International Journal of Computational Linguistics and Applications, Vol. 6, No. 1, 2015, pp. 81–102
Received 17/02/2015, Accepted 14/05/2015, Final 18/06/2015.
ISSN 0976-0962, http://ijcla.bahripublications.com

Extracting Sentences Using Lexical Cohesion for Arabic Text Summarization

HAMZA ZIDOUM
AHMED AL-MAAMARI
NASSER AL-AMRI
AHMED AL-YAHYAI
SAID AL-RAMADHANI
Sultan Qaboos University, Muscat, Sultanate of Oman

ABSTRACT

Automatic Text Summarization has received a great deal of attention in the past couple of decades. It has gained a lot of interest especially with the proliferation of the Internet and the new technologies. Arabic as a language still lacks research in the field of Information Retrieval. In this paper, we explore lexical cohesion using lexical chains for an extractive summarization system for Arabic documents.

INTRODUCTION

Summary as defined by [21] is a "text that is produced from one or more texts, that conveys important information in the original text(s), and that is no longer than half of the original text(s) and usually significantly less than that." Summarization dates back to the late fifties where the first attempts relied entirely on statistical approaches. The sentence consisting of words with a high frequency were given a higher weight than the others indicating the importance of these sentences. Other than the mentioned approach, many different approaches were devised to tackle the problem of summarization [4], [16]. Cue phrases and lead method are one of the many approaches, the former extracts sentences containing words or phrases for e.g., "significant," "In this paper” etc. The latter extracts first sentences of paragraphs assuming they contain the main idea. These methods rely on shallow approaches to indicate the importance of sentences to be included in the summary. Other approaches look at deeper levels like similarity that occurs when two words share a common stem, as in for instance the *thesaural* relationships that identify the different semantic relations existing between words, or the Rhetorical Structure Theory which identifies the relationship between text units.

There have been a few studies done on summarizing Arabic documents. Lakhas [6] attempted generating summary using a hybrid approach. The developed system relied on shallow approaches; frequency calculation, indicative expressions (cue words), lead method and title method. The system was evaluated in DUC 2004 (Document Understanding Conference). Systems that produce user focused summaries such as the one developed by [10] generates query based and concept based summaries. Arabic Query Based Text Summarization System (AQBTSS) is a query based single document summarizer that generates summaries relevant to the user query. Each sentence in the document is compared against the user query and only the relevant sentences are extracted. The other system Arabic Concept Based Text Summarization System (ACBTSS) is a concept based summarizer that generates a summary by matching each sentence against a set of keywords entered by the user, and these words represent a specific concept. The system uses a vector space model (VSM) that makes use of two measures; term frequency (tf) and inverse document frequency (idf) to weighing sentences.

A different approach by [12] was devised using clustering techniques. In this technique the roots are extracted for each word and are placed in the appropriate cluster. The words are assigned weights based on the number of words in the cluster it belongs. In addition to that it makes use of cue words which can enhance the weight of the sentence. The system then extracts sentences with highest scores, and the number of sentences depends on the size of the document.

Summarization can be described as a two-step process: (1) Building a source representation from the original text, and (2) Converting the source representation to an intermediate representation.

The input to the summarization systems can be in the form of textual data or other types of multimedia such as audio, video or images. Furthermore, summarization can even be performed on single documents or multiple documents consisting of a single language or more than one language also called multi-lingual. Output of the summarization system can be categorized into two groups: extracts and abstracts. Extract summaries consist of sentences from the original document whereas abstract summaries paraphrase some sections of the text or formed from the generated sentences. This requires language generation techniques and has some challenges.

Presenting a user with an adequate summary requires capturing the main theme of the document. This can be accomplished by looking at the related words in the document. Lexical cohesion can be created by semantically related words and represented by lexical chains. Lexical chains groups together semantically related words.

In comparison to English, and despite the recent interest due to geopolitic issues, Arabic still lacks research in the field of Information Retrieval. The factors contributing to this challenge of automatic processing of Arabic is the Arabic script itself due to the lack of dedicated letters to represent short vowels, changes in the form of the letter depending on its place in the word, and the absence of capitalization and minimal punctuation, Arabic words can be ambiguous as diacritics have been disappearing in contemporary writings [11]. Another factor is the normalization due to the inconsistency in the use of diacritic marks and certain letters in contemporary Arabic texts. Some Arabic letters share the same shape and are only differentiated by adding certain marks such as a dot, a hamza or a madda placed above or below the letter. For example, the “alif” in Arabic (ا) may be three different letters depending on whether it has a hamza above as in (أ) or a Hamza below as in (إ) or a madda above as in (آ). Recognizing these marks above or below a letter is essential to be able to distinguish between apparently similar letters. But texts written in MSA often do not incorporate voweling as mentioned earlier nor do they adhere to the “proper” inclusion of marks above or beneath some Arabic letters. To manage this problem, the common practice in Arabic NLP systems is to normalize the input text [14]. For example, in order to handle the different variations in Arabic script, Larkeyand Connell (2001) replace the initial alif with a hamza above or below withsimply an alif, or bare alif. They also normalize the alif madda with a bare alif.Further, they normalize the final taa marbuuTa (ةor ىة) with a final haa (هorىه) and the alif maqsuura (ى) with the yaa (ي).

In this paper we implement an algorithm that generates an extractive summary for Arabic single document using lexical chains. It makes use of WordNet [19], a lexical resource database containing nouns, verbs, adverbs and adjectives and groups each of them into set of synonyms called *synsets*. These synsets are related to others via semantic relations. The system will also make re-use of a parser and part-of-speech tagger to identify nouns.

The rest of the paper is organized as follows: Section 2 discusses previous works on lexical chains especially in the context of Summarization. Section 3 gives an overview of lexical cohesion and lexical chains. Section 4 presents the system architecture. Finaly section 5 is dedicated to discussing the results.

2. RELATED WORKS

Lexical chains was first introduced by [20]. They did not implement this algorithm as a machine-readable dictionary was not available then.

Barzilay & Elhedad [2] attempted to implement this algorithm using Wordnet. They performed segmentation on the source text and built lexical chains in every segment. All the interpretations were computed for every sense of the word to determine the correct one. This gave it an exponential runtime complexity even for short documents.

[22] implemented a linear time algorithm to compute lexical chains. The computation was done by creating a structure to store the interpretations without actually creating them. This two phases algorithm built an internal representation in the first phase and computed the chain score and performed Word Sense Disambiguation (WSD) in the second phase.

This algorithm was not yet accurate enough when performing word sense disambiguation as claimed by [8]. The algorithm divided the computation of lexical chains into three steps as opposed to two done by Silber and McCoy. In the first step, similar to Siber and McCoy they built a representation of all possible interpretations. This was the only time they made a pass through the document. The latter steps depended on this representation. The following step performed disambiguation where the word was assigned a correct sense based on sum of the weights leaving all the senses of the word, and this depends on the type of relation and the distance factor. The final step is the actual building of the lexical chains where the word whose sense is different than the assigned sense is removed. This algorithm was compared against Barzilay et al. [2] and Silber et al. [22] algorithm and performed better in terms of word sense disambiguation.

Medelyan [18] came up with a new approach to compute lexical chains with graph clustering. Lexical chains are treated as graphs where the nodes are the terms and the edges represent the relations. The chains cohesive strength is computed by the diameter of the graph; which can be defined as the longest of the shortest distance between any two nodes in the graph. This diameter is strong if it is fully connected where each node is connected to all the others, weak or moderately cohesive. The weak cohesive chains are decomposed into several highly cohesive chains by using a graph clustering algorithm. After strong chains are identified by using the chain scores, sentences are retrieved by summing up the weights of all the words in the sentence which correspond to the chain weights in which these words occur.

The previous methods assume that each chain represents a specific topic, but [7] claimed that a single chain cannot represent a whole topic. A cluster of lexical chains might represent a specific topic where they could define the "what," "where" and "why." They computed lexical chains using Galley et al. algorithm and filtered out the weak chains that had a score lower than a defined threshold. The remaining chains represent the strong ones and are clustered using cosine similarity. The connected sequences of sentences in these clusters are identified as segments and are scored. The best scoring segments are retrieved and the first sentence of each segment is included in the summary. The number of sentences to be retrieved in the summary depends on the number of unique sentences in the best scoring segments as only one sentence per segment is selected so this number could be less than the number of sentences required in the summary.

Cohesion introduced by Halliday and Hasan (1976) is a device used to *glue* together different parts of the text by means of cohesive relations, thus giving it a sense of unity. These relations can be classified into *co-reference*, *ellipsis*, *conjunction* and *semantic word relations*. Lexical cohesion can be created by semantically related words and is one of the most identifiable types of cohesion. Halliday and Hasan classified lexical cohesion into *reiteration* and *collocation*.

Reiteration can be achieved by *repetitions*, *superordinates*, *subordinates* and *synonyms*. *Collocation* defines semantic relations between words that often tend to occur in similar lexical contents (e.g., "The girl was digging the garden"). *Collocation* as a relation is more problematic to identify than reiteration.

Lexical cohesion does not occur between a pair of words but over a number of related words spanning a topical unit of text. This grouping of words is called lexical chains. To identify a relationship between words we make use of Arabic Wordnet.

Wordnet is a lexical resource database consisting of syntactic categories such as nouns, verb, adjective and adverbs. It groups words in each syntactic category into a set of synonyms called *synsets*. Furthermore, the synsets can be related to others in terms of semantic relations. Wordnet identifies the following relations:

synonyms are the basic and most important relation in Wordnet. Words are synonymous if they share the same sense and can be used interchangeably in the text. *Hypernyms/Hyponyms* also known as superordinate and subordinate. *Antonyms*(opposing name) are mainly described for adjectives and adverbs and *meronym/holonym* known as sub-part/whole part.

3. System Architecture and Implementation

Automatic extractive summarization is a complex task that involves several challenging subtasks. The performance in each of these subtasks affects the performance for generating high quality summaries. First, Aramorph [3] proved to be faster than Stanford Parser [17]. Moreover, Aramorph works as Part of Speech tagger as well as word stemmer. The Aramorph module offers a better API which facilitates the integration. Diacritics add difficulties in comparing words and identifying their relations using Arabic Wordnet, so we used "Diacritic Remover" (reference) module to remove diacritics. Also we used "Arabic stem analyzer" to extract the stems. In general there is a lack of electronic lexical resources for Arabic, for example Arabic Wordnet is not as rich as the English counterpart for Wordnet, and the only available JAVA API for the Arabic Wordnet is is written by Abobakr Ahmed, available from sourceforge: <http://sourceforge.net/projects/javasourcecodeapiarabicwordnet/>.

The design of the system can be summarized in the following steps as illustrated in Figure 2 below. Each module takes as input the file produced by the previous module starting with the Tokenization module which takes a single Arabic text file.

1. Tokenization: Process where each sentence is partitioned into a list of tokens. Before the tokenization process we break the text into sentences. We adopted the Stanford document pre-processor module.
2. Part of Speech Tagging: It consists of classification of the tokens according to the best part of speech they represent; noun, verb, adverb, etc. Arabic stem analyzer which is called

Aramorph is used in this step. Aramorph PoS tagger produces the standard tag set as well as the extended tag set, so a noun could be assigned a tag set NN indicating a noun, ADJ indicating Adjective or NNS indicating a singular noun. Aramorph then returns the stem of the original word. The outputs are the stemmed words with diacritics.

3. Noun Filtering and Normalization: Nouns need to be filtered out prior to computing lexical chains. Regular Expressions have been used to accomplish this task by identifying the tags assigned as nouns by the toolkit then, we used "Arabic normalizer" to normalize stems as follows:

    - Normalization of *hamzated alif* to a bare alif.
    - Normalization of taa marbutah to haa.
    - Normalization of dotless yaa to yaa.
    - Removal of tatweel (stretching character).

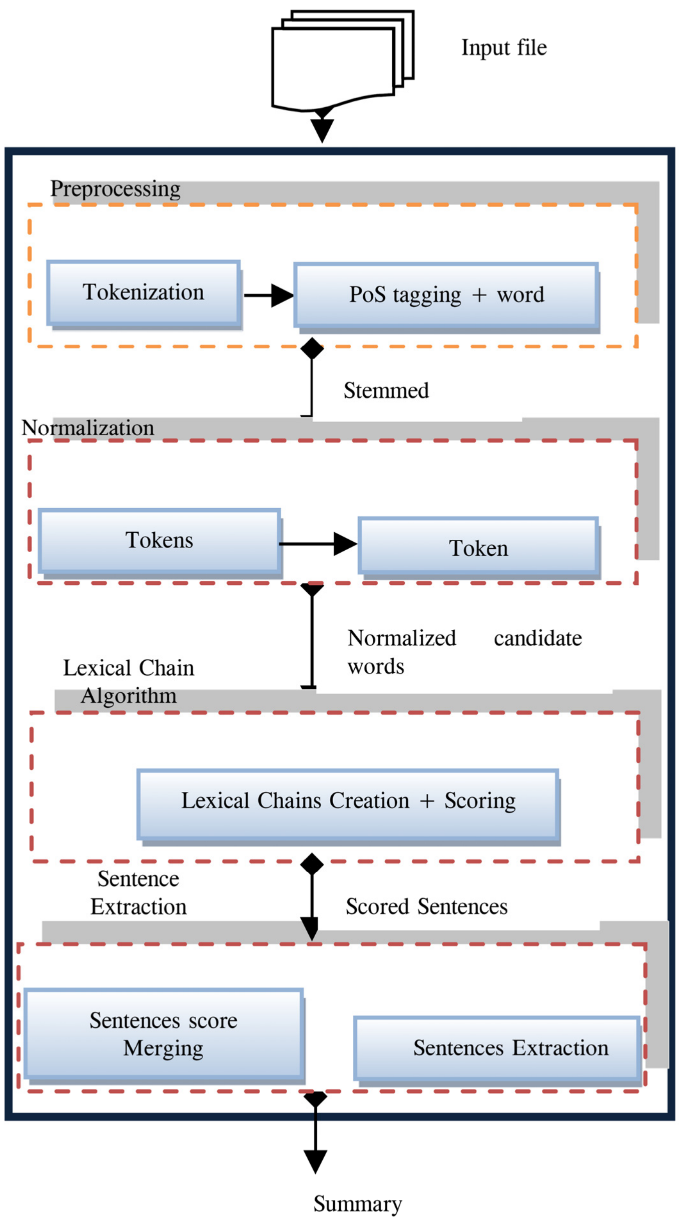
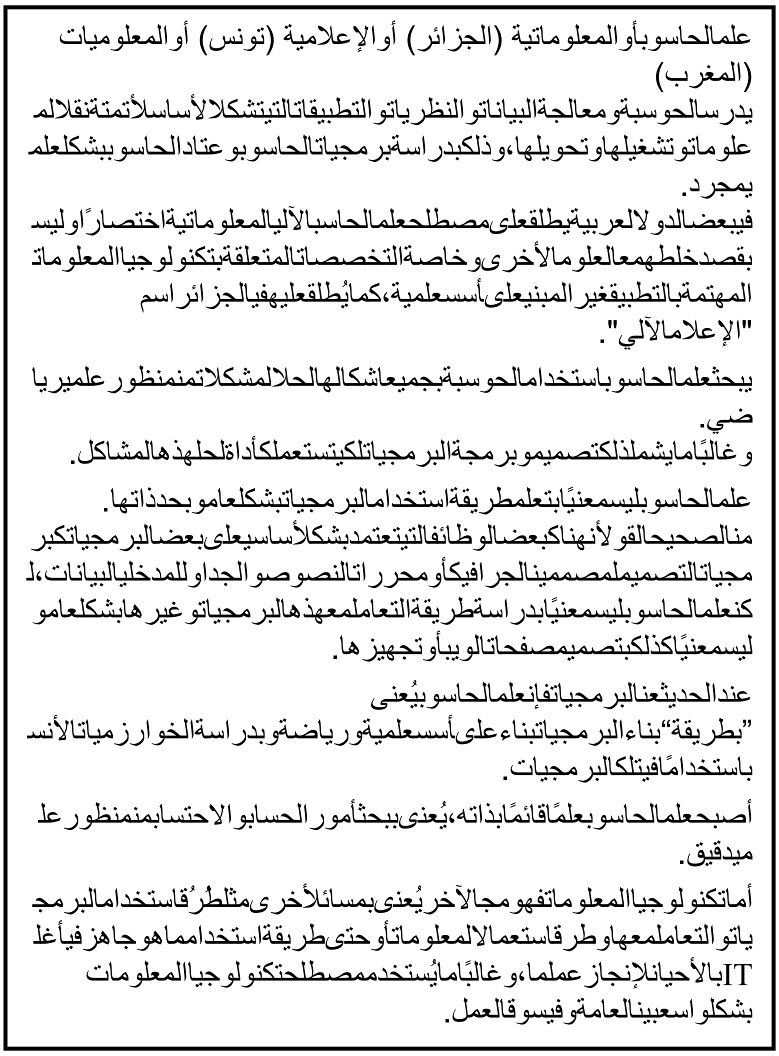


Figure 1. *System Architecture*



علمالحاسوبأو المعلوماتية (الجزائر) أوالإعلامية (تونس) أوالمعلوميات (المغرب)
يدرسالحوسبةومعالجةالبياناتوالنظرياتوالتطبيقاتالتيتشكلالأساسلأتمتةنقلالم
علوماتوتشغيلهاوتحويلها،وذلكبدراسةبرمجياتالحاسوبوعتادالحاسوببشكلعلم
يمجرد.
فيبعضالدولالعربيةيطلقعلىمصطلحعلمالحاسبالآليالمعلوماتيةاختصارًاوليس
بقصدخلطهمعالعلومالأخرىوخاصةالتخصصاتالمتعلقةبتكنولوجياالمعلومات
المهتمةبالتطبيقغيرالمبنيعلىأسسعلمية،كمايُطلقعليهفيالجزائراسم
"الإعلامالآلي".

يبحثعلمالحاسوباستخدامالحوسبةبجميعاشكالهالحلالمشكلاتمنمنظورعلميريا
ضي.
وغالبًامايشملذلكتصميموبرمجةالبرمجياتلكيتستعملكأداةلحلهذهالمشاكل.

علمالحاسوبليسمعنيًابتعلمطريقةاستخدامالبرمجياتبشكلعاموبحدذاتها.
منالصحيحالقولأنهناكبعضالوظائفالتيتعتمدبشكلأساسيعلىبعضالبرمجياتكبر
مجياتالتصميملمصممينالجرافيكأومحرراتالنصوصوالجداولللمدخليالبيانات،ل
كنعلمالحاسوبليسمعنيًابدراسةطريقةالتعاملمعهذهالبرمجياتوغيرهابشكلعامو
ليسمعنيًاكذلكبتصميمصفحاتالويبأوتجهيزها.

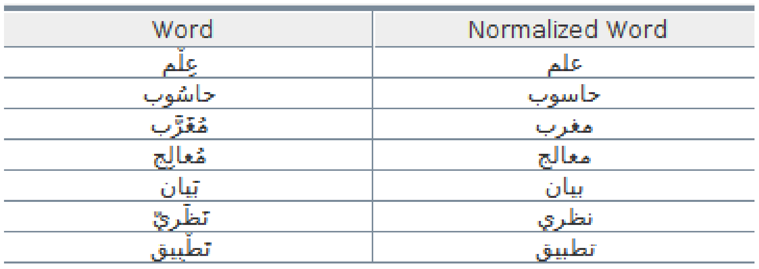
عندالحديثعنالبرمجياتفإنعلمالحاسوبيُعنى
"بطريقة"بناءالبرمجياتبناءعلىأسسعلميةورياضةوبدراسةالخوارزمياتالأنسب
باستخدامًافيتلكالبرمجيات.

أصبحعلمالحاسوبعلمًاقائمًابذاته،يُعنىببحثأمورالحسابوالاحتسابمنمنظورعل
ميدقيق.

أماتكنولوجياالمعلوماتفهومجالآخريُعنىبمسائلأخرىمثلطُرُقاستخدامالبرمج
ياتوالتعاملمعهاوطرقاستعمالالمعلوماتأوحتىطريقةاستخداممماهوجاهزفيأغل
ITبالأحيانلإنجازعملما،وغالبًامايُستخدممصطلحتكنولوجياالمعلومات
بشكلواسعبينالعامةوفيسوقالعمل.

Figure 2. *Arabic text input sample*

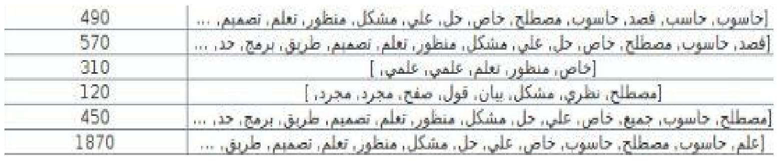
We used “*DiacriticRemover*” to remove Diacritic of the words (1) remove excessive whitespace sequences in the Arabic text, and (2) remove all diacritic marks like (TANWEEN_ALFATHA, TANWEEN_ALDAMMA, TANWEEN_ALKASRA, FATHA, DAMMA, KASRA, SHADDA, SUKUN). The Output is the Normalized candidate words.



Word	Normalized Word
عِلّم	علم
حاسُوب	حاسوب
مُغَرَّب	مغرب
مُعالج	معالج
بَيان	بيان
نَظَرِيّ	نظري
تَطْبيق	تطبيق

3.1. *Lexical chain computation*

For the first step, we first process all the sentences to produce a chain list. For each candidate word in each sentence, we try to add the candidate word to a chain. If the candidate word hasalready been added to a chain, we increase the chain weight by adding *weight1*. This number represents a strong relation and a repetition of the word. Then we add the sentence id of the candidate word to the chain and update the score of the sentence inside the chain. In case the word is not added to a chain, we create a new chain. After creating the new chain, we generate a sense list of the word by using Arabic WordNet (AWN) and add this list to the newly created chain and we mark the word as a used word by adding it to a list named used words. Finally, if the chain sense list is not empty, we add the chain to our chain list. Otherwise, we ignore the chain, since it does not contain any sense list and it cannot create any relation with other chains. Examples of lexical chains are given in Figure 2.



490	[حاسوب، حاسب، قصد، حاسوب، مصطلح، خاص، حل، علي، مشكل، منظور، تعلم، تصميم، ...
570	[قصد، حاسوب، مصطلح، خاص، حل، علي، مشكل، منظور، تعلم، تصميم، طريق، برمج، حد، ...
310	[خاص، منظور، تعلم، علمي، علمي،]
120	[مصطلح، نظري، مشكل، بيان، قول، صفح، مجرد، مجرد،]
450	[مصطلح، حاسوب، جميع، خاص، علي، حل، مشكل، منظور، تعلم، تصميم، طريق، برمج، حد، ...
1870	[علم، حاسوب، مصطلح، حاسوب، خاص، علي، حل، مشكل، منظور، تعلم، تصميم، طريق، ...

Figure 3. *Lexical chains and their weights*

After constructing the chains, we try to find relations between them. For each chain, we create a new thread to create a relation list of the chain with the other chains. We have two kinds of relations with different weight, IN-relation and OUT-relation. To simplify these relations, let us say that we have two words: “word1” and “word2.” If we could find a relation between the “word1” and one of the senses of “word2,” than this will represent an IN-relation for “word1” and an OUT-relation for “word2.” The ID of the chain that contains the “word1” will be added to the chain of the “word2” and vice versa. The IN-relation adds *weight2* to the weight and it is considered as a medium relation. The OUT-relation is considered as a weak relation and it adds *weight3* to the weight. Notice that *weight1> weight2> weight3*.

```
for each sentence { //First step
        for each word {
                if (word is repeated)   // already in UsedWord list
                {
                        add weight1 to chain weight; // weight1for repetition
                        add sentence id to chain;
                        update sentence score;
                }
                else {
                        Create new chain for the candidate word;
                        Generate a senseList taking all relations from Wordnet;
                        Add word to UsedWord list;
                        if (new chain senseList is not empty)
                                Add chain to chainList;
                }
        } // end words loop
} // end sentences loop
for each chain { // Second step
        Inside a new thread {
                if (chain word has a relation with other chain senseList) {
                Link chain with other chain;
                        add weight2 to chain weight; // IN relation
                        add weight2 to the other chain weight; // OUT relation
                } // end if
        }  // end thread
 }  // end chains loop
```

3.2. *Strong chains identification and extraction*

By this stage, each chain contains a chain number, a list of sentence numbers and weight. The algorithm for the sentences scoring and extracting is in Figure 6. Before extraction, we calculate sentences score by adding a fraction of chain weight to each sentence in the chain. The fraction is to take an equal fixed part of the chains weight instead of taking the whole weight which sometimes can be large numbers.

After adding the final scores of sentences inside each chain, we reach the extraction stage. In this stage we have two main steps: (1) Merging the sentences from all chains with their scores to be extracted, and (2) Sentences with highest score will be extracted for the user depending on the extraction rate that he needs. The number of the sentences which will be extracted will be counted as follows:

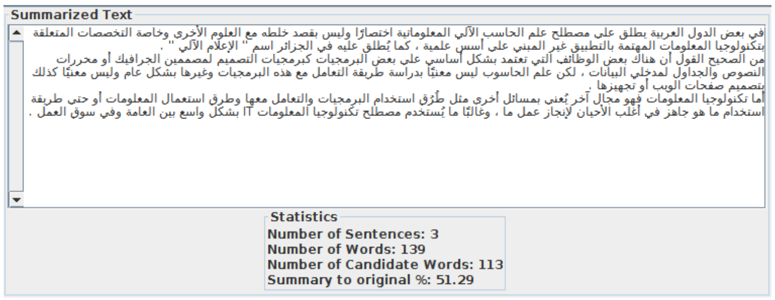
```
Add 3% of each chain weight to score of sentences;
Create table for sentences and their scores;
For each chain {
    For each sentence {
        If (sentence in table)
            Increment sentence score in the table;
        Else
        Add new record to the table;
    }
}
Sort table by the score;
Extract of sentences according to rating;
```

Number of extracted sentences =
*The ceil of (Extraction rate * Number of sentences in original text)*

For example, if the user needs 10% of extraction and the original text includes 22 sentences. The number of the sentences which will be extracted is:

Number of extracted sentences = 0.10 22 = ceil (2.2) =3 sentences.*
(Note: we used the ceil to prevent losing part of sentences)

The output is a combination of sentences with high score in the correct sequence which represent the final summary.

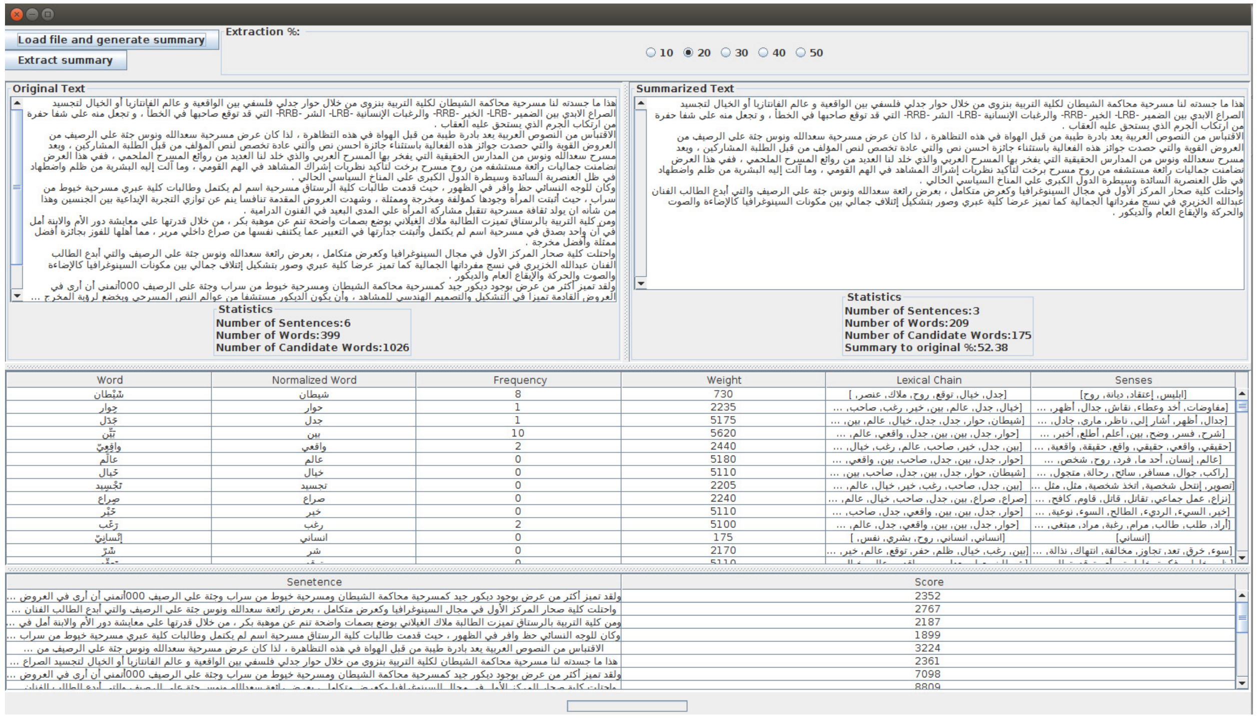


في بعض الدول العربية يطلق على مصطلح علم الحاسب الآلي المعلوماتية اختصارًا وليس بقصد خلطه مع العلوم الأخرى وخاصة التخصصات المتعلقة بتكنولوجيا المعلومات المهتمة بالتطبيق غير المبني على أسس علمية ، كما يُطلق عليه في الجزائر اسم " الإعلام الآلي " .
من الصحيح القول أن هناك بعض الوظائف التي تعتمد بشكل أساسي على بعض البرمجيات كبرمجيات التصميم لمصممين الجرافيك أو محررات النصوص والجداول لمدخلي البيانات ، لكن علم الحاسوب ليس معنيًا بدراسة طريقة التعامل مع هذه البرمجيات وغيرها بشكل عام وليس معنيًا كذلك بتصميم صفحات الويب أو تجهيزها .
أما تكنولوجيا المعلومات فهو مجال آخر يُعني بمسائل أخرى مثل طُرُق استخدام البرمجيات والتعامل معها وطرق استعمال المعلومات أو حتى طريقة استخدام ما هو جاهز في أغلب الأحيان لإنجاز عمل ما ، وغالبًا ما يُستخدم مصطلح تكنولوجيا المعلومات IT بشكل واسع بين العامة وفي سوق العمل .

Number of Sentences: 3
Number of Words: 139
Number of Candidate Words: 113
Summary to original %: 51.29

Figure 4. *Extracted Sentences*

The system graphical user interface is given in Figure 4, where the most prominent items are, for instance, loading a text file for summarization, original text, extracting the summary, e**xtraction rate in percentage,** displaying the generated summery, o**riginal text statistics (n**umber of sentences, number of words, number of candidate words), **summarized text statistics,** candidates words from the original text, n**ormalized words, their frequency,** senses from Arabic wordnet**, and the** weight of candidate words' lexical chains from the original text, lexical chains, candidate word's, sentences from original text and their weight**,** the score of sentence from the original text.



Original Text

هذا ما جسدته لنا مسرحية محاكمة الشيطان لكلية التربية بنزوى من خلال حوار جدلي فلسفي بين الواقعية و عالم الفانتازيا أو الخيال لتجسيد
الصراع الابدي بين الضمير -LRB- الخير -RRB- والرغبات الإنسانية -LRB- الشر -RRB- التي قد توقع صاحبها في الخطأ ، و تجعل منه على شفا حفرة
من ارتكاب الجرم الذي يستحق عليه العقاب .
الاقتباس من النصوص العربية يعد بادرة طيبة من قبل الهواة في هذه التظاهرة ، لذا كان عرض مسرحية سعدالله ونوس جثة علي الرصيف من
العروض القوية والتي حصدت جوائز هذه الفعالية باستثناء جائزة احسن نص والتي عادة تخصص لنص المؤلف من قبل الطلبة المشاركين ، وبعد
مسرح سعدالله ونوس من المدارس الحقيقية التي يفخر بها المسرح العربي والذي خلد لنا العديد من روائع المسرح الملحمي ، ففي هذا العرض
تضامنت جماليات رائعة مستشفه من روح مسرح برخت لتأكيد نظريات إشراك المشاهد في الهم القومي ، وما آلت إليه البشرية من ظلم واضطهاد
في ظل العنصرية السائدة وسيطرة الدول الكبرى على المناخ السياسي الحالي .
وكان للوجه النسائي حظ وافر في الظهور ، حيث قدمت طالبات كلية الرستاق مسرحية اسم لم يكتمل وطالبات كلية عبري مسرحية خيوط من
سراب ، حيث أثبتت المرأة وجودها كمؤلفة ومخرجة وممثلة ، وشهدت العروض المقدمة تنافسا ينم عن توازي التجربة الإبداعية بين الجنسين وهذا
من شأنه ان يولد ثقافة مسرحية تتقبل مشاركة المرأة علي المدى البعيد في الفنون الدرامية .
ومن كلية التربية بالرستاق تميزت الطالبة ملاك الغيلاني بوضع بصمات واضحة تنم عن موهبة بكر ، من خلال قدرتها على معايشة دور الأم والابنة أمل
في أن واحد بصدق في مسرحية اسم لم يكتمل وأثبتت جدارتها في التعبير عما يكتنف نفسها من صراع داخلي مرير ، مما أهلها للفوز بجائزة أفضل
ممثلة وأفضل مخرجة .
واحتلت كلية صحار المركز الأول في مجال السينوغرافيا وكعرض متكامل ، بعرض رائعة سعدالله ونوس جثة علي الرصيف والتي أبدع الطالب
الفنان عبدالله الخزيري في نسج مفرداتها الجمالية كما تميز عرضا كلية عبري وصور بتشكيل إئتلاف جمالي بين مكونات السينوغرافيا كالإضاءة
والصوت والحركة والإيقاع العام والديكور .
ولقد تميز أكثر من عرض بوجود ديكور جيد كمسرحية محاكمة الشيطان ومسرحية خيوط من سراب وجثة علي الرصيف 000أتمنى أن أرى في
العروض القادمة تميزا في التشكيل والتصميم الهندسي للمشاهد ، وأن يكون الديكور مستشفا من عوالم النص المسرحي ويخضع لرؤية المخرج ...

Statistics
Number of Sentences:6
Number of Words:399
Number of Candidate Words:1026

Summarized Text

هذا ما جسدته لنا مسرحية محاكمة الشيطان لكلية التربية بنزوى من خلال حوار جدلي فلسفي بين الواقعية و عالم الفانتازيا أو الخيال لتجسيد
الصراع الابدي بين الضمير -LRB- الخير -RRB- والرغبات الإنسانية -LRB- الشر -RRB- التي قد توقع صاحبها في الخطأ ، و تجعل منه على شفا حفرة
من ارتكاب الجرم الذي يستحق عليه العقاب .
الاقتباس من النصوص العربية يعد بادرة طيبة من قبل الهواة في هذه التظاهرة ، لذا كان عرض مسرحية سعدالله ونوس جثة علي الرصيف من
العروض القوية والتي حصدت جوائز هذه الفعالية باستثناء جائزة احسن نص والتي عادة تخصص لنص المؤلف من قبل الطلبة المشاركين ، وبعد
مسرح سعدالله ونوس من المدارس الحقيقية التي يفخر بها المسرح العربي والذي خلد لنا العديد من روائع المسرح الملحمي ، ففي هذا العرض
تضامنت جماليات رائعة مستشفه من روح مسرح برخت لتأكيد نظريات إشراك المشاهد في الهم القومي ، وما آلت إليه البشرية من ظلم واضطهاد
في ظل العنصرية السائدة وسيطرة الدول الكبرى على المناخ السياسي الحالي .
واحتلت كلية صحار المركز الأول في مجال السينوغرافيا وكعرض متكامل ، بعرض رائعة سعدالله ونوس جثة علي الرصيف والتي أبدع الطالب الفنان
عبدالله الخزيري في نسج مفرداتها الجمالية كما تميز عرضا كلية عبري وصور بتشكيل إئتلاف جمالي بين مكونات السينوغرافيا كالإضاءة والصوت
والحركة والإيقاع العام والديكور .

Statistics
Number of Sentences:3
Number of Words:209
Number of Candidate Words:175
Summary to original %:52.38

Word	Normalized Word	Frequency	Weight	Lexical Chain	Senses
شَيْطان	شيطان	8	730	[جدل، خيال، توقع، روح، ملاك، عنصر،]	[ابليس، إعتقاد، ديانة، روح]
حِوار	حوار	1	2235	[خيال، جدل، عالم، بين، خير، رغب، صاحب، ...	[مفاوضات، أخذ وعطاء، نقاش، جدال، أظهر، ...
جَدَل	جدل	1	5175	[شيطان، حوار، جدل، جدل، خيال، عالم، بين، ...	[جدال، أظهر، أشار إلى، ناظر، مارى، جادل، ...
بَيَّن	بين	10	5620	[حوار، جدل، بين، بين، جدل، واقعي، عالم، ...	[شرح، فسر، وضح، بين، أعلم، أطلع، أخبر، ...
واقِعِيّ	واقعي	2	2440	[بين، جدل، خير، صاحب، عالم، رغب، خيال، ...	[حقيقي، واقعي، حقيقي، واقع، حقيقة، واقعية، ...
عالَم	عالم	0	5180	[حوار، جدل، بين، جدل، صاحب، بين، واقعي، ...	[عالم، إنسان، أحد ما، فرد، روح، شخص، ...
خَيال	خيال	0	5110	[شيطان، حوار، جدل، بين، جدل، صاحب، بين، ...	[راكب، جوال، مسافر، سائح، رحالة، متجول، ...
تَجْسِيد	تجسيد	0	2205	[بين، جدل، صاحب، رغب، خير، خيال، عالم، ...	[تصوير، إنتحل شخصية، اتخذ شخصية، مثل، مثل ...
صِراع	صراع	0	2240	[صراع، صراع، بين، جدل، صاحب، خيال، عالم، ...	[نزاع، عمل جماعي، تقاتل، قاتل، قاوم، كافح، ...
خَيْر	خير	0	5110	[حوار، جدل، بين، بين، واقعي، جدل، صاحب، ...	[خير، السيء، الرديء، الطالح، السوء، نوعية، ...
رَغَب	رغب	2	5100	[حوار، جدل، بين، بين، واقعي، جدل، عالم، ...	[أراد، طلب، طالب، مرام، رغبة، مراد، مبتغى، ...
إِنْسانِيّ	انساني	0	175	[انساني، انساني، روح، بشري، نفس،]	[انساني]
شَرّ	شر	0	2170	[بين، رغب، خيال، ظلم، حفر، توقع، عالم، خير، ...	[سوء، خرق، تعد، تجاوز، مخالفة، انتهاك، نذالة، ...
[illegible]	[illegible]	0	5110	[illegible]	[illegible]

Senetence	Score
ولقد تميز أكثر من عرض بوجود ديكور جيد كمسرحية محاكمة الشيطان ومسرحية خيوط من سراب وجثة علي الرصيف 000أتمنى أن أرى في العروض ...	2352
واحتلت كلية صحار المركز الأول في مجال السينوغرافيا وكعرض متكامل ، بعرض رائعة سعدالله ونوس جثة علي الرصيف والتي أبدع الطالب الفنان ...	2767
ومن كلية التربية بالرستاق تميزت الطالبة ملاك الغيلاني بوضع بصمات واضحة تنم عن موهبة بكر ، من خلال قدرتها على معايشة دور الأم والابنة أمل في ...	2187
وكان للوجه النسائي حظ وافر في الظهور ، حيث قدمت طالبات كلية الرستاق مسرحية اسم لم يكتمل وطالبات كلية عبري مسرحية خيوط من سراب ...	1899
الاقتباس من النصوص العربية يعد بادرة طيبة من قبل الهواة في هذه التظاهرة ، لذا كان عرض مسرحية سعدالله ونوس جثة علي الرصيف من ...	3224
هذا ما جسدته لنا مسرحية محاكمة الشيطان لكلية التربية بنزوى من خلال حوار جدلي فلسفي بين الواقعية و عالم الفانتازيا أو الخيال لتجسيد الصراع ...	2361
ولقد تميز أكثر من عرض بوجود ديكور جيد كمسرحية محاكمة الشيطان ومسرحية خيوط من سراب وجثة علي الرصيف 000أتمنى أن أرى في العروض ...	7098
[illegible]	8809

Figure 5. *Graphical User Interface (GUI)*

4. Summary and Discussion

Evaluating summaries and automatic text summarization systems is not a straightforward process. When evaluating a summary, generally we measure two properties: the Compression Ratio and the Retension Ratio, i.e. how much of the central information is retained. We can also evaluate the qualitative properties of the summaries, such as how coherence, and readability. This is usually done by using a panel of human judges [13].

4.1. *Comparing with human summaries*

We measured how our system performed relative to different human summaries. We choose a sample text file which includes 34 sentences in total. First, we generate summaries using our system for this text in different extraction rates which is as follows:

Table 1. *Number of summary sentences*

Rate	#number of sentences un the summary
10%	4
20%	7
30%	10
40%	14
50%	17

We gave five native english-speaking persons the same sample to summarize. Let's call them person1 to person5.

Total # of sentences in the original text is = 34 sentences.
A: # of our project's summary sentences (*differentiate depending on extraction rate*).
p1: # of person1 summary sentences = 12
p2: # of person2 summary sentences = 17
p3: # of person3 summary sentences = 14
p4: # of person4 summary sentences = 18
p5: # of person5 summary sentences = 19

Comparing human summaries with our system summary, we have found that the intersection between the summary sentences of two of them is as shown in Table 2. (Note: The intersections represent the number of sentences which are common in human summary and our system's summary)

Table 2. *Human and our system summary intersection*

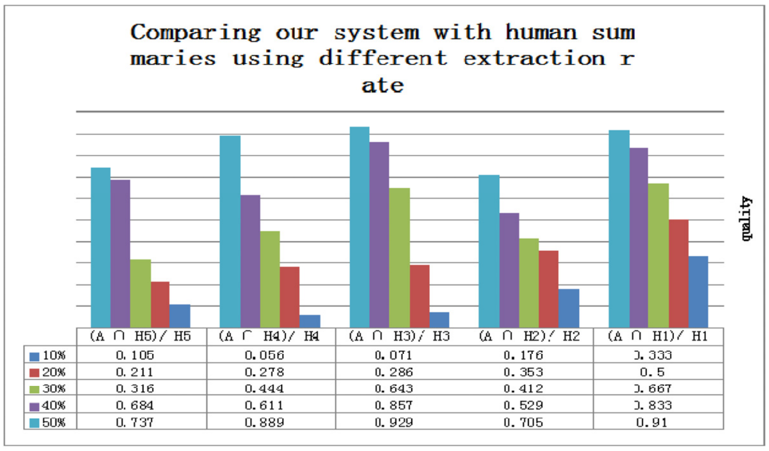
Ex. rate	A	A ∩ p1	A ∩ p2	A ∩ p3	A ∩ p4	A ∩ p5
10%	4	4	3	1	1	2
20%	7	6	6	4	5	4
30%	10	8	7	9	8	6
40%	14	10	9	12	11	13
50%	17	11	12	13	16	14

Finally, to measure the quality of our system depending on human summaries we have divide the intersection between the summary sentences of two of them by the total number of human summaries sentences in every percentage of extraction. The results are as shown in Table 3. We conclude that the accuracy of our system will increase when the rate of extraction is increasing.

Table 3. *Projects's summary quality comparing to human summaries*

Ex. rate	(A ∩ H1)/ H1	(A ∩ H2)/ H2	(A ∩ H3)/ H3	(A ∩ H4)/ H4	(A ∩ H5)/ H5
10%	0.333	0.176	0.071	0.056	0.105
20%	0.500	0.353	0.286	0.278	0.211
30%	0.667	0.412	0.643	0.444	0.316
40%	0.833	0.529	0.857	0.611	0.684
50%	0.910	0.705	0.929	0.889	0.737

Figure 6 gathers the whole information and results of comparing our system with 5 human summaries using different extraction rate. The vertical coordinate show the quality percentage and the horizontal coordinate represents the intersection of ours system's summaries and human's summaries to human summaries ration using different extraction rate summaries.



	(A ∩ H5)/ H5	(A ∩ H4)/ H4	(A ∩ H3)/ H3	(A ∩ H2)/ H2	(A ∩ H1)/ H1
🟦 10%	0. 105	0.056	0. 071	0. 176	0. 333
🟥 20%	0. 211	0. 278	0. 286	0. 353	0. 5
🟩 30%	0. 316	0. 444	0. 643	0. 412	0. 667
🟪 40%	0. 684	0. 611	0. 857	0. 529	0. 833
🟦 50%	0. 737	0. 889	0. 929	0. 705	0. 91

Figure 6. *Comparing ours system with human summaries using different extraction rate*

How the extraction rates effects the result? Having high extraction rates like 60% for example in small texts does not make sense because in such a case the summary may contains the most of those texts which is not reasonable. On the other hand, having small extraction rates like 20% in a summarization of large texts may result in loss of critical information of the original texts.

4.2. *Evaluation using human judgment*

Using human judgments to have another way of evaluating a summary even that the expensive cost of human judgment. We ask 7 different human to judge the summaries of our system using 6 different texts from different categories which are culture, economy, international, local, religion and sport.

First, we generate summaries for those 6 texts with different sizes and we get different numbers of summary sentences as follows:

# of summary sentences using culture text = 8
# of summary sentences using economy text = 10
# of summary sentences using international text = 7

# of summary sentences using local text = 5
# of summary sentences using religion text = 9
# of summary sentences using sport text = 4

Then, we give these all summaries for each human involved in this evaluation and we ask them to count the number sentences that describe the entire text to the total number of summary sentences in different types of texts. Finally, we find the ratio of these numbers of sentences to the total number of sentences in the summaries.

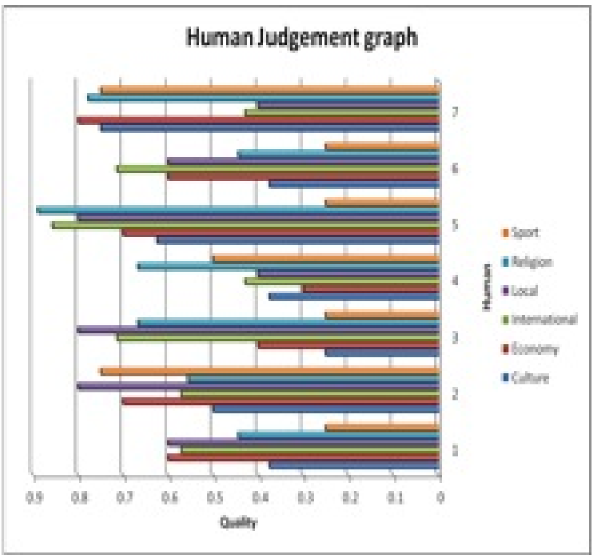


Figure 7. *Human judgment on six texts*

Table 4. Number of sentences that describe the entire text to the total number of summary sentences in different types of texts

	Culture	Economy	International	Local	Religion	Sport
Human1	0.375	0.600	0.571	0.600	0.444	0.250
Human2	0.500	0.700	0.571	0.800	0.556	0.750
Human3	0.250	0.400	0.714	0.800	0.667	0.250
Human4	0.375	0.300	0.429	0.400	0.667	0.500
Human5	0.625	0.700	0.857	0.800	0.889	0.250
Human6	0.375	0.600	0.714	0.600	0.444	0.250
Human7	0.750	0.800	0.429	0.400	0.778	0.750

Figure 7 shows human judgment of our system using 6 different texts from different categories which are culture, economy, international, local, religion and sport. We can see from the graph that the human judgment will depend on the type of the text.

REFERENCES

1. Abbas, M., Smaili, K. & Berkani. 2011. Evaluation of topic identification methods on Arabic corpora. *Journal of Digital Information Management*, 9/5, 185-192.
2. Barzilay, R. & Elhedad, M. 1997. *Using* lexical chains for text summarization. In proceedings of the *Intelligent Scalable Text Summarization Workshop (ISTS'97), ACL*
3. Buckwalter, T. 2001. ARAMORPH Arabic morphological analyzer. *Linguistic Data Consortium.*
4. Das, D. & Martins, A. 2007. A survey on automatic text summarization. *Literature Survey for the Language and Statistics Course at CMU.*
5. Doran, W., Stokes, N., Carthy, J. & Dunnion, J. 2004. Assessing the impact of lexical chain scoring methods on summarization. In proc of *CICLING'04* (pp. 627-635).
6. Douzidia, F. S. & Lapalme, G. 2004. *Lakhas, an Arabic Summarization System*.
7. Ercan, G. & Cicekli, I. 2008. Lexical cohesion based topic modeling for summarization. *CICLing 2008, LNCS 4919* (pp. 582-592).
8. Galley, M. & McKeown, K. 2003. Improving word sense disambiguation in lexical chaining. In proceeding of *18th International Joint Conference on Artificial Intelligence pages 1486-1488*
9. El-Haj, M., Kruschwitz, U. & Fox, C. 2011. Experimenting with automatic text summarization for Arabic. *Human Language Technology. Challenges for Computer Science and Linguistic* (pp. 490-499), Springer Berlin Heidelberg.
10. El-Haj, M., Kruschwitz, U. & Fox, C. 2010. Using mechanical turk to create a corpus of Arabic summaries. In the *7th International Language Resources and Evaluation Conference (LREC 2010)* (pp. 36-39), Valletta, Malta, LREC.
11. Farghaly, A. & Shaalan, A. 2009. Arabic natural language processing: Challenges and solutions. *ACM Transactions on Asian Language Information Processing*, 8/4, Article 14.

12. Haboush, A., Al-Zoubi, M., Momani, A. & Tarazi, M. 2012. Arabic text summarization model using clustering techniques. *World of Computer Science and Information Technology Journal*, 2/3, 62-67
13. Hassel, M. 2004. Evaluation of automatic text summarization. Licentiate Thesis. *KTH Numerisk analys och datalogi*, Stockholm, Sweden.
14. Larkey, L. & Connell, M. E. 2001. Arabic information retrieval at UMASS in TREC-10. In proceedings *of the 10th Text Retrieval Conference (TREC'01).*
15. Lehal, G. 2010. A Survey of text summarization extractive techniques. *Journal of Emerging Technologies in Web Intelligence*, 2/3, 258-268
16. Jezek, K. & Steinberger, J. 2007. Automatic text summarization (The state of the art 2007 and new challenges). *Information Processing and Management*, 42/6.
17. Klein, D. & Manning, C. D. (2002). Fast exact inference with a factored model for natural language parsing. *In NIPS.*
18. Medelyan, O. 2007. Computing lexical chains with graph clustering. In proceedings of the *Association for Computational Linguistics* (pp. 85-90).
19. Miller, G. 1995. WordNet: A lexical database for English. *Communications of the ACM*, 38/11, 39-41
20. Morris, J. & Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Association for Computational Linguistics*, 17/1.
21. Radev, D., McKeown, K. & Hovy, E. 2002. Introduction to the Special Issue on Summarization. *Computational Linguistics*, 28/4, 399-408.
22. Silber, G. & McCoy, K. 2002. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Association for Computational Linguistics*, 29/4, 487-496.

HAMZA ZIDOUM
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF SCIENCE,
SULTAN QABOOS UNIVERSITY,
MUSCAT, SULTANATE OF OMAN.

AHMED AL-MAAMARI
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF SCIENCE,
SULTAN QABOOS UNIVERSITY,
MUSCAT, SULTANATE OF OMAN.

NASSER AL-AMRI
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF SCIENCE,
SULTAN QABOOS UNIVERSITY,
MUSCAT, SULTANATE OF OMAN.

AHMED AL-YAHYAI
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF SCIENCE,
SULTAN QABOOS UNIVERSITY,
MUSCAT, SULTANATE OF OMAN.

SAID AL-RAMADHANI
DEPARTMENT OF COMPUTER SCIENCE,
COLLEGE OF SCIENCE,
SULTAN QABOOS UNIVERSITY,
MUSCAT, SULTANATE OF OMAN.