IMPROVE THE AUTOMATIC SUMMARIZATION OF ARABIC TEXT DEPENDING ON RHETORICAL STRUCTURE THEORY

Ahmed Ibrahim [1], Tarek Elghazaly [2]

Department of Computer and Information Sciences,
Institute of Statistical Studies and Research,
Cairo University, Egypt
[1]a.ibr@live.com
[2]tarek.elghazaly@cu.edu.eg

Abstract— **this paper uses a semantic technique by adopting a Rhetorical Structure Theory (RST) for summarization purpose, to discover the most significant paragraphs based on functional and semantic criteria. However, the quality of RST summarization suffers when dealing with large documents. This paper proposes a new hybrid summarization model for Arabic text, which mingles two sub-models: The first sub-model produces a primary summary by using Rhetorical Structure Theory for identifying a range of the most significant parts of the text (the nucleus). Then the second sub-model ranks the significant parts in the primary rhetorical-summary based on the cosine similarity feature. To evaluate the proposed model, a prototype was developed on a range of articles, which have been classified into three groups different in size. The final output summary was evaluated in relation to its manual counterpart. In terms of enhancement of the rhetorical-summary precision, the experiment shows that proposed model HSM average precision is 71.6%, superior over the primary rhetorical-summary precision 56.3%.**

Keywords— **Arabic text summarization, Rhetorical Structure Theory, RST, Vector Space Model, VSM.**

I. INTRODUCTION AND PREVIOUS WORK

Automatic Text Summarization is one of the most difficult problems in Natural Language Processing (NLP) as in [1]. The key problem in the text automatic summarization process is that the target summarized text is incoherent and deviates from the context of the original text. This problem emerges when statistical techniques are used for summarization. Recent researches use different linguistic methods to treat these issues. Hammo in [1] presents a hybrid technique based on text structure and topic identification, Abdel Fattah in [2] proposes an approach based on several features to generate summaries. Also, Iraky in [3] present a technique to segment Arabic discourse into complete sentences based on RST. RST is a linguistically useful method for the summarization purpose, by extracting semantics behind the text. Al-Sanie in [4] attempts to develop an infrastructure for Arabic text summarization based on RST. Also, RST is used in [5] for the summarization purpose by identifying the rhetorical relationship between the paragraphs and extract the most significant paragraphs as a summary. Ibrahim and Elgazaly in [6] are providing a comparative study between the benefits of both the Rhetorical and Vector Representation in Arabic text summarization. Statistical results show that Rhetorical Representation is superior to Vector Representation. Moreover, the rhetorical summary keeps the text in context, without leading to lack of cohesion in which the anaphoric reference is not broken i.e. improving the ability of extracting the semantics behind the text. But the limitation of RST is appeared with huge text; also the experimental result's precision in [4] of the small-sized articles is 65%, where medium size is 62%. To overcome this limitation, this paper presents a hybrid model using two summarization techniques the (RST and VSM), to extract the most significant paragraphs, and evaluating the output summary in relation to its manual counterpart, in terms of enhancing the rhetorical-summary precision.

II. MODEL COMPONENTS

The proposed model consists of two sub-models, the first one is a semantic on RST and the second one is statistically based on VSM. A framework as shown in figure 1 illustrates the proposed model.

A. The First Sub-Model (RST)

The RST consists of four, the RST parsing process, rhetorical relationship identification process, building the RS-Tree and RST-Selector process.

1) RST Parsing Process Process: is designed to determine the elementary unit. The proposed model segment the original input text into paragraphs (as indicated by the HTML <p> tags). Paragraphs are extracted and segmented into sentence and words. Identifying sentences in Arabic is not an easy task [1]. This is mainly due to the morphological complexity of the language, missing punctuation marks (i.e. "."، "؛"), and the fact that sentences do not start with capital letters as in English. We solve the word segmentation problem of discriminating the Arabic cues. In more details, the problem is occurring when the writer adds a connector "cue" close to its successive word. For example, it discriminates connector "و" "waw" from the letter "و" "waw". In Arabic typing, the connector "و" "waw" is typed closed to its successive word, without separating them with a space. Looking at this example: "وقال" which means ("and said"), the connector "و" "waw" is directly typed before the verb "قال" which means ("said") without adding a space. The morphological analyzer is used to solve the word segmentation problem, by detecting the word-root and the word-type. The proposed model uses an

978-1-4799-2605-3/13 $31.00 © 2013 IEEE
DOI 10.1109/MICAI.2013.35

automatic Arabic part-of-speech tagger as in [7]. Khoja toolkit (It is based on a subset of the JavaScript Programming Language, Open Source license). If Khoja algorithm fails to detect the word-root and the word-type, the model uses the Xerox online service for easy access as in [8]. For more clarification, the following example (1) is a one of the BBC online news as shown in Table 1, and the article[1] titled is "Khamenei: We will face any attack from Israeli or U.S. strongly".

Table 1. The discourse of example (1)

Paragraph Number	Article Paragraphs Text
1	حذر مرشد الجمهورية الاسلامية الايرانية آية الله علي خامنئي الولايات المتحدة واسرائيل من مهاجمة بلاده، وقال إن الايرانيين سيواجهون هذه الهجمات "بقبضات من حديد." Ayatollah Ali Khamenei is warned , the United States and Israel from attacking his country, and said that the Iranians will face these attacks, "iron fists".
2	ونقل التلفزيون الايراني الرسمي عن المرشد قوله إن "على اعدائنا، وعلى وجه الخصوص النظام الصهيوني وامريكا وحلفائها، ان يعلموا اننا سنرد بحزم على اي تهديد او هجوم او حتى فكرة شن هجوم علينا." Iranian state television quoted the official leader as saying that "our enemies, and in particular the Zionist regime, America and its allies, should know that we will respond firmly to any threat or attack, or even the idea of launching an attack on us".
3	ومضى خامنئي للقول "سيرد الحرس الثوري والجيش والامة على اي هجوم بصفعات قوية وقبضات من حديد." And Khamenei went on to say "will refund the Revolutionary Guards and the army and the nation to any attack via strong slaps and iron fists".
4	وكانت التكهنات حول احتمال شن هجوم يستهدف منشآت ايران النووية قد تصاعدت في الاسبوع الماضي اثر قيام اسرائيل باختبار صاروخ بعيد المدى، وبعد التصريحات التي ادلى بها مسؤولون اسرائيليون وصفوا فيها برنامج ايران النووي بأنه "يشكل تهديدا مباشرا وخطيرا." The speculation about the possibility of an attack aimed at Iran's nuclear facilities has escalated in the last week following the creation of Israel a long-range missile test, and after the statements made by Israeli officials and described Iran's nuclear program as "showing a direct and serious threat".
5	كما ارتفعت حدة التوتر يوم الثلاثاء المنصرم بعد قيام الوكالة الدولية للطاقة الذرية بنشر تقرير قالت فيه إن لديها معلومات تفيد بان ايران اجرت تجارب واختبارات "ذات صلة بتطوير سلاح نووي". As tensions rose last Tuesday after the International Atomic Energy Agency (IAEA) published a report in which it said it had information that Iran has conducted experiments and tests "relevant to the development of a nuclear weapon".
6	وادانت ايران ما جاء في التقرير، واصفة اياه "بغير المتوازن" وبأنه "يخفي دوافع سياسية." Iran condemned the statement in the report, describing it "Without a balanced," and as a "to hide a political motives".

[1] Published in http://www.bbc.co.uk/arabic/middleeast/2011/11/111110_iran_nuclear_khamenei.shtml, at: 11/10/2011.

2) Rhetorical Relationship Identification Process: is considered the main process of applying RST between text spans. This process can match the connector between paragraphs based on the predefined Arabic rhetorical relations, in order to identify the relationship between paragraphs. We are collecting the Arabic rhetorical relations collecting the Arabic rhetorical relations from different resources. Al-Sanie in [4] provides a list of rhetorical relations and cue phrases. In addition, Hemida in [9] presents a synthesis theory for the Arabic sentences and provided a part of the list of the Arabic rhetorical relations as shown in Table 2. The rhetorical relationship identification process uses the predefined Arabic rhetorical relations to classify paragraphs into the nucleus or satellites, as shown in the following conditions:

> *If a paragraph P_n begins with any of the defined cues, then there is a direct relation with the previous paragraph P_{n-1}. Therefore, paragraph P_n is called a **satellite**.*
>
> *Else*
>
> *If the paragraph P_n begins without any of the defined cues, then there is a relation to the first paragraph (root). In this case, the paragraph P_n is called a **nucleus***

In this study, we use a universal data structures JSON (JavaScript Object Notation) to produce the text structure based on the previous conditions. JSON is a text format that is completely language independent, and it's an ideal data-interchange language. It provides significant performance gains over XML, JSON parse faster than one hundred times than XML in modern browsers, as a comparison in [10]. The following is a JSON code for example (1):

var json = "{id:"Root Node", name:"1", data:{}, children:{id:"node 2" , name:"2", data:{}, children:[{id:"nodeP22",name:"2",dta:{},children:[]}, {id:"nodeP13",name:"3",data:{},children:[{id:"nodeP23", name:"3",data:{},children:[]},{id:"nodeP14", name:" 4 ",

Table 2. Part of Rhetorical Relation as Defined in (Hemida , 1997)

Relations		Cues											
Justification	السببية و التعليل	اي	كي	بمعنى	لان	نستنتج	لذا	نتيجة	بسبب	حيث	بفضل	اذن
Exception	الاستثناء	لاسيما	بيد	حاشا	خلا	عدا	ليس	غير	ألا			
Antithesis	نقض	لكن	رغم	إنما	لا	بل	لكن					
Joint	علاقة معنوية	أما	أو	الفاء	الواو	أم						
Condition	شرط	لو	إذا	إذ								
Timing	الزمنية		بعد	قبل								
Sequence	ترتيب		ثم	حتى								

data:{}, children: [{id:"nodeP24",name:"4",data:{}, children:[]}, {id:"nodeP15",name:"5", data:{}, children: [{id:"nodeP25",name:"5", data:{},children:[]},{id:"nodeP16", name:" 6 ", data:{},children:[]}]}]}]}, {id:"End root node\", name:\"1\", data:{},children:[]}]}".

2) Building the RS-Tree Process: the abstract structure of most texts is a tree. Most articles and text theories explicitly or implicitly mention that trees are good mathematical abstractions as in [11]. After the units are identified and the rhetorical relations between units are recognized, now the proposed model can represent and building a rhetorical structure by using a binary tree topology called the rhetorical structure tree (RS-Tree). According to [12] the main goal of data visualization is to communicate information clearly and effectively through graphical means. In this study, we use the JavaScript InfoVis Toolkit [2] , an open source license; it provides tools for creating a graphical Data visualization. The RST Visualizing process has required developing this toolkit to be a convent to visualize RS-Tree. First we adepts, the Space-Tree model toolkit to be convened to visualize RS-Tree topology (binary tree). Then, we render a JSON code, to visualize RS-Tree, as shown in Fig1.

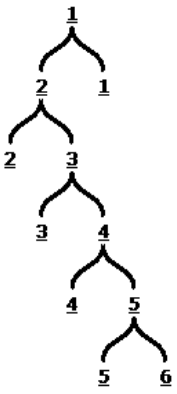

Fig.1 Rhetorical Structure Tree for Example 1

3) RST Selecting Process: The final step, is selecting the nucleus paragraphs depending on the relationship to root node (P_1). The significant paragraphs (the nucleus nodes) as shown in Figure 1 are P_1 and P_2.

B. The Second Sub-Model (VSM)

Thus far, Section 2.1 is focused on the primary sub-model, which can extract the significant paragraphs based on RST. Now turn to a second sub-model, which uses the Vector Space Model. It is a popular choice in many applications because of its simplicity and robustness. It is fundamental to a host of information vector space model retrieval operations ranging from scoring documents on a query, document classification and document clustering [13]. Vector space model or term vector model is an algebraic model for representing text documents and any objects in general as vectors. In this paper, we are representing the article parts (title, paragraphs) as vectors. For scoring comparable weights, long and short paragraphs should be normalized then computing the cosine similarity score and selecting the top as in [13].

1) Vector Representation Process: is responsible for transfer text parts *(title, paragraphs)* to vectors based on the *tf.idf* weights. The *tf.idf* weights scheme assigns to term *t* a weight in paragraph *p*. The *tf.idf* is a composite weight combines the definitions of term frequency *tf* and inverse document frequency *idf*

2) Length-Normalize: For scouring comparable weights, the long and short paragraphs should be normalized by dividing each of its components by its length. The normalization process is used for each title vector *V(t)* and each paragraph vector $V(P_s)$ to have complete length normalization process. The set of paragraphs in one article will be viewed as a set of vectors in a vector space, in which there is one axis for each term. Table 3, illustrates the Term-Paragraph Matrix as a collection of *N* paragraphs as a collection of vectors leads to a neutral view of an article as a term- paragraph matrix this an $M \times N$ matrix whose rows represents the M terms (dimensions) of the *N* columns, each of them corresponds to a paragraph.

[2] http://philogb.github.com/jit/ last visit at 11/10/2012

Table 3. Term-Paragraph Matrix in Example 1, After Length-Normalization

V(P)	*LN(V(P)*	اسرائيل	حديد	خامنئي	قبض	هجم	وجه	امريكا	**SUM**
P_1	1.09	0.43	0.43	0.43	0.43	0.20	0.43	0.00	2.36
P_2	0.53	0.00	0.00	0.00	0.00	0.42	0.89	0.00	1.30
P_3	0.84	0.00	0.56	0.56	0.56	0.20	0.00	0.00	1.87
P_4	0.65	0.96	0.00	0.00	0.00	0.26	0.00	0.00	1.22
P_5	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
P_6	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

3) Cosine Similarity: is the cosine of the angle θ between vectors (each paragraphs $V(P_n)$, with title $V(t)$), as shown in Fig.2, by using the equation 1

$$Cosine\ Similarity(V_t, V_p) = \frac{V_t \cdot V_p}{|V_t|\ |V_p|} = \frac{V_t}{|V_t|} \cdot \frac{V_p}{|V_p|} = \frac{\sum_{i=1}^{|V|} t_i\ p_i}{\sqrt{\sum_{i=1}^{|V|} (t_i)^2}\ \sqrt{\sum_{i=1}^{|V|} (p_i)^2}}$$

Where:

V_t is the *tf•idf* weight of term i in the title.
V_p is the *tf•idf* weight of term i in the paragraph.

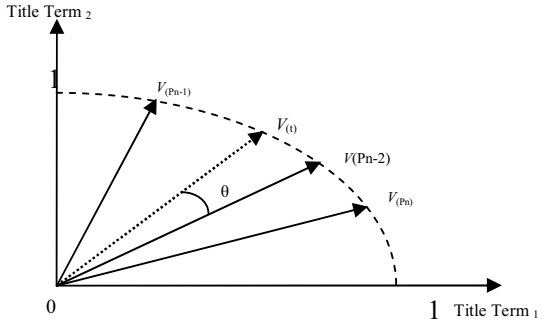

Fig. 2 Vector representation and Cosine Similarity after length-Normalization

.

4) Ranking Paragraphs: in the example (1) the paragraph P1 is the top-scoring paragraph in article title (query) with a score of 0.997. The second rank is paragraph P_3 with a score of 0.759, and the third rank is paragraph P_2 with a score of 0.449.

III. The Evaluation

The test set is extracted using the Really Simple Syndication is known as RSS. It provides an RSS-document (which is called a "feed", "web feed" or "channel") includes full or summarized text, plus XML metadata such as publishing dates and authorship. In this study, the RSS reader uses BBC online Arabic news portal, to extract fully news articles. Articles include general news, political news, business news, regional news, entertainment news, niche-oriented news (health, science and technology), crime reporting, and sensationalism. The average paragraphs of the article are five and the average words in the paragraph are 24 words. The test set is classified into three groups: small-sized articles (1-10 paragraph), medium-sized articles (11-20 paragraph), and large-sized articles (21-40 paragraph). The overall Figures of the test set are illustrated in Table 4. The Precision is used to assess the performance of the rhetorical summary result. The precision is simply the division of the number of paragraphs that were correctly identified by the proposed model result as being important, over the total number of important paragraphs in the manual summarization as in [4]. The precision is based on the result of experiments illustrated in Fig3. The Y-axis represents the precision, and the X-axis represents the text size groups based on article size (paragraphs number P_n). The proposed model improves the output summary precision of the medium sized articles' group from 55% to 71%. And at the large sized articles' group from 39% to 69%.

Table 4. Statistics the Test Set

Category	**Figures**
Corpus Textual Size	25.06 MB
No. of Articles	212
No. of Paragraphs	2249
No. of Sentences	2521
No. of words (exact)	66448
No of Word (root)	41260
No. of Stop word	15673
No. of small sized articles (Less than 10 paragraphs)	104
No. of medium sized articles (10 - 20 paragraphs)	79
No. of large sized articles (More than 20 paragraphs)	29

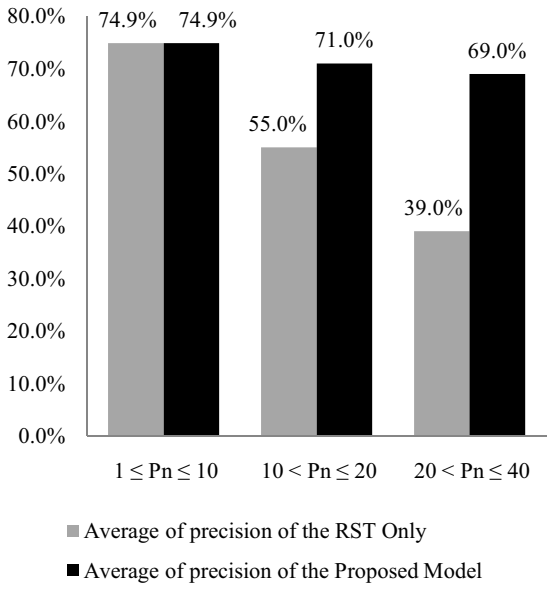

Fig.3 Precision with Different Articles Size

IV. Conclusion

The experiments show that the proposed hybrid model for Arabic text summarization combining RST and VSM is able to take the advantages of both. As the proposed model uses VSM for ranking text spans based on the cosine similarity feature, the average precision of the hybrid model's output summary is 71.6% which was 56.3% by using RST only. Furthermore, the proposed model has still the advantage of RST of improving the ability of extracting the semantics behind the text. Consequently, it produces an output summary not out of context, without the lack of cohesion and with non broken anaphoric references.

References

[1] B. H. Hammo, H. Abu-Salem, and M.W. Evens, "A Hybrid Arabic Text Summarization Technique Based on Text Structure and Topic Identification". In Proceedings of Int. J. Comput. Proc. Oriental Lang. pp 39-65. DOI: 10.1142/S1793840611002206. 2011.

[2] M. Abd-Elfattah, and R. Fuji. "Automatic text summarization". Proceeding of world academy of science, engineering and technology, pp. 192–195. Cairo, Egypt. 2008.

[3] K. Iraky, F. Zakareya and F. Abdelfatah. "Arabic Discourse Segmentation Based on Rhetorical Methods", International Journal of Electric & Computer Sciences IJECS-IJENS, vol: 11, no: 01, pp. 10-16. 2011.

[4] W. Al-Sanie, A. Touir and H. Mathkour. "Towards an infrastructure for Arabic text summarization using rhetorical structure theory". M.Sc. Thesis, King Saud University, Riyadh, Saudi Arabia. 2005.

[5] A. Ibrahim and T. Elghazaly. "Arabic text summarization using Rhetorical Structure Theory". Informatics and Systems (INFOS), 8th International Conference, vol., no., pp.NLP-34,NLP-38. 2012.

[6] A. Ibrahim and T. Elghazaly. "Rhetorical Representation and Vector Representation in Summarizing Arabic Text". Natural Language Processing and Information Systems. Lecture Notes in Computer ScienceVolume 7934, 20s13, pp 421-424, 2013.

[7] S. Khoja. "An Automatic Arabic Part-of-Speech Tagger". UK. Doctoral Thesis, Computer Science Department, University of Lancaster. 2003.

[8] K. R. Beesley. and L. Karttunen. "Finite State Morphology", Palo Alto, CA: CSLI Publications. 2003.

[9] M. Hemida, "Nzam alartbat WA alrabt fi trkip algomla alarabia". Egypt: Dar Nubar for printing, ISBN 977-16-0235-7. 1997.

[10] N.Nurseitov, M. Paulson, R. Reynolds, and C. Izurieta, "Comparison of JSON and XML Data Interchange Formats: A Case Study". CAINE'09. pp. 157–162. 2009.

[11] D. Marcu. "Discourse trees are good indicators of their importance in text". In Advances in Automatic Text Summarization, pp123–136. Cambridge, MA: The MIT Press.1999.

[12] V. Friedman. "Data Visualization and Infographics" in: Graphics, Monday Inspiration, 2008.

[13] C. D. Manning, P. Raghavan and H. Schütze. "An Introduction to Information Retrieval". Cambridge University. Press New York, USA: Cambridge University Press,ISBN:0521865719 9780521865715. 2008.