorded.

Figure m is rerun and results are recorded. Figure shows the change of the ROUGE-L scores for the HMM-D-LM automatic headline-generation system with different values of the LM scale factor. When the LM scale factor is 0, the system is equivalent to the HMM-F-LM because the probabilistic relationships between word pairs were completely ignored.

The best performance was achieved at the LM scale factor of 11 (ROUGE-L = 0.2332). But it can be seen that the system performance settled after the LM scale factor of 6. It is worth mentioning that all results of HMM-based

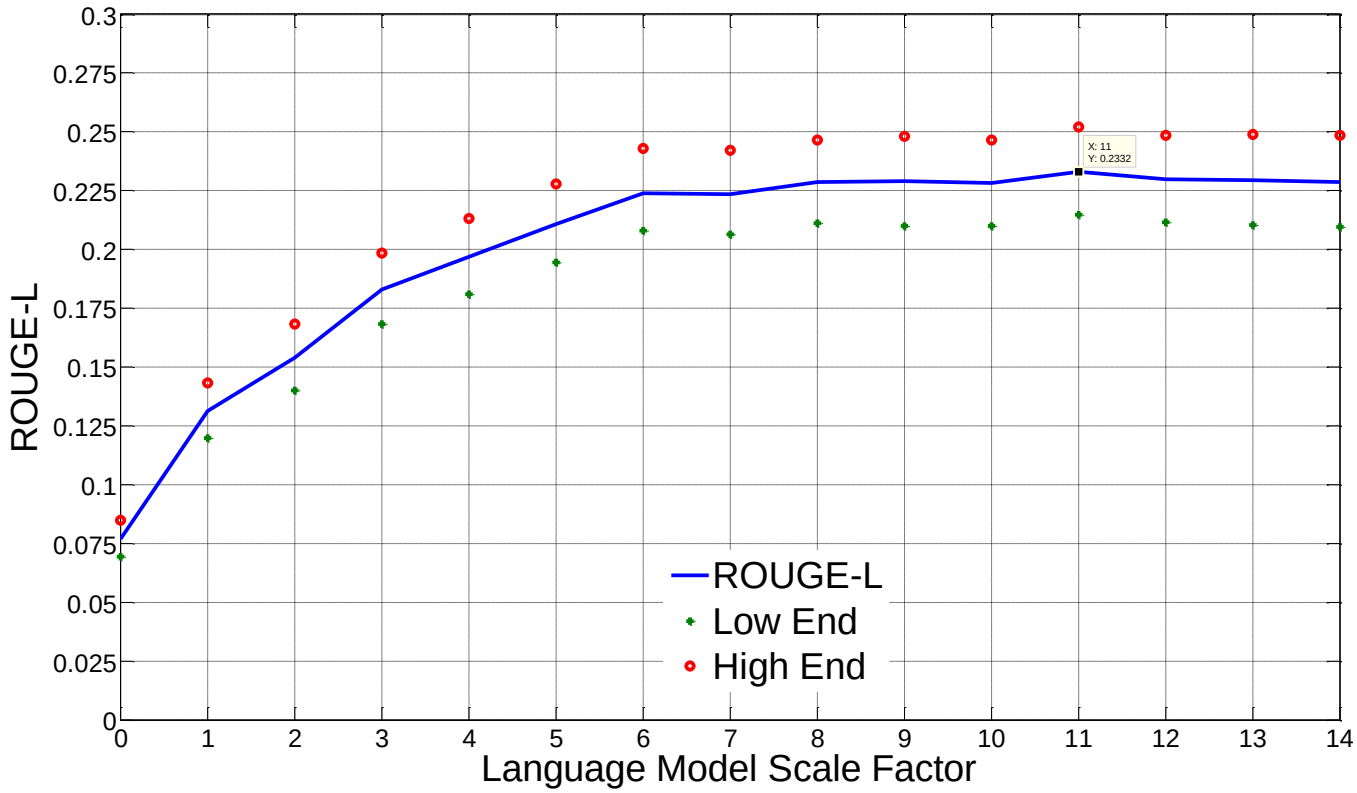

Figure 11. ROUGE-L scores for different values of LM scale factor.

Table 3. Overall evaluation results of the human examiners.

Headline source	Readability (%)	Consistency (%)	Overall (%)
Original	97.90	94.40	96.15
HMM-D-LM	62.20	76.30	69.25
MAX-CCC	77.00	69.20	73.10

systems in Table 2 were generated with a language model scale factor of 11.

Manual evaluation

One of the main criticisms of automatic evaluation metrics is that they do not give a global absolute score. However, they provide a reliable indication when used to compare systems. Therefore, the 258 Original, HMM-D-LM and MAX-CCC headlines were evaluated according to the readability and consistency of the headline. As a result, ROUGE scores become more interpretable. The examiners were asked to assign a score from 1 to 10 for readability and consistency; in which 1 is the lowest score and 10 is the highest. Table 3 shows the overall results of the human evaluation. Few original headlines contained grammatical or spelling mistakes that made them less readable. At the same time, more of them did not perfectly represent the corresponding documents. Obviously, the *HMM-D-LM* headlines were less accurate than the originals, but it is remarkable that their consistency score is higher than their readability score. Since the MAX-CCC headlines were extractive, their readability score is high. The trimming of the MAX-CCC headlines to 10 words is the major factor that reduces their readability. After reviewing the human evaluation in detail, it seems that the examiners could not completely discriminate between readability and consistency. The less readable headline received a lower consistency score, even if it is constructed out of well-represented words.

In this paper, the effectiveness of using character cross-correlation in choosing the best headline from nominated sentences in Arabic documents has been shown. The advantage of using character cross-correlation is that it can overcome the complex morphology of the Arabic language. In the comparative

experiment, character cross-correlation registered ROUGE-L = 0.19384 and outperformed the exact word match, which registered ROUGE-L = 0.17252. Therefore, we can conclude that character cross-correlation is effective when comparing words in morphologically complex languages such as Arabic. Also a new HMM-based approach to automatic headline generation for Arabic news stories was proposed. In this approach, headline words were modeled. In addition, 10 features for every observed word in the document were used as observation vectors. The proposed approach was applied using three different language models. The HMM-based approach with a bigram language model computed from the document for which the headline was generated gave the best score among other automatic systems. The registered ROUGE-L scores were 0.2332 for HMM-D-LM, 0.13092 for HMM-G-LM and 0.08772 for HMM-F-LM. The increase in scores from Flat-LM to Gen-LM to DS-LM shows the strong effect of the language model in building a statistical automatic headline-generation system. Therefore, introducing a higher level of statistical language models than bigram language models may produce a great improvement in the readability of the final headline.

ACKNOWLEDGEMENT

This work has been supported by a direct grant from His Excellency the Rector of King Saud University, Prof. Abdullah Bin Abdulrahman Al-Othman and by a grant from the Research Center, College of Engineering, King Saud University.

REFERENCES

Allen J (1995). Natural Language Understanding. Benjamin/Cummings Pub. Co., Michigan, USA.

Alotaiby F, Alkharashi I, Foda S (2009). Processing large Arabic text corpora: Preliminary analysis and results. In Proceedings of the Second International Conference on Arabic Language Resources and Tools. Cairo, Egypt, pp. 78-82.

Alotaiby F, Foda S, Alkharashi I (2010). Clitics in Arabic language: a statistical study. In Proceedings of Pacific Asia Conference on Language, Information and Computation 24 (PACLIC 24). Sendai, Japan, pp. 595-602.

Buckwalter T (2004a). Issues in Arabic orthography and morphology analysis. In Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages. Geneva, Switzerland.

Buckwalter T (2004b). Buckwalter Arabic morphological analyzer version 2.0, Linguistic Data Consortium (LDC) catalogue number LDC2004L02, ISBN 1-58563-324-0, Philadelphia, USA.

Conroy JM, O'Leary DP (2001). Text summarization via hidden Markov models. In Proceedings of SIGIR 2001. New York, USA, pp. 406-407.

Douzidia F, Lapalme G (2004). Lakhas, an Arabic summarization system. In Proceedings of Document Understanding Conference (DUC), Boston, MA, USA.

DUC Document Understanding Conference (2004). http://duc.nist.gov/duc2004/tasks.html.

Graff D (2007). Arabic Gigaword (3rd ed.). Linguistic Data Consortium. Philadelphia, USA.

Grundy WN (1997). Modeling biological). Linguistic Data Consortium. Philadelphia, USA.

Grundy WN (1997). Modeling biological sequences using HTK. Technical report prepared for Entropic Research Laboratory, Washington, DC.

HTK, Hidden Markov Model Toolkit version 3.4 (2009). http://htk.eng.cam.ac.uk/index.shtml.

Khorsheed M (2007). Offline recognition of omnifont Arabic text using the HMM Toolkit (HTK). Journal of Pattern Recognition Letters, 28(12), Elsevier Science Inc. New York, USA.

Lin CY (2004a). ROUGE: a package for automatic evaluation of summaries. In Proceedings of the Workshop on Text Summarization Branches Out. Barcelona, Spain, pp. 56-60.

Lin CY (2004b). Looking for a few good metrics: ROUGE and its evaluation. In Proceedings of the Fourth NTCIR Workshop on Research in Information Access Technologies, Information Retrieval, Question Answering and Summarization. Tokyo, Japan.

Lloret E, Palomar M (2011). Analyzing the use of word graphs for abstractive text summarization. In Proceedings of The First International Conference on Advances in Information Mining and Management (IMMM 2011), Barcelona, Spain, pp. 61-66.

Manning CD, Schütze H (1999). Foundations of Statistical Natural Language Processing. The MIT Press, Cambridge, MA, USA.

Papineni K, Roukos S, Ward T, Zhu WJ (2002). BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL). Philadelphia, USA.

Reddy PV, Vardhan BV, Govardhan A, Babu MY (2011). Statistical translation based headline generation for Telugu. Int. J. Comput. Sci. Netw. Security, 11(6): 295-299.

Songhua X, Shaohui Y, Francis CM (2010). Keyword extraction and headline generation using novel word features. In Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-10), Georgia, USA, pp. 1461-1466.

Tokuda K, Yoshimura T, Masuko T, Kobayashi T, Kitamura T (2000). Speech parameter generation algorithms for HMM-based speech synthesis. In Proc. ICASSP 2000. Istanbul, Turkey, pp. 1315-1318.

Yang C, Chen H, Honga K (2003). Visualization of large category map for internet browsing. J. Decis. Support Syst., 35(1): 89-102.

Zajic D, Dorr B, Schwartz R (2002). Automatic headline generation for newspaper stories. In Workshop on Automatic Summarization. Philadelphia, USA, pp. 78-85.