World of Computer Science and Information Technology Journal (WCSIT)
ISSN: 2221-0741
Vol. 5, No. 3, 51, 60, 2015

Lexical Cohesion and Entailment based Segmentation for Arabic Text Summarization (LCEAS)

Fatima T. AL-Khawaldeh
Department of Computer Science
Al-Albayt University
Al-Mafraq, Jordan

Venus W. Samawi
Department of Computer Information System
Amman Arab University
Amman, Jordan

Abstract—Text summarization is the process of creating a short description of a specified text while preserving its information context. This paper tackles Arabic text summarization problem. The semantic redundancy and insignificance will be removed from the summarized text. This can be achieved by checking the text entailment relation, and lexical cohesion. Accordingly, a text summarization approach (called LCEAS) based on lexical cohesion and text entailment relation is developed. In LCEAS, text entailment approach is enhanced to suit Arabic language. Roots and semantic-relations are used between the senses of the words to extract the common words. New threshold values are specified to suit entailment based segmentation for Arabic text. LCEAS is a single document summarization, which is constructed using extraction technique. To evaluate LCEAS, its performance is compared with previous Arabic text summarization systems. Each system output is compared against Essex Arabic Summaries Corpus (EASC) corpus (the model summaries), using Recall-Oriented Understudy for Gisting Evaluation (ROUGE) and Automatic Summarization Engineering (AutoSummEng) metrics. The outcome of LCEAS indicates that the developed approach outperforms the previous Arabic text summarization systems.

Keywords- Text Summarization; Text Segmentation; Lexical Cohesion; Text Entailment; Natural Language Processing.

I. INTRODUCTION

The increasing availability of online information has revived the interest in automatic text summarization. Text summarization is one of the Natural Language Processing (NLP) fields, which is employed in different domains such as decision making, and search engines. Selecting the most important portions of the text and generating coherent summaries is the aim of text summarization systems [1]. Text summarization system should produce document (or multi-documents) summaries that contains most (better to be all) of the significant information included in the original text [2]. Automatic summarization must produce a summary that includes the informative information from single, or multi documents [3]. Essentially, two Automatic text summarization techniques could be used: *extraction technique*, or *abstraction technique*. In extraction technique, the most important sentences are selected based on the statistical and linguistic features. In Abstraction technique, the summary is generated using linguistic methods to interpret the text and paraphrase it [4]. Text summarization system could be single document or multi documents summarization system.

In Arabic language, very few researchers tackled text summarization problem. The main problem of most previous Arabic text summarization systems is that the summarized text contains many sentences with redundant information, or it may contain unimportant information. A measure of lexical cohesion (semantically related words) can be used to detect and remove the insignificant information to improve the value of the text summary. The challenge is to find a way to measure if the meaning of a sentence can be inferred from information available in the preceding sentence. In this case, the second sentence is considered to be redundant information, which should be excluded from the summary [13].

In this work, we tackle the problem of developing an Arabic text summarization system that produces most informative of the original text without redundancy. To solve the redundancy problem and the poor information in the previous Arabic text summarization systems, lexical-cohesion and entailment-based-segmentation will be utilized. Accordingly, an Arabic text summarization system (called LCEAS) is developed by making use of text entailment and lexical cohesion in order to get a balanced text summary. A measure of lexical cohesion (semantically related words) is used to detect and remove the unimportant information in order to improve the quality of the summary. Text entailment is a method for matching two texts in order to check if the statement of one text is logically inferred by another.

In LCEAS, the text entailment algorithm suggested in [14] is enhanced to make it suitable for Arabic Language. This enhancement includes using roots and semantic relations between the senses of the words to extract the common words, and specify new threshold values (empirically) for Arabic texts. The quality of the summary produced by LCEAS is measured using ROUGE and AutoSummEng metrics. LCEAS is evaluated by comparing its performance with the performance of other two popular approaches suggested in [9] and [10].

This paper is organized in 5 sections. Related works are illustrated in section 2. The LCEAS phases are summarized in section 3. The experimental results are discussed in section 4. Finally, we concluded in section 6.

II. RELATED WORKS

The authors in [5] employed the Rhetorical Structure Theory (RST) to parse Arabic text into a tree based on rhetorical relations. A suitable level of the tree is selected to represent the summary. The performance of the system fits small and medium sized articles. In [6] the author examined set of eleven statistical features, which is then reduced to five useful features. Naive Bayesian classifier and genetic programming classifier are used to extract high quality summary. Arabic Query-Based Text summarization System AQBTSS, proposed by the authors of [7], is a query-based single document summarizer system. An Arabic document and query are used to generate summary by using standard retrieval methods for the document around this query. AQBTSS performance is compared with Sakhr summarizer (Sakhr Software Company is the pioneer in Arabic NLP technologies) [8]. In Sakhr summarizer, the important sentences within documents are extracted by applying some features (Keywords Distribution, Weighting, Sentence Type, Document Categories, Sentence Length, Position, and Title) (Sakhr website). AQBTSS outperforms Sakhr summarizer from the human assessor's point of view. Another single document summarization system is Concept-Based Text summarization System (ACBTSS) by the authors of [7]. In this system, each sentence is matched against a set of keywords which represents a concept. AQBTSS and ACBTSS systems are compared. The overall performance shows that AQBTSS outperforms ACBTSS. The authors of [9] proposed an Arabic text summarization approach based on an aggregate similarity method (noun/verb categorization method), which is originally proposed for the Korean language text. The frequencies for each noun in each sentence, and in the whole document are computed. The sentence similarity between the noun frequency in the sentence and the document is calculated using the cosine equation. The summation of all similarities of every sentence represents a total similarity. The sentences with highest value of similarity are selected to represent the summary. The authors of [10] implemented a technique of word root clustering. They adopt cluster weight of word roots instead of the word weight itself. All words with the same root are put in the same cluster, and the number of words in that cluster is calculated. Term frequencies, in the sentence and in the text, are calculated to compute the score of each sentence. The highest score sentences are selected to represent the summary. The authors of [11] suggested a two pass Arabic text summarization algorithm. In the first pass, a primary summary using Rhetorical Structure Theory (RST) is produced. In the second pass, a score to each of the sentences in the primary summary are assigned in order to generate the final summary. Experiments on sample texts proved that this system outperform some of the existing Arabic summarization systems including those that used machine learning. Research concerning languages closely related to the Arabic language, such Farsi, showed that the result of applying Farsi summarization system to summarize Arabic text was unsatisfactory [11]. Although Farsi includes Arabic words, but the structure of the sentence is different due to different origins of the words. For Arabic text classification, Arabic text summarization was used as an effective feature selection technique which addressed its effectiveness in Arabic text classification [12].

III. LCEAS SYSTEM MODEL

This section aims to summarize the main phases of the suggested system (LCEAS). Fig. 1 illustrates LCEAS architecture, which clarifies the main stages of LCEAS (preprocessing stage, word sense disambiguation (WSD) stage, lexical cohesion -based segmentation for summarization stage, and text entailment based segmentation for summarization stage).

A. *Pre-processing Stage*

This stage is based on two steps. These steps are concerned with the analysis of the summarized text, and converting it from unstructured form to structured form. Step one is *removing stop words* (words that are considered as unimportant or irrelevant words). The stop words list of [15] is employed in LCEAS. The performance of eliminating these lists improves the retrieval effectiveness in Arabic language texts [15]. Step two is *word stemming*. The Arabic language is highly derivational language. Many different forms for the same word could occur. In word stemming, every word in the sentence is changed to its root. In LCEAS, the stemming approach suggested in [16] is used as they claimed that the hybrid method is more effective than the previous methods, where the obtained average accuracy is 57%.

B. *Word Sense Disambiguation (WSD) Stage*

In this phase, the correct meaning (sense) of each word in the text is identified based on the context of the text. Arabic WordNet (a lexical resource for Arabic language) is used to perform the pre-processing stage in order to find the lexical relations. In this stage, three techniques of WSD are combined.

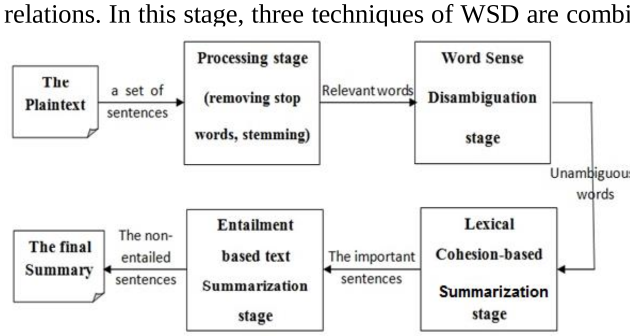

Figure 1. LCEAS Architecture

These techniques are the WSD technique used in [17], the semantic relations proposed in [18], and the improvement of using the levels of senses of words in first, middle, and last sentences suggested in [19]. The used semantic relations are: gloss relation (a textual definition of the synonym set with a set of examples), holonym relation (which means that something is a part of another thing), and meronym relation (opposite to holonym).

The developed WSD algorithm consists of five steps (as illustrated in Fig. 2): 1) Extract all the possible interpretations (senses) of each word. 2) Extract the three levels of senses (the beginning, middle, and final) and establish connections between the related senses. The first level is the senses of a word; the second level is the senses for each sense in the first level and so on. 3) Calculate the strength of the connections. 4) Summing all the strength of the connections. 5) Select the highest summation sense.

For example, suppose that we have (“Qaraa' al-talbe al-dars”, “قرأ الطالب الدرس”) as the first sentence of the text. In order to extract the correct sense of the word (“Qara’a'”, “قرأ”), its senses must be extracted in 3 levels, as illustrated in Fig. 3.

The senses of each word are compared against all the senses of all words in the text. A connection (link) is established if there is a semantic relation between the senses of the current word and any sense of the other words. If the sentence is first, middle, or last in the text, then we start comparing the senses of each word with the senses of the other words starting from level-3, then Level-2, and Level-1(as shown in Fig. 3). The semantic relations and their assigned weights, as suggested by [18], are:

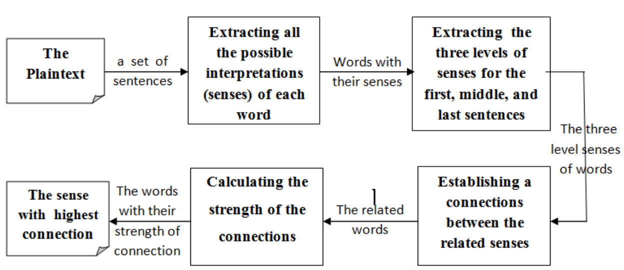

Figure 2. The main steps of the proposed WSD

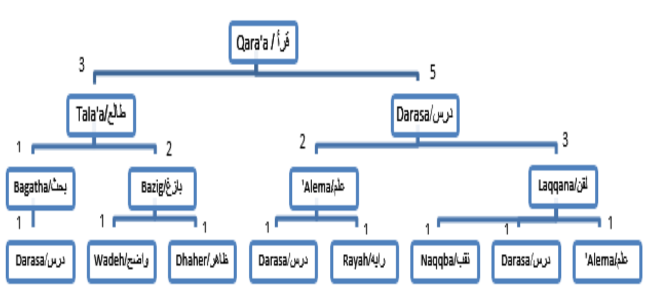

Figure 3. WSD on the sentence (“Qaraa' al-talbe al-dars”, “قرأ الطالب الدرس”)

- Repetition relation (same occurrences of the word), weight=1.
- Synonym relation (weight=1). In the example above, the word (“Alema”, “ علم“) has a synonym semantic relation with the sense (“Darasa”, “درس”).
- Hypernym and Hyponym relation (weight=0.5): Y is a hypernym of X, if X is a (kind of) Y; X is a hyponym of Y, if X is a (kind of) Y.(e.g. X=(computer, كمبيوتر), Y=(a'leh, آلة).
- Holonym and Meronym relation (weight=0.5): holonymy relation is (whole of) and meronymy relation is (part of). Y is a holonym of X, if Y is a whole of X; X is a meronym of Y, if X is a part of Y. X=(“lauhaat-almafatih”, “لوحة المفاتيح”),Y=(“computer”, “كمبيوتر”).
- Gloss relation (definition and/or example sentences for a synset), (weight=0.5): consider the word=(shaja'a, الشجاعة), gloss=(“alihsas be-adam alkhowf”, “ الإحساس بعدم الخوف”).

Each sense has a number of weighted links with the related senses of other words. The weighted links between the senses are summed. The sense which has the highest sum is the correct sense of the word. As shown in the example (Fig. 3), the correct sense of the word (“Qara'a'”, “قرأ”) is (“Darasa”, “درس”) as its weighted link is 5.

C. *Lexical Cohesion based Segmentation Stage*

Lexical cohesion is applied to distinguish the important sentences from the unimportant sentences in the text. Lexical cohesion divides the text into segments. Each segment consists of set of sentences with a specific topic. The most important segments include the main topics. The selected sentences to represent the summary are the sentences of the most important segments. Computing the lexical cohesion among the parts of text helps identifying poor information. Poor information is removed to enhance the quality of summary before applying the text entailment based segmentation for summarization. In this work, the lexical chain procedure suggested by [19] is used as it improves the performance compared to state-of-art summarization procedures. After the proposed WSD is implemented in the previous stage, the algorithm divides the text into segments (related sentences) based on topics as shown in Fig. (4):

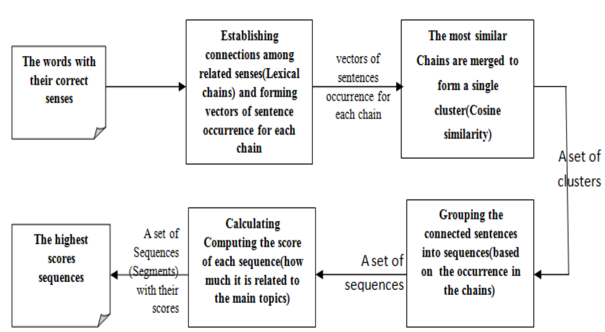

Figure 4. Lexical cohesion-based segmentation for summarization

The procedure of lexical cohesion-based segmentation for Arabic text summarization is as suggested by [19]. If, for example, we have the following text from Essex Arabic Summaries Corpus EASC:

{دعا(Da'a) معهد(ma'had) ثربانتس(Therbantis) بعمان(be-Amman) للمشاركة(lelmosharakah) بمسابقة(be-Mosabaqt) الرسم(alrasm) الاولى(al'oula) للاطفال(lelatfal) المقيمين(almoqimien) في(fe) الاردن(al'ordon) والذين(wallathien) تتراوح(tatarawah) اعمارهم(a'amarohum) ما(ma) بين(bayn) 4 و12 سنة (sanah)، وعنوان(wa'enwan) مسابقة(mosabaqt) الرسم(alrasm) هو(howa) «عام(a'am) الكيخوتي(alkaykhouti)». وتضمنت(watadammanat) شروط(shorout) المسابقة(almosabaqah)، تنظيم(tantheem) مسابقة(mosabaqt) رسم(rasm) للاطفال(lelatfal) حول(hawl) موضوع(mawdou') الكيخوتي(alkaykhouti) حيث(haytho) بامكان(be'emkan) جميع(jamee') الاطفال(alatfal) الذين(allathein) تتراوح(tatarawah) اعمارهم(a'amarohum) ما(ma) بين(bayna) 4 و12 سنة(sanah) من(men) المقيمين(almoqemeen) في(fi) عمان(Amman) او(aw) في(fi) مدن(modon) المملكة(almamlakah) الاخرى(al'okhra) المشاركة(almosharekah) فيها(fiha). الجوائز(aljawa'ez) مقدمة(moqaddamah) من(men) قبل(qebal) شركة(sharekat) الدياسا(aldeyasa) الاسبانية(al'espanyyah) «الاسواق(alaswaq) الحرة(alhorrah) لمطار(lematar) الملكة(almalekah) علياء (alya') الدولي(aldowaly)» وشركة(washareqt) ايبيريا(eyberya) «خطوط(khotout) الطيران(altayaran) الاسبانية(al'espanyyah)». لاختيار(le'ekhteyar) الرسومات(alrosoumat) الفائزة(alfa'ezah) تتألف(tata'allaf) لجنة(lajnat) التحكيم(attahkeem) من(men) اعضاء(a'ada') اسبان(espan) واردنيين(wa'ordonyeen).}

After applying the proposed WSD algorithm on the text above, each word is assigned with the correct sense. Lexical chains (LC_i) are created by establishing connections among senses (meanings) of the words. If there are semantic relations between these senses, then the words are related [19].

LC_1={(عمان Amman, الاردن AlOrdon, عمان Amman, اردنيين Ordonyeen)}.

LC_2={(مشاركة Mosharakah, مشاركة Mosharakah, مسابقة Mosabaqt, مسابقة Mosabaqt, مسابقة Mosabaqt, مسابقة Mosabaqt)}.

LC_3={(الاطفال AlAtfal, الاطفال AlAtfal, الاطفال AlAtfal)}.

LC_4={(الكيخوتي AlKaykhouti, الكيخوتي AlKaykhouti)}.

LC_5={(سنه Sanah, عام A'am, سنة Sanah)}.

LC_6= {(عنوان 'enwan, موضوع Mawdou')}.

LC_7={(الرسم AlRasm, رسم Rasm, رسم Rasm, الرسومات AlRosoumat)}.

LC_8= {(اعمار a'amar, اعمار a'amar)}.

LC_9= {(مطار Matar, طيران Tayaran)}.

LC_{10} = {(الإسبانية Al'espanyyah, الإسبانية Al'espanyyah, إسبان Espan)}.

The score of each lexical chain is calculated using equation (1) and (2) [20].

Score (*Chain*)= *Length* × *Homogeneity* (1)

Homogeneity=1-*DistinctMembers*/*Length*) (2)

Where, Length is the number of occurrences of members of the chain; Homogeneity index represents the number of different occurrences divided by the length. The score of each chain is: (LC_1, 1), (LC_2, 4), (LC_3, 2), (LC_4, 1), (LC_5, 1), (LC_6, 0), (LC_7, 2), (LC_8, 1), (LC_9, 0), (LC_{10}, 1).

For each chain, a vector of sentence-occurrence in the chain is formed. $V_i=(s_{1i}, s_{2i}, s_{3i},,\dots s_{mi})$. For example, LC_1 appears twice in 1st sentence, once in the 2nd sentence, not appeared in sentence 3, and appeared once in the 4th sentence. Therefore, V_1=(2 1 0 1). The content of the vectors are:

V_1=(2 1 0 1), V_2=(3 3 0 0), V_3=(1 2 0 0), V_4=(1 1 0 0), V_5=(2 1 0 0), V_6=(1 1 0 0), V_7=(2 1 0 1), V_8=(1 1 0 0), V_9=(0 0 2 0), V_{10}= (0 0 2 1).

Each LC_i (vector) is considered as a cluster. The most similar clusters pairs are found and they are merged to form a single cluster. The related lexical chains are clustered to represent main topics in the text. The authors of [21] assumed that "if two lexical chains tend to appear in same sentences, then there may be a relation between two sets in the given context". Cosine equation is used to find the degree of similarity between each pairs of vectors. Given two vectors, $V_i=\vec{x}$ and $V_j=\vec{y}$, Cosine similarity is calculated using equation (3):

$$\cos(\vec{x},\vec{y}) = \frac{\vec{x}\cdot\vec{y}}{|\vec{x}||\vec{y}|} = \frac{\sum_{i=1}^{n} x_i\, y_i}{\sqrt{\sum_{i=1}^{n} x_i^2}\,\sqrt{\sum_{i=1}^{n} y_i^2}} \tag{3}$$

In this work, vectors with degree of similarity≥0.8 (0.8 is specified empirically) are merged to form a cluster. The sentences are not talking about the same topic when the degree of similarity is<0.8.

Cluster$_1$:
V_1=(2 1 0 1), V_2=(3 3 0 0), V_3=(1 2 0 0), V_4=(1 1 0 0), V_5=(2 1 0 0), V_6=(1 1 0 0), V_7=(2 1 0 1), V_8=(1 1 0 0),
Cluster$_2$:
V_9=(0 0 2 0), V_{10}= (0 0 2 1).

For each cluster, connected sequences of sentences are extracted separately to represent the segments (topics). If the sentence has a member in the chain and the current sequence is opened, then add the sentence to the sequence. If the previous sequence is closed, a new sequence is created. The sequence is kept open until we reach the sentence that does not contain a member in the chain. Each sequence of the chain is merged with its peer sentences. In the given example, there are two sequences in Cluster$_1$: Sequence$_1$ (s_1) includes the 1st and 2nd sentences, Sequence$_2$ (s_2) includes the 4th sentence. Cluster2 has Sequence$_3$ (s_3), which includes the 3rd and 4th sentences.

Cluster$_1$	s_1	s_2	**Cluster$_2$**	s_3
V_1 =	2 1	0 1	V_9 = (0 0	2 0)
V_2 =	3 3	0 0	V_{10}= (0 0	2 1)
V_3 =	1 2	0 0		
V_4 =	1 1	0 0		
V_5 =	2 1	0 0		
V_6 =	1 1	0 0		
V_7 =	2 1	0 1		
V_8 =	1 1	0 0		

The average score of the lexical chains in each cluster is calculated using equation (4) [19]:

$$S(Cl_i) = \sum \frac{Sclc}{N} \qquad (4)$$

Where, $S(Cl_i)$ is average score of the lexical chains in a $cluster_x$, Scl_c is the summation of scores of lexical chains in $cluster_x$, and N is number of lexical chains in $cluster_x$.

In the above example, the average of Cluster1 $S(Cl_i)=1.5$, and the average of $Cluster_2$ $S(Cl_i)=0.5$. The score of each sequence si in each cluster is calculated using equation (5) [19]:

$$S(s_i) = S(Cl_i) \times L_i \times (1 + SLC_i) \times PLC_i / f_i \qquad (5)$$

Where *i* is a sequence number, $S(Cl_i)$ is the average score of the lexical chains in $cluster_x$. L_i is the number of sentences in s_i. SLC_i is the number of lexical chains that starts in s_i between all extracted chains in all clusters (same sequence of sentences is the beginning sequence of the lexical chain). PLC_i is the number of lexical chains having a member (at least one sentence) in s_i in all clusters, and f_i is the number of lexical chains in the cluster.

Cluster₁:

In $Sequence_1$ $S(Cl_1)= 1.5$, $L_1=2$, $SLC_1= 8$, $PLC_1= 8$, $f_1=8$.

The score of $sequence_1$ $S(s_1) =27$.

In $Sequence_2$ $S(Cl_2)= 1.5$, $L_2=1$, $SLC_2=2$, $PLC_2=3$, $f_2=8$.

The score of $sequence_2$ $S(s_2) =1.687$.

Cluster₂:

In $Sequence_3$ $S(Cl_3)= 0.5$, $L_3=2$, $SLC_3=2$, $PLC_3=4$, $f_3=2$.

The score of $sequence_3$ $S(s_3) =6$.

The average score of sequences = (27+1.687+6)/3 =11.562.

The poor sequences are removed as they include poor sentences. If the score of a sequence < average score of sequences, then it is poor and will be eliminated. In this example, s_2 and s_3 are poor sequences (i.e. sentences 3 and 4 are poor and will be removed). The first and the second sentences are part of the obtained summary (these sentences are part of $sequence_1$, which is a strong sequence):

{ دعا(Da'a) معهد(ma'had) ثربانتس(Therbantis) بعمان(be-Amman) للمشاركة(lelmosharakah) بمسابقة(be-Mosabaqt) الرسم(alrasm) الاولى(al'oula) للاطفال(lelatfal) المقيمين(almoqimien) في(fe) الاردن(al'ordon) والذين(wallathien) تتراوح(tatarawah) اعمارهم(a'amarohum) مابين(mabayn) 4 و(wa) 12 سنة (sanah)، وعنوان(wa'enwan) مسابقة(mosabaqt) الرسم(alrasm) هو(howa) «عام(a'am) الكيخوتي(alkaykhouti)». وتضمنت(watadammanat) شروط(shorout) المسابقة(almosabaqah)، تنظيم(tantheem) مسابقة(mosabaqt) رسم(rasm) للاطفال(lelatfal) حول(hawl) موضوع(mawdou') الكيخوتي(alkaykhouti) حيث(haytho) بامكان(be'emkan) جميع(jamee') الاطفال(alatfal) الذين(allathein) تتراوح(tatarawah) اعمارهم(a'amarohum) مابين(mabayna) 4 و(wa) 12 سنة(sanah) من(men) المقيمين(almoqemeen) في(fi) عمان(Amman) او(aw) في(fi) مدن(modon) المملكة(almamlakah) الاخرى(al'okhra) المشاركة(almosharekah) فيها(fiha).}

D. *Text Entailment Based Segmentation for Summarization Stage*

Text entailment relation is used to decide whether the meaning of one sentence is inferred from another sentence. The entailment relation helps in checking the existence of a semantic connectedness (entailment) between two sentences. The summary obtained by using the entailment inferences only includes the sentences that are not entailed-by any of the sentences in the previously accumulated summary. The sentence with a meaning included in another sentence is described by the entailed-by sentence. The entailed-by sentence will be less informative than the sentence entails it. The higher dependency of meaning among several terms of the text indicates redundant information. The objective of this stage is to decide which sentences are not redundant according to text entailment relation. Three methods for textual entailment, proposed in [22] and [23] are used. These methods are:

- Textual entailment using similarity of texts obtained from the text-to-text metric [22].
- Cosine directional similarity for textual entailment, based on cosine calculation [23].
- Modified Levenshtein distance for textual entailment verification. It calculates the minimal number of transformations (deletions, insertions and substitutions) between the two portions of the text [23].

Cosine directional similarity for textual entailment is used in this research, as it has the highest computational precision compared with the other methods [23]. In this work, we enhanced the text entailment based segmentation for summarization method used by Tatar [14] as follows:

- Using the roots and semantic relations between the senses of the words to extract the common words.
- Specify new threshold values (empirically) to suit Arabic language texts.

The proposed text entailment based segmentation for summarization procedure is illustrated in Fig. 5.

For the example shown in lexical cohesion phase, calculate the common words between the sentences. Each word in the first sentence that has at least one semantic relation with any word in the second sentence is considered a common word. If the common word appears more one time in any sentence (T or H), which means more semantic relations (links) between the two sentences, it is repeated. By considering the same example

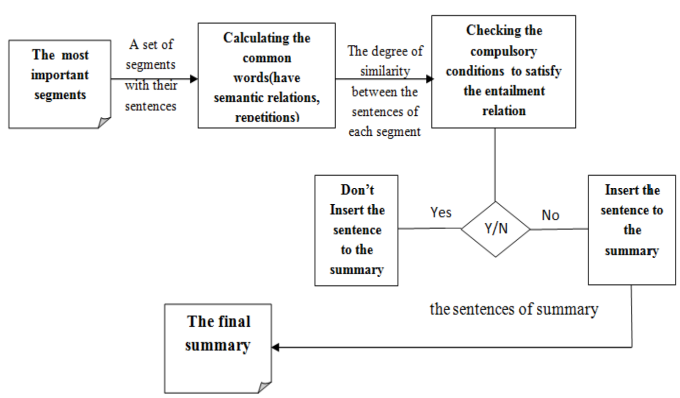

Figure 5. The proposed text entailment based segmentation for summarization procedure

above, the common words between the first sentence (T) and the second sentence (H) are 14 (as shown in Table 1).

Calculate the degree of similarity between the sentences of each segment using cosine similarity. Cosine directional similarity for text entailment is used in this research, as it has the highest computational precision compared with the other methods [14]. Cosine measures (cost(T,H)) consider the words of Text T that entails Hypothesis (i.e. T→H). T is the first sentence and H is the second one. T words are (t1, t2…tm), and H words are (h_1, h_2…h_n). Using equations (6,7), the two vectors that are used for calculating cost (T,H) are:

Let T= (1, 1,…1) (m-dimensional vector) and

H = (1, 1, …1) (n-dimensional vector)

$$H_i = \begin{cases} 1, & \text{i f } t_i \text{ is a word in sentnce H} \\ 0, & \text{otherwise} \end{cases} \qquad (6)$$

$$T_i = \begin{cases} 1, & \text{i f } h_i \text{ is a word in sentnce T} \\ 0, & \text{otherwise} \end{cases} \qquad (7)$$

For cosH∪T(T,H), the first vector is attained from the words of T contained in T∪H. The second vector is attained from the words of H contained in T∪H. This method considers c as the number of common words between T and H. The common words have a semantic relation. The three equations (8-10) in [14] are applied in LCEAS:

$$\text{cosT(T,H)}=\sqrt{c/m} \qquad (8)$$

$$\text{cosH(T,H)}=\sqrt{c/n} \qquad (9)$$

$$\text{cosH}\cup\text{T(T,H)}=\sqrt{4c^2/(n+c)(m+c)} \qquad (10)$$

By considering the above example, T is the 1st sentence, H is the 2nd sentence. Number of words in T (after eliminating the stop words and numbers) m=19, while number of words in H n=18. The common words between the two sentences c=14. By applying equations (8-10), cosT(T,H)=0.8583, cosH(T,H)=0.8819, and cosH ∪ T(T,H)=0.8616. The text entailment relation is satisfied between the two sentences if cosH(T,H)≥cosH∪T(T,H)≥cosT(T,H), taking in consideration that m ≥ n ≥c. To satisfy the entailment condition, T entails H if H is not informative with respect to T [14]. In other words, T entails H if the following conditions are satisfied:

TABLE I. THE COMMON WORDS BETWENN T &H

الاطفال Al atfal	الرسم Al rasm	مسابقة mosabaqt	المشاركة Al-mosharakah	عمان Amman
الاطفال Al atfal	الرسم Al rasm	مسابقة mosabaqt	عنوان 'enwan	سنة Sanah
	الكيخوتي Al kaykhouti	اعمارهم a'amarohum	عام a'am	الاردن al'ordon

$$\text{cosH}\cup\text{T} - \text{cosT} \le \tau 1$$

$$\text{cosH - cosH}\cup\text{T} \le \tau$$

$$\text{Max\{cost,cosH,cosH}\cup\text{T\}} \ge \tau 3$$

New threshold values are specified (empirically) to suit Arabic text. These thresholds are τ_1=0.095, τ_2=0.2, and τ_3=0.5. In the experiments on Arabic texts summarization, it was found that:

- Using $\tau_1 \le 0.095$, $\tau_2 \le 0.20$, and $\tau_3 \le 0.5$ leads to removing sentences which are redundant.
- Using $\tau_1 > 0.095$, $\tau_2 > 0.20$, and $\tau_3 > 0.5$, will keeps redundant sentences in the summary.

In the example above, the results of compulsory conditions are:

cosH∪T-cosT≤0.095.

cosH-cosH∪T =0.0203≤0.2.

Max{cosT;cosH;cosH∪T } is 0.8819≥0.5.

Since all the compulsory conditions are satisfied, the first sentence entails the second sentence. If the sentence is not entailed by other sentences, it is added to the summary. The summarization method based on the text entailment method is that the most important sentences of the most important sequences are selected and be part of the summary. The final summary holds the important sentences without redundancy. In the example above, the final summary includes only the first sentence, which matches the summary of the EASC text:

(be-Amman)بعمان (Therbantis)ثربانتس (ma'had)معهد (Da'a)دعا}
(alrasm)الرسم (be-Mosabaqt)بمسابقة (lelmosharakah)للمشاركة
(fe)في (almoqimien)المقيمين (lelatfal)للاطفال (al'oula)الاولى
(tatarawah)تتراوح (wallathien)والذين (al'ordon)الاردن
،(sanah) سنة 12و 4 (bayn)بين (ma)ما (a'amarohum)اعمارهم
(a'am)عام« (howa)هو (alrasm)الرسم (mosabaqt)مسابقة (wa'enwan)وعنوان
.{»(alkaykhouti)الكيخوتي

IV. EVALUATION OF LCEAS

Evaluating summaries and automatic text summarization systems is not a simple process. There are no ideal standards to evaluate the results of text summarization system [24] since there is no obvious "ideal" summary to evaluate the quality of the obtained summary. When an ideal extract has been created by human(s), extractive summaries are easy to evaluate [25]. NLP systems must address the relevancy issue in its evaluation to assess the reliability and beneficiary of these systems [26] .One of the most common ways for summary evaluation is by comparing the informative of automatic summaries against human made model [1]. Information retrieval metrics of precision, recall, and F-measure (equations 11 to 13) are used to evaluate LCEAS summaries against human summaries [27], [28] and [29].

$$\text{Recall} = \frac{\text{system - human choice overlap}}{\text{sentences chosen by human}} \qquad (11)$$

$$\text{Precision} = \frac{\text{system - human choice overlap}}{\text{sentences chosen by system}} \qquad (12)$$

$$F - measure = \frac{2 \times \text{Recall} \times \text{Precision}}{(\text{Recall} + \text{Precision})} \qquad (13)$$

To assess LCEAS as automatic text segmentation and summarization system, its performance is compared against: human judge, previous systems, and the improvement on computational task as measuring the influence of the proposed text segmentation on summarization. ROUGE and AutoSummENG are used to evaluate LCEAS. ROUGE is used because it is considered as the main evaluation metric in Document Understanding Conference (DUC 2004, 2005, 2006, 2007), (Web1, 2012). In Text Analysis Conference TAC (2008, 2009, 2010, 2011), AutoSummENG metric is used. AutoSummENG is a summarization evaluation method that evaluates summaries by comparing graphs of character and word N-gram.

Two datasets are used to evaluate LCEAS, Arabic Reuters newswire and Essex Arabic Summaries Corpus (EASC). Arabic Reuters newswire is used to evaluate LCEAS against human judge. Fifty documents are selected from Arabic Reuters newswire that cover different topics (computer, education, sport, health, and politics) written by native speakers. On the other hand, Essex Arabic Summaries Corpus (EASC) is used to evaluate the proposed system compared against Al-Radaideh and Haboush systems using TAC2011 metrics. EASC contains 153 Arabic articles and 765 human-generated summaries [30] (Lancaster university website).

A. *Evaluation of Text Segmentation Using Manual Segmented Texts*

Fifty specialized participants of five groups from Arab Open University instructors and Al-Tamyouz school instructors in Kuwait with different majors are selected to evaluate LCEAS system. In order to evaluate the segmentation method used in LCEAS and compare it against human segmentation, each group is asked to segment the texts of the fifty articles that are selected from Arabic Reuters newswire. Each text is segmented using LCEAS, and then it is compared against the text which is segmented by the fifty professional individuals (the gold standards). Table 2 illustrates how to calculate the recall, precision, and F-measure between the manual method (fifty professional individuals) and the proposed segmentation method in LCEAS. The example is a text consists of 33 sentences is segmented to 11 segments by humans, and to 9 segments by the LCEAS.

The number of correct gaps is the number of the beginning of the segments that have been identified by LCEAS, which differ by -1.0, +1 from the beginning of the segments that have been identified manually [14]. The human evaluation of the summaries obtained by applying LCEAS, with and without WSD method, is illustrated in table 3:

TABLE II. EXAMPLE COMPARISION BETWEEN MANUAL SEGMENTATION AND LCEAS SEGMENTATION

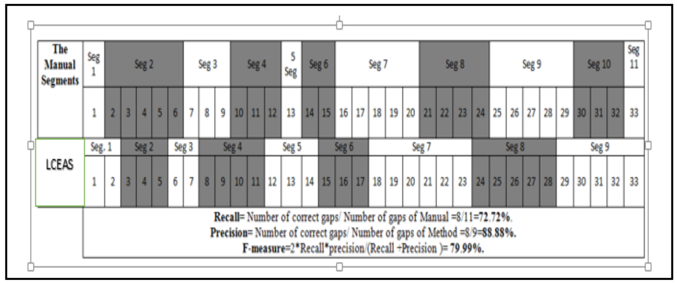

TABLE III. HUMAN EVALUATION OF LCEAS TEXT SEGMENTATION

LCEAS segmentation	Recall	Precision	F-Measure
with WSD	73.6%	66.7%	69.98%
without WSD	68.4%	57.8%	62.65%

B. *Human Judge Evaluation on LCEAS*

Fifty articles are selected from Arabic Reuters newswire and summarized by three Arabic text summarization systems: LCEAS, Al-Radaideh, and Haboush system. Fifty professional persons are partitioned into five groups. Each group has to judge 10 summarized texts obtained by the three systems. For human judgment, five evaluation scales are used. Theses scales are stated in DUC 2005: very poor (the summary doesn't hold the main ideas of plaintext), poor (some topics are hold, but most of them are missed), fair (most of the main ideas are hold, but still missing ideas), good (most of the main ideas are hold with redundancy information), and very good (the main ideas are hold without redundancy). Each individual evaluates the summaries obtained by Al-Radaideh, Haboush, and LCEAS. By comparing the evaluation results of the three text summarization systems (see Fig. 6), it is clearly seen that LCEAS performance is high in good and very good scales compared with the other two systems. This indicates that LCEAS outperforms Al-Radaideh, and Haboush systems.

C. *Automatic Summarization Engineering (AutoSummEng)*

The AutoSummENG system is used to evaluate set of summarizing systems with respect to a given set of model summaries. AutoSummENG evaluates summaries by extracting and comparing graphs of character N-grams [31]. It is robust and effective and directly comparable (or even better

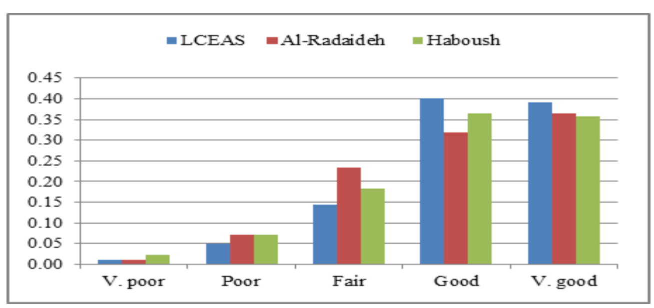

Figure 6. The Human Judge Results

than) ROUGE [32]. Reference [33] generated extractive summaries of the same set of documents of EASC corpus using a number of Arabic text summarization systems. Th ese Arabic summarization systems are: Sakhr (an online Arabic summarizer), AQBTSS [7], Gen-Summ (similar to AQBTSS except that the query is replaced by the document's first sentence), and LSA-Summ (similar to Gen-Summ, but the vector space is transformed and reduced by applying Latent Semantic Analysis (LSA) to both document and query). The values obtained by applying those systems with AutoSummENG for "CharGraphValue" is within the range 0.516–0.586. AutoSummENG is also used to evaluate LCEAS, Al-Radaideh, and Haboush after applying them on EASC corpus. Fig. (7) Shows that LCEAS AutoSummENG value is 0.725, which outperforms the other systems.

D. Recall-Oriented Understudy for Gisting Evaluation (ROUGE)

ROUGE includes measures to automatically evaluate the quality of a summary by comparing it with other ideal summaries created by humans. ROUGE counts the number of overlapping units (such as N-gram, word sequences, and word pairs) between the automatically-generated text summary and the ideal text summary produced by humans [34] ROUGE-N (i.e. for N-gram, N=2), ROUGE-L, ROUGE-W, and ROUGE-S worked well in single document summarization tasks [34]. ROUGE evaluation procedure is based on N-gram co-occurrence, longest common subsequence and weighted longest common subsequence between the ideal summary (human summary) and the automatically generated summary.

1) ROUGE-N (N-gram Co-Occurrence Statistics): N-gram is a sequence of terms, with the length of N (the length of the N-gram). ROUGE-N is an N-gram recall between a candidate summary and a set of reference summaries. ROUGE-N is computed using equation (14):

$$\text{ROUGE\_N} = \frac{\sum \text{Match\_N\_gram with references summarie}}{\sum \text{Match\_N\_gram with references summarie}} \quad (14)$$

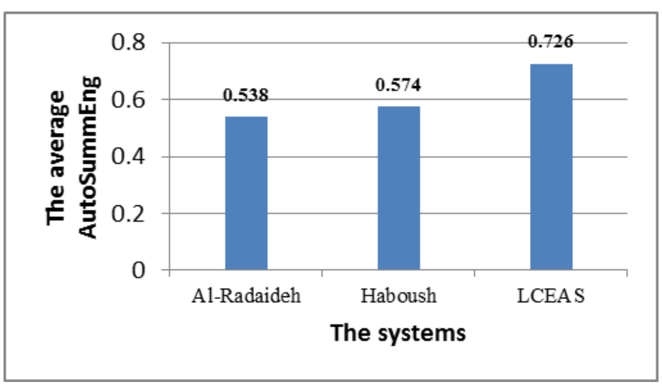

Figure 7. The average AutoSummEng of LCEAS, Al-Radaideh, and Haboush

Where, (match_N-gram) is the maximum number of N-grams co-occurring in a candidate summary C and a set of reference summaries R.

Example 1:

C_1:السيجارة هي منتج مصنوع من أوراق التبغ .

Alsegara heya montaj masnou' mn 'awraq altebegh.

R_1:قد تكون السيجارة من أوراق التبغ او من ساق النبات.

Qad takoon alsegara mn awraq altebegh aw mn saaq alnabat.

R_2:تحتوي السيجارة على أول أكسيد الكربون .

Tahtawi alsegara ala awal oxid alcarboun.

Table .4 shows the results of applying ROUGE-N (ROUGE-1 and ROUGE-2 between the candidate summary (C) and set of reference summaries (R1 and R2) for example-1.

1) ROUGE-L (Longest Common Subsequence (LCS)): The Longest Common Subsequence (LCS) of 2 given sequences *X* and *Y* is a common subsequence with maximum length. Reference [34] found the LCS of two sequences of length m and n using standard dynamic programming technique in O(mn) time. Recall is applied to calculate ROUGE-L using equation (15). LCS does not require consecutive matches, assuming the least N-gram value is 2. By considering example 1, table .5 shows the result of applying ROUGE-L between a candidate summary (C) and a set of reference summaries (R1 and R2).

$$\text{ROUGE\_L} = \frac{LCS(X,Y)}{m} \quad (15)$$

Where, LCS(X,Y) is the length of the longest common subsequence of X and Y, and *m* is the length of reference summary.

TABLE IV. APPLY ROUGE-1 AND ROUGE-2 BETWEEN CANDIDATE SUMMARY AND SET OF REFERENCES SUMMARIE

	$R_1 \cap C_1$	$R_2 \cap C_1$	Match R_1+R_2	Total / N-gram	Score match/ total
ROUGE-1	السيجارة، من، أوراق، التبغ alsegara, min, 'awraq, altebegh	السيجارة Alsegara	5	16	0.3125
ROUGE-2	alsegara min السيجارة من min awraq من أوراق awraq altebegh أوراق التبغ	No Matches	3	15	0.2000

TABLE V. APPLYING ROUGE-1 BATEEWN CANDIDATE SUMMARY AND REFERENCE SUMMRIES (R1, R2)

	LCS Words	LCS(X,Y)	M	ROUGE_L
R1	السيجارة، من، أوراق التبغ Alsegara, min, 'awraq, altebegh	4	10	0. 4
R2	No-Matches	0	6	0.0000

2) ROUGE-W (Weighted Longest Common Subsequence): ROUGE-W favors strings with consecutive matches. It can be computed efficiently using dynamic programming. Consider Example 2, C_1 is favored than C_2 because it has more consecutive LCSs.

Example 2

C_1: قد تكون السيجارة من أوراق التبغ او من ساق النبات

Qad takoon alsegara mn awraq altebegh aw mn saaq alnabat.

C_2: السيجارة تصنع من ساق النبات او اوراق التبغ.

Alsegara tosna' mn saaq alnabat aw awraq altebegh.

R_1: السيجارة هي من أوراق التبغ Alsegara heya mn awraq altebegh.

3) ROUGE-S: Skip-Bigram Co-Occurrence Statistics: In ROUGE-**S**, the S letter shows Skip-bigram which is any pair of words in their sentence order, allowing for random gaps.

Example 3

C_1: الشرطي قتل جميع اللصوص. Ashorti qatala jameea' allosoos.

R_1: الشرطي قتل اللصوص. Ashorti qatala allosoos.

C_1: has the following 6 Skip-bigrams

(الشرطي، قتل), (الشرطي، جميع), (الشرطي، اللصوص),(قتل، جميع) , (قتل، اللصوص), (جميع، اللصوص).

(ashorti, qatala), (ashorti,jameea'), (allosoos, ashorti), (jameea', qatala), (allosoos,qatala), (allosoos,jameea').

R_1: has the following 3 Skip-bigrams

(الشرطي، قتل), (الشرطي، اللصوص), (قتل، اللصوص).

(qatala, ashorti), (allosoos, ashorti), (allosoos, qatala).

R_1 has three skip-bigram matches with C_1.

For each evaluation, ROUGE generates 3 scores (Recall, Precision, and F-measure). The metrics used are ROUGE-2, ROUGE-L, ROUGE-W, and ROUGE-S as they worked well in single document summarization tasks [34]. Figs (8-11) show that LCEAS outperforms Al-Radaideh, and Haboush systems from the four ROUGE matrices point of views.

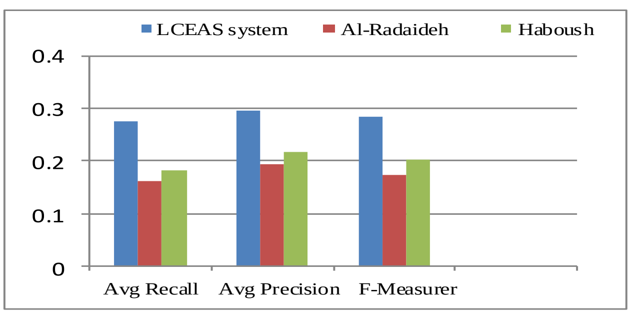

Figure 8. The average recall, precision, and F-measure using ROUGE-2

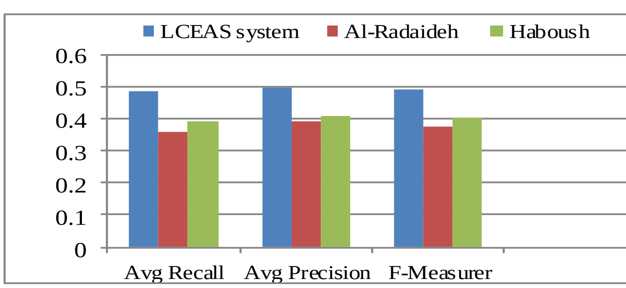

Figure 9. The average recall, precision, and F-measure using ROUGE-L

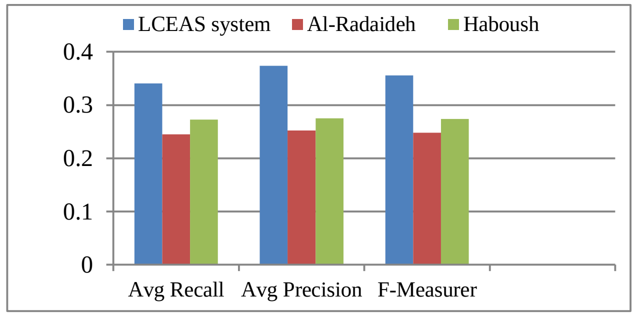

Figure 10. The average recall, precision, and F-measure using ROUGE-W

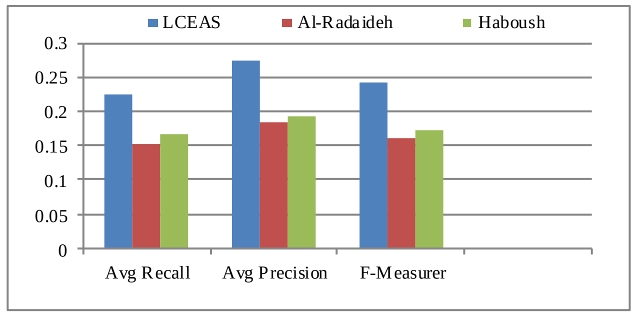

Figure (11) The average recall, precision, and F-measure using ROUGE-S

From Figs (7-11), the values of AutoSumEng and ROUGE shows that LCEAS has better results than Haboush and Al-Radaidah systems (i.e. the summarized text produced by LCEAS contain more significant sentences, and less redundancy).

V. Conclusion

In this paper, we developed Arabic text summarization system that contains the main topics in the text document without redundancy. Summarizing texts saves reading time, facilitates the document searches, improves the indexing efficiency, and serves the Question Answering systems. Two main procedures, lexical cohesion and text entailment, are used to develop the proposed text summarization system (LCEAS). Lexical cohesion is implemented to extract the main topics in text, while text entailment procedure is applied to reduce the redundancy in summary. LCEAS is tested and evaluated using two standard metrics AutoSumEng, ROUGE, in addition to human judge. Two datasets are used (Reuter's newswire and Essex Arabic Summaries Corpus (EASC)) for evaluation purposes. The experimental results show that the performance of LCEAS for Arabic text summarization highly improves the text summarization performance compared with the other Arabic text summarization systems. This improvement is due to eliminating the poor sentences in lexical cohesion with word senses disambiguation (WSD) segmentation and eliminating the redundant sentences. The AutoSumEng mean of LCEAS is 0.726, about 15% improvement compared with the best system (Haboush). The average recall, precision, and F-measure using ROUGE is improved about 8%-10% compared with Haboush system. Concerning text segmentation, WSD in corporation with lexical chains is used in LCEAS. It was found that using WSD in text segmentation algorithm improves recall, precision, and F-measure values of the segmentation process by 5%, 8%, and 7% respectively. Using lexical cohesion based segmentation improves Arabic text summarization performance. The results of human evaluation of LCEAS system showed highr performance in good and very good scales, compared with Al-Radaideh and Haboush systems. For further improvement, we suggest adding more semantic relations in the process of analyzing the sentences.

REFERENCES

[1] Lloret E, and Palomar M., "Challenging issues of automatic summarization: relevance detection and quality-based evaluation", Informatics, vol. 34, pp 29–35, 2010.

[2] Luhn H., "The automatic creation of literature abstracts," In IBM Journal of Research Development, vol.. 2(2):pp. 159–165, 1958.

[3] Das D. and Martins A., "A survey on automatic text summarization. literature survey for language and statistics", Carnegie Mellon University, 2007.

[4] Gupta V., Lehal G. S., "A survey of text summarization extractive techniques", Journal of Emerging Technologies in Web Intelligence, vol. 2(3), August 2010.

[5] Al-Sanie W., "Towards an infrastructure for arabic text summarization using rhetorical structure theory," Master thesis. Riyadh, Kingdom of Saudi Arabia, 2005.

[6] Sobh I., An optimized dual classification system for arabic extractive generic text summarization, MSc thesis, Giza, Egypt, 2009.

[7] El-Haj M., Kruschwitzc U., and Fox C., "Experimenting with automatic text summarization for Arabic", LTC'09 Proceedings of the 4th conference on Human language technology: challenges for computer science and linguistics, 2009.

[8] Sakhr company website: http//:www.sakhr.com. Last visited-January, 2015.

[9] Al-Radaideh Q., and Afif M., "Arabic text summarization using aggregate similarity", the international Arab conference on information Technology, 2011.

[10] Haboush A., Al-Zoubi M., Momani A., and Tarazi M., "Arabic text summerization model using clustering techniques", World of Computer Science and Information Technology Journal (WCSIT) ISSN: 2221-0741 vol. 2 (3), pp. 62 – 67, 2012.

[11] Azmi A., Al-Thanyyan S., "A text summarizer for Arabic", Sciencedirect, Computer Speech and Language, pp. 260–273, 2012.

[12] Al-Thwaib E., "Text summarization as feature selection for arabic text classification", World of Computer Science and Information Technology Journal (WCSIT), vol. 4(7), pp 101-104, 2014.

[13] Lloret E., Ferrández O., Munoz R., Palomar M., "A Text summarization approach under the influence of textual entailment", Natural Language Processing and Cognitive Science, Proceedings of the 5th International Workshop on Natural Language Processing and Cognitive Science, NLPCS 2008, In conjunction with ICEIS 2008, Spain, June ,2008.

[14] Tatar D., Mihis A., and Lupsa D., "Entailment–based Linear Segmentation in Summarization", International Journal of Software Engineering and Knowledge Engineering vol. 19(80, pp 1023–1038, 2009.

[15] Abu-Elkhair I., "Effects of stop words elimination for arabic information retrieval: a comparative study", International Journal of Computing and Information Science, vol. 4(3), pp 119-133, 2006.

[16] Dilekh T., and Behloul A., "Implementation of a new hybrid method for stemming of Arabic text", International Journal of Computer Applications, vol.46(8), 2012

[17] Galley M., and McKeown K., "Improving word sense disambiguation in lexical chaining", Proceeding in IJCAI'03 Proceedings of the 18th international joint conference on Artificial intelligence, 2003.

[18] Yllias, Chali, and Joty, Shafiq, "Word sense disambiguation using lexical cohesion", Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007), Association for Computational Linguistics, pp. 476–479, 2007.

[19] Pourvali M., Abadeh M., and Etemadian M., "Automated text summarization base on lexical chain and use of wordnet knowledge base", International Conference on Advancements in Information Technology With workshop of ICBMG 2011 IPCSIT vol.20 IACSIT Press, Singapore, 2011.

[20] Barzilay R., and Elhadad M., "Using lexical chains for text summarization", Proc. ACL Workshop Intelligent Scalable Text summarization, , pp. 10-17, 1997.

[21] Ercan G., and Cicekli I., "Lexical cohesion based topic modeling for summarization", Springer-Verlag Berlin Heidelberg, pp. 582–592, 2008

[22] Corley C., and Mihalcea R., "Measuring the semantic similarity of texts", Proceedings of ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment, ACL, 2005.

[23] Tatar D., Serban G., Mihis A., and Mihalcea R., "Text entailment as directional relation", Proceedings of CALP07 Workshop at RANLP2007, Borovets, Bulgaria, pp. 53-58, 2007.

[24] Fort F., "A flexible multitask summarizer for documents from different media, domain, and language", PhD thesis, 2008.

[25] Radev D., Hovy E., and McKeown K., "Introduction to the special issues on summarization", Computational Linguistics, 2002, pp 399–408.

[26] Spark-Jones K., and Galliers J., "Evaluating natural language processing systems: an analysis and review", Springer, 1995.

[27] Nenkova A., "Summarization evaluation for text and speech: issues and approaches", in Proceedings In. Interspeech, 2006.

[28] Pevzner L., and Hearst M., "A critique and improvement of an evaluation metric for text segmentation", Computational Linguistics, 2002.

[29] Brodic D., "Methodology for the evaluation of the algorithms for text line segmentation based on extended binary classification" , Measurement Science Review, vol. 11(3), 2011.

[30] http://www.lancaster.ac.uk/staff/elhaj/corpora.htm. Last visited in January,2015

[31] Giannakopoulos G., Karkaletsis V., Vouros G., and Stamatopoulos P., "Summarization system evaluation revisited: n-gram graphs", ACM Transactions on Speech and Language Processing, 2008.

[32] Giannakopoulos G., and Karkaletsis V., " Summarization system evaluation variations based on N-gram graphs", In TAC 2010 Workshop, Maryland MD, USA, November, 2010.

[33] El-Haj M., Kruschwitz U., and Fox C., "Using mechanical turk to create a corpus of arabic summaries", proceedings in the Language Resources (LRs) and Human Language Technologies (HLT) for Semitic Languages workshop held in conjunction with the 7th International Language Resources and Evaluation Conference (LREC 2010), pp 36-39, Valletta, Malta, 2010.

[34] Chin-Yew Lin, "Rouge: A package for automatic evaluation of summaries", Proceedings of the Workshop on Text summarization Branches Out (WAS2004),pp. 25–26, 2004.