A New Approach for Arabic Text Summarization

Samira Lagrini
Labged Laboratory, Computer Science Department,
Badji Mokhtar University, P.O. Box 12,
Annaba, 23000, Algeria

samira.lagrini@univ-Annaba.dz

Mohammed Redjimi
LICUS Laboratory, Computer science Department
Université 20 Aout 1955, Skikda,
21000, Algeria

medredjimi@gmail.com

Abstract

Due to the increasing number of online textual information, acquiring relevant information quickly has become a challenging task. Automatic text summarization (TS) offers a powerful solution for the quick exploitation of these resources. It consists of producing a short representation of an input text while preserving its relevant information and overall meaning. Automatic text summarization has seen a great attention for Indo-European languages. However, for Arabic, researches in this field have not yet attaint a notable progress. Most of the existing approaches in Arabic text summarization literature rely mainly on numerical techniques and neglect semantic and rhetorical relations connecting text units. This affects negatively the global coherence of the generated summary and its readability. In this paper, we attempt to overcome this limitation by proposing a new approach that combines a rhetorical analysis following the rhetorical structure theory (RST) and a statistical-based method. The proposed approach relies on exploiting rhetorical relations linking text units to generate a primary summary, which will be pass by a second phase where a statistical processing is applied in order to produce the final summary. Evaluation results on Essex Arabic Summaries Corpus (EASC) using ROUGE-N measures are very promising and prove the effectiveness of the proposed approach.

1. Introduction

Automatic text summarization is a fundamental task in Natural Language Processing (NLP). It consists of producing a brief representation of an input text covering its relevant content and overall meaning. This allows researcher acquiring needed information with minimum effort and accurately exploiting available resources.

The idea of designing text summarization systems dates back to 1950s (Luhn, 1958; Baxendale, 1958; Edmundson, 1969) in order to satisfy the first needs in term of automatic text summaries. But this need has become even more excessive with the advent of the Internet and the exponential increase of textual information in electronic format. This situation has sparked a great attention within the Natural Language Processing (NLP) community. Many researchers have fully invested in this field and a lot of research works have been devoted to produce automatic text summarizers in different languages. However, up to date there are several problems without effective solutions.

Generally, producing automatic text summaries can be done following two main paradigms: extractive summarization and abstractive summarization. In abstractive summarization, producing a summary involves an in-depth analysis of the source text in order to select the relevant content and to product the summary (abstract) using other words not necessary presented in the source text. This requires many advanced linguistic resources for text representation, sentences fusion and automatic text generation. However, in extractive summarization, relevant sentences in the source text are selected and directly assembled without any reformulation to produce the summary. Compared to abstractive summarization, extractive approaches are simple to implement and require only certain linguistic aspects. This is why most researches in this field focuses on extractive text summarization. The approach developed here is also based on extractive Arabic text summarization.

Several extractive summarization approaches have been developed to date to produce Arabic extracts including numerical, linguistic and hybrid methods. Numerical methods rely on computational values or a

statistical distribution of particular features to judge the relevance of textual segments including statistical methods (Douzidia, and Lapalme, 2004; Alotaiby et al., 2012), supervised learning based methods (Sobh et al., 2006; Boudabous et al. 2010; Belkebir and Guessoum, 2015; Qaroush et al., 2019), clustering based methods (El Haj et al., 2011; Oufaida et al., 2014; Waheeb et al., 2020; Alqaisi et al., 2020; Alami et al., 2021) and graph based methods (Alami et al., 2017).
Linguistic methods rely on semantic relations or discursive structure to assess the relevance of sentences in the text (Kumar et al., 2016). The hybrid method is a combination of the two former methods used to produce summary (Azmi and Al-Thanyyan, 2012; Al Khawaldeh and Samawi, 2015; Azmi and Altmami, 2018).
By analyzing Arabic text summarization literature, we can notice that most of reported research rely on numerical methods based on traditional bag of word representation and don't take into account semantic and rhetorical relations linking textual units. This affect the global coherence of the generated summary and its readability.

In this research we try to overcome this limitation by proposing a new approach that combine a rhetorical analysis following the rhetorical structure theory (Mann and Thompson, 1988) and a statistical method. The proposed approach rely on exploiting rhetorical relations linking elementary discourse units to generate a primary summary, which will be pass by a second phase where a statistical processing is applied in order to produce the final summary.

The remainder of this paper is as follows: the second section copes with a general overview of the rhetorical structure theory (RST). In Section 3, the proposed approach will be described. In section 4, the results and evaluation of the proposed approach are presented, and finally the conclusion and future works are addressed in section 5.

2. Rhetorical structure theory

The Rhetorical structure theory (RST) (Mann and Thompson, 1988) is a prominent theory in discourse analysis (Taboada and Mann, 2006). It focuses on rhetorical analysis, which aims to represent an input text in a hierarchical form of rhetorical relations linking its text units. In the RST framework, if two non-overlapping atomic textual units called elementary discourse units (EDUs) are linked via a discourse relation (also called rhetorical or coherence relation), they constitute together another discourse unit which can in turn participates in another discourse relation (Mann and Thompson, 1988). Under a full analysis, a coherent text can be represented as a labelled tree, called discourse tree (or RST-tree).

The Rhetorical analysis in RST framework involves three tasks:

- Segmenting the text into elementary discourse units.
-Identifying discourse relations between adjacent discourse units.
- Linking all discourse units into labelled trees (RST-trees).

Figure 1 shows a sample of an RST-tree for the following text segment consisting of two sentences segmented into five EDUs.

[*The impact won't be that great,*]$_1$ [*said Graeme Lidgerwood of First Boston Corp.*]$_2$ [*This is in part because of the effect*]$_3$ [*of having to average the number of shares outstanding,*]$_4$ [*she said.*]$_5$

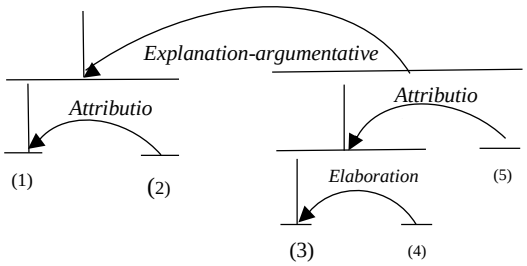

Figure 1. Example of an RST- tree for two sentences in RST–DT (Carlson et al., 2003)

Discourse Units (EDUs or larger discourse unit) linked via rhetorical relations are assigned a nuclearity attribute 'Nucleus' or 'satellite' depending on their relative importance in the text. The 'Nucleus' expresses what is more important for the author purpose, while the satellite provides a secondary information. In Figure 1, the 'Nucleus' are denoted by vertical line, Horizontal lines indicate discourse units, the Satellites are linked to their nucleus by curved arrows.

Rhetorical relations can be either '*mononuclear*' when they connect two discourse units having different status: 'Nucleus' and 'satellite', or '*multi-nuclear*' linking discourse units of equal importance, all nucleus.

The authors of RST defined a set of 24 relations, including 21 mononuclear relations. This set was extended by (Carlson et al., 2003) to 78 relation grouped into 16 class, which allows a high level of expression.

3. Proposed approach

In this research, we propose a new approach for Arabic single document summarization that combines rhetorical structure theory (RST) and a statistical-based method. Our aim is to exploit rhetorical relations linking text units in order to produce a coherent extracts. To this end, a rhetorical analysis is firstly performed to produce a primary summary relying mainly on rhetorical relations that exist between elementary discourse units (EDUs) rather than the discourse structure of the text. Then each sentence within the primary summary is assigned a score based on some statistical and linguistic features. Sentences having the best score will be selected to produce the output summary. Thus, the production of the summary go through two main phases; rhetorical analysis phase and statistical processing phase. Each phase consist of three main steps as shown in Figure 2.

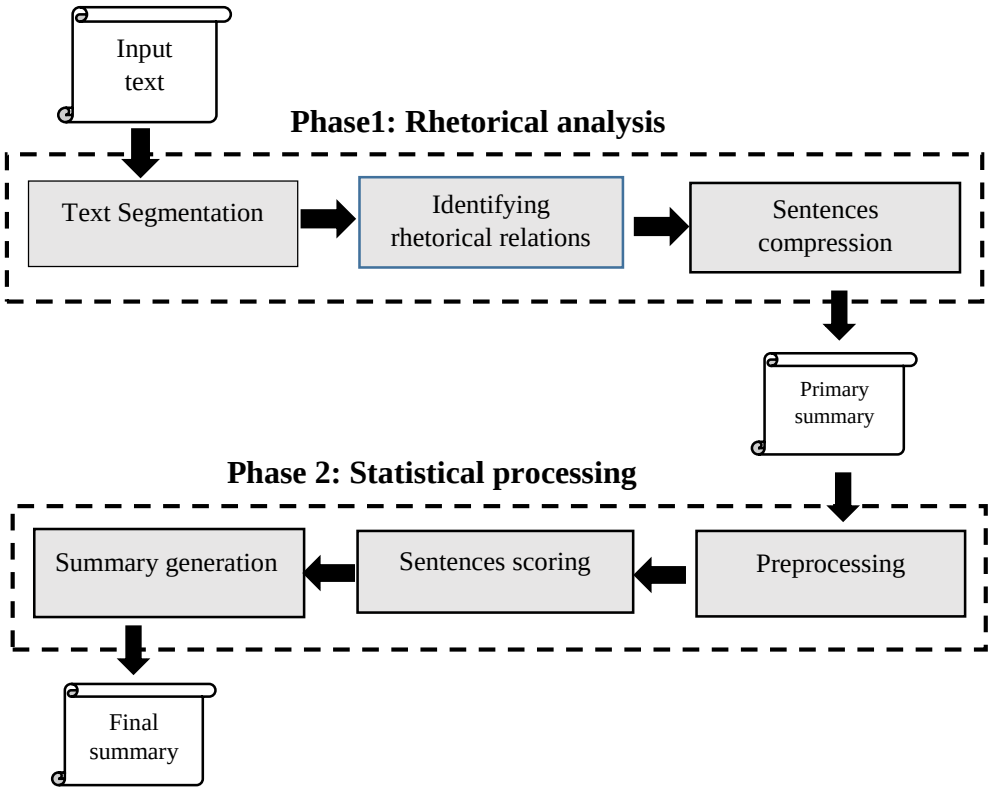

Figure 2: Proposed approach main steps.

3.1. Rhetorical analysis phase

The rhetorical analysis of the text includes the following subtasks: text segmentation, identifying rhetorical relations, and sentences compression.

3.1.1 Text segmentation

Text segmentation consists of breaking the text into non overlapping clauses called

elementary discourse units (EDUs). In our segmentation rules, an EDU can be a verbal or a nominal clause that begin with a discourse markers. Thus, segmenting the text into EDUs is performed as follow:

- First, segmenting the text into sentences. A sentence is defined as a textual passage delimited by (.).
- Then segmenting each sentence into EDUs based on Arabic discourse markers and a set of segmentation rules.

Arabic discourse markers such as (' وبالتالي'/therefore,' حيث'/ where,' حتى'/even, 'كذلك'/ as well as, 'بالرغم'/though) are already defined in a rich lexicon during the annotation process of our Arabic RST annotated corpus (Lagrini et al., 2019).

The following example presents a sentence segmented into four EDUs (between brackets), discourse markers are written in red bold.

[أكد سفير دولة فلسطين] $_1$[**ان** الأوضاع الصعبة التي يعيشها الشعب الفلسطيني]$_2$ [**هي نتيجة** الفرقة التي يشهدها الشارع الفلسطيني]$_3$[**و**تراجع الموقف العربي عن نصرة القدس]$_4$

[The ambassador of Palestine state confirmed $_1$[that the difficult circumstances that the Palestinian people live in,]$_2$ [are the result of the division seen in Palestinian street]$_3$[and the retreat of the Arab people on the support of Jerusalem]$_4$

3.1.2 Identifying rhetorical relations

Identifying rhetorical relations between two text segments has been shown to be useful in many Natural Language Processing tasks. Discourse markers or discourse connective such as: *because, although, since, but …, etc* strongly indicate the sense of explicit relations. For example 'because' is a strong indicator for causal relation. However, in the absence of such connectives the relation is called implicit and identifying such relation is still a big challenge. In our summarization process, we focus on identifying explicit relations and more precisely fine-grained relations between two adjacent elementary discourse units within the same sentence. In our previous work (Lagrini et al., 2019a), we have already defined a set of 23 fine-grained Arabic relations that can hold between two adjacent EDUs at the intra- sentential level. These relation are grouped into seven classes: /causal, المقارنة /comparison, العطف' /joint, تفصيل /elaboration, توضيح /explanation, اسناد/ attribution, الشرط /conditional. For more details see (Lagrini et al., 2019a).

In the RST framework, fine-grained relations are enriched with nuclearity annotation:' SN', 'NS 'for mononuclear relations and 'NN' for multinuclear relations. These notations specify the rhetorical status of the connected discourse units. Taking as an example the following sentence composed of two EDUs:

[العلم يتقدم نافيا ما سبقه] [بمعنى ان اخر ما ينجزه العلم هو الأكثر صحة]

[Science advances in denial of what preceded it]$_1$[that's mean that the last thing that science accomplishes is the most correct]$_2$

These two EDUs are linked by the fine grained relation ' تفسير' /' explanation/NS' signaled by the discourse marker 'بمعنى ان'/ that's mean'. The notation NS attached to the name of the relation means that the first EDU is the most important segment (Nucleus) denoted by N. while the second is the satellite, (denoted by S). It provides an optional information about the nucleus.

The nuclearity annotation attached to the name of relations is very interesting in our work, since it provides us with information about the relative importance of linked EDUs.

To automatically identify fine-grained rhetorical relations between adjacent EDUs, we have used a multi-class supervised learning approach based on a multi-layer perceptron model (Lagrini et al., 2019). We proceeded as follows:

- For each pair of adjacent EDUs, a feature vector is computed. We used ten group of lexical and semantic features. See (Lagrini et al., 2019) for a detailed description of used features.
- Then, all features vectors are fed as input to the model in order to predict fine-grained relations classes. Figure 3 summarizes this process.

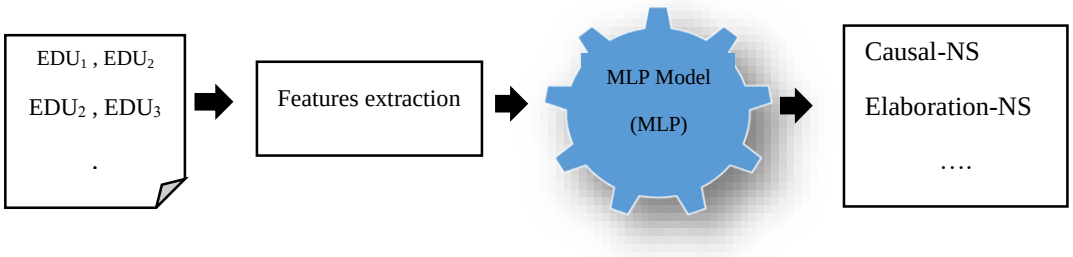

Figure 3: Identifying rhetorical relations process

3.1.3 Sentences compression

Once rhetorical relations between each pair of EDUs were identified, we proceeded to sentences compression. This step consists of removing satellite EDUs from each sentence while taking into consideration its overall coherence.

Sentence compression relies not only on the rhetorical status of its constituents EDU, but also on rhetorical relations. Some relations such as: تمثيل/example-NS, تشبيه/simile-NS, مقابلة/contrast-NN are not useful for the summary task. Thus, the presence of such relations involves the deletion of its constituents EDUs even if these EDUs are nucleus. Consider as an example the following sentence:

[ولقد دفع ذلك التوتر هذا البلد الى شن غارات قاتلة على جيرانه]$_1$[مما ولد رد فعل قاس لدى السلطات العسكرية بزعامة ضباطها الذين، كانوا يحاولون تهدئة الأوضاع]$_2$ [مراعاة لخاطر الاميركيين من جهة،]3[وتأجيلاً للانفجار المحتمل من جهة ثانية] $_4$.

This tension prompted this country to launch deadly raids on its neighbors (1) which generated a harsh reaction on the part of the military authorities led by its officers, who were trying to calm the situation, (2) taking into account the dangers of the Americans on the one hand, and (3) postponing the possible explosion on the other hand (4)

The sentence is composed of four EDUs linked by the following fine-grained relations:

Relation (1,2)= نتيجة/ result-NN;
Relation (2,3)=غاية /purpose-NS;
Relation (3,4)=وصل / joint-NN

Sentence compression involves removing EDU3 because it is a segment satellite and removing EDU4 since it is linked by a multinuclear relation with EDU3. That's means, they have the same level of importance. Therefore the compressed version of the sentence is as follow:

ولقد دفع ذلك التوتر هذا البلد الى شن غارات قاتلة على جيرانه مما ولد رد فعل قاس لدى السلطات العسكرية بزعامة ضباطها الذين، كانوا يحاولون تهدئة الأوضاع .

Compressing all sentences in the input text results in an abbreviate version of the source text which we consider as a primary summary.

3.2. Statistical processing phase

The goal of this phase is to reduce the number of compressed sentences in the primary summary and selecting the most relevant ones to produce the final summary. This phase includes the following subtasks: preprocessing the primary summary, sentence scoring, and summary generation.

3.2.1 Preprocessing of the primary summary

Preprocessing consists of three sequenced steps: tokenization, stop-words removal, and stemming.

Tokenization: consists of segmenting an input text into paragraphs, sentences, and words called tokens (Attia, 2007). As the primary summary is already segmented into sentences, AraNLP tool (Althobaiti et al., 2014) was used to segment each sentence into list of tokens.

Stop-words removal: Stop words are non-informative words that are frequently used in the text such as conjunctions, pronoun, prepositions, .. etc. They serve only a syntactic function but not indicate any relevant information. Removing these words is necessary to avoid affecting words weighting process (El-Khair, 2006). In our system, we have used the general stop-words list of AraNLP tool (Althobaiti et al., 2014) containing environ 168 words.

Stemming: Stemming is a morphological technique that consists of reducing inflected words to their stem or root by removing affixes

attached to them. For example the words ' استعمل' 'استعمالات' 'عامل' ' can be stemmed to word 'عمل'. For Arabic language, there are two main approaches for stemming: *Light-Based Stemming* and *root based stemming*. Following a comparative study between these two approaches regarding text summarization (Alami et al., 2016), it has been shown that root based stemming performs better than light stemming. This is why we choose to use in our summarizer khoja stemmer (Khoja and Garside, 1999) as a root based stemmer.

3.2.2 Sentence scoring

After preprocessing, each sentence was assigned a score based on certain features to assess its relevance. In text summarization literature, several features have been explored including key terms, indicative phrases, sentence position, sentence cohesions ... etc. In our summarization method, we used the following features: sentence position, sentence length and title similarity. These features have been successfully used in several works reported in Arabic summarization literature (Al-Radaideh and Bataineh, 2018; Al-Abdallah and Al-Taani, 2017; Douzidia and Lapalme, 2004; Fattah et al., 2009).

Title Similarity: As the title usually covers the main topic covered in the text, title words can be considered as key terms. Therefore, sentences that contain title words are relevant sentences and should be included in the final summary. Title similarity score assigned to each sentence is a function of the co-occurrences of title words in the sentence. This score is calculated using the following formula:

$$\boldsymbol{T}itle - sim(Si,T) = \frac{\text{title words} \cap \text{sentence (Si)}}{\text{title words}} \quad (1)$$

Sentence position: Sentence position in the text can be a good indicator that reflects its degree of importance. Generally in news articles, relevant sentences are either located at the beginning of the document or at the end. This is why we consider the first and the last sentence in the primary summary very important and should be included in the final summary. The position score assigned to each sentence is calculated as follows:

$$pos(Si) = \begin{Bmatrix} 1 & if\ i = 1\ or\ i = N \\ 0.5 & otherwise \end{Bmatrix} \quad (2)$$

With:
i: sentence number
N: total number of sentences within the input text.

Sentence length: As the majority of sentences in the primary summary only contain nucleus EDUs, we can say that the longer sentences are those which are most likely to contain more relevant information. Thus the score assigned to each sentence based on its length (the length in terms of words) is calculated as follows:

$$length(Si) = \frac{N(Si)}{N(Sl)} \quad (3)$$

With:
N (Si): number of words in the sentence Si
N (SL): number of words in the longest sentence in the primary summary.

The final score of each sentence is a linear combination of its scores assigned for each feature. This score represents the degree of relevance of the sentence in the primary summary. It is calculated using the following formula:

Score (Si) = Title-similarity (Si, T) + length (Si) + position (Si) *(4)*

3.2.3 Summary generation

In this phase, Sentences were ranked in descending order according to their final scores. Best scored sentences were then selected to produce the final summary. The selected sentences were assembled and arranged according to their order of appearing in the primary summary. The number of selected sentences depends on user's compression rate.

4. Evaluation and results

To evaluate the performance of our system we relied on intrinsic evaluation. Such evaluation seeks to evaluate automatic summaries based on their forms and contents. Content assessment measures the ability of the system to identify relevant sentences from the source document, this can be done automatically by comparing the generated summaries with reference summaries produced by human experts.

4.1 Evaluation dataset

For automatic evaluation we used Essex Arabic Summaries Corpus (EASC) (El Haj et al., 2010). The corpus consists of 153 documents extracted from Wikipedia and two Arabic newspapers: ALRai and Alwatan, covering ten different topics: art and music, education, environment, finance, health, politic, religion, science and technology, sport and tourism.

For each document in EASC, five reference summaries produced by humans are available. That is, the corpus is composed of a 765 reference summaries whose size does not exceed 50% of the size of the source text. EASC is available online with two encodings: UTF-8 and ISO-Arabic.

To evaluate our system, we selected a collection of 40 news articles from EASC corpus with their fives references summaries.

4.2 Evaluation measures

We used ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric (Lin, 2004) to evaluate our system. ROUGE is an automatic evaluation method that assess the quality of generated summary by comparing its content against one or more reference summaries. This comparison can be made by computing overleaping words such as Ngram (ROUGE-N) or word pairs (ROUGE-S and ROUGE-SU) or word sequence (Rouge-L, ROUGE-W)

ROUGE-N calculates the number of overleaping N-grams (N successive words) between the machine summary and reference summaries. Different metrics can be used such as ROUGE-1 (unigrams) , ROUGE-2(bi-grams), ROUGE-3 (trigrams)..etc.

In our system evaluation we used two metrics: ROUGE-1 and ROUGE-2 of ROUGE-N. For each metric, the precision, recall and F-score are calculated in order to provide a complete information about the system.

Recall: indicates the coverage of the system. It is calculated using the following formula:

$$Recall = \frac{nomber\ of\ overloapping\ N-grams}{nomber\ of\ N-grams\ in\ the\ set\ of\ refrence\ summaries} \quad (5)$$

Precision: Indicates the accuracy or the exactitude of the system. It is calculated as follow:

$$precision = \frac{nomber\ of\ overloapping\ n-grams}{nomber\ of\ \ n-grams\ in\ generated\ summary} \quad (6)$$

F-score: combines precision and recall, this measure is calculated as follows:

$$F-score = \frac{2*precision*rapell}{precision+rapell} \quad (7)$$

4.3 Results and analysis

Figure 4 shows performance evaluation of our summarization system on a collection of 40 articles from EASC with a compression ratio CR=50%. Each generated summary was compared against five reference summaries in EASC corpus using ROUGE-1 and ROUGE-2 metrics.

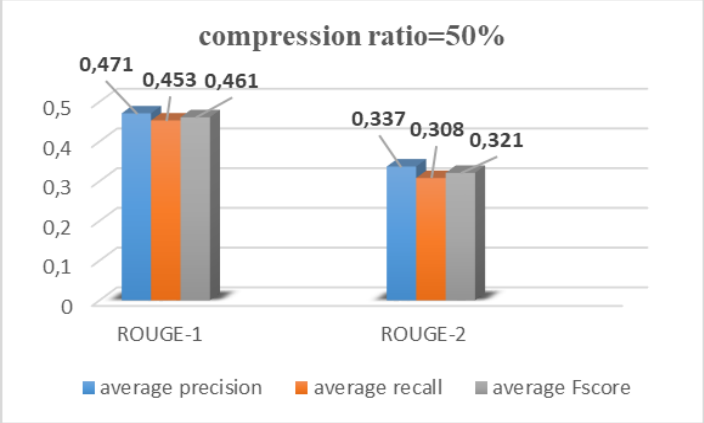

Figure 4: Performance of the proposed system with compression ratio=50%

Results analysis shows that our system achieves very good performance in term of precision, recall and F-score for both metrics ROUGE-1 and ROUGE-2. The average recall of our system attaint 0.453 using ROUGE-1 and 0.308 using ROUGE-2 indicating a high level of completeness and coverage. We can also note that the average precision of our system for both metrics is very good (average precision = 0.471 using ROUGE-1 and 0.337 using ROUGE-2) this means that the proposed system is quite preferment in excluding irrelevant sentences.

5. Conclusion and future works

To conclude, in this article, we have presented a new approach for automatic Arabic text summarization. The proposed approach combines a linguistic processing based on rhetorical analysis with a statistical processing. Rhetorical analysis is firstly applied to compress text sentences while keeping relevant segments. Sentence compression task is based on the exploitation of rhetorical relations defined within the rhetorical structure theory framework. Statistical processing is then used to reduce the number of compressed sentences based on three features: sentence position, similarity to the text title and sentence length. Results analysis on a text collection from EASC Data set proved clearly the efficacy of the proposed approach in terms of ROUGE-1 and ROUGE-2 measures.

As a future work, we will investigate the use of more linguistic features in statistical processing phase as well as exploiting both Intra-sentence and inter-sentence rhetorical relations to produce Arabic extracts.

References

Luhn, H. P. 1958. The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2): 159-165.

Baxendale, P. B. 1958. Machine-made index for technical literature—an experiment. *IBM Journal of Research and Development*, *2*(4): 354-361.

Edmundson, H. P. 1969. New methods in automatic extracting. *Journal of the ACM (JACM)*, *16*(2): 264-285

Fouad Soufiane, Douzidia, and Guy Lapalme. (2004). Lakhas, an Arabic summarization system. In *Proceedings of DUC* (Vol. 4), pages 128-135.

Alotaiby, F., Foda, S., Alkharashi, I. 2012. New approaches to automatic headline generation for Arabic documents. *Journal of Engineering and Computer Innovations*, *3*(1):11-25.

Ibrahim, Sobh, Nevin Darwish, and Magda Fayek.. 2006. An optimized dual classification system for Arabic extractive generic text summarization. *In Proceedings of the 7th Conference on Language Engineering, pages 149–154.*

Fattah, M. A., Ren, F. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text summarization. *Computer Speech & Language*, *23*(1):126-144.

Mohamed Mahdi, Boudabous, Mohamed Hédi Maaloul, and Lamia Hadrich Belguith. 2010. Digital learning for summarizing Arabic documents. *In proceeding of International Conference on Natural Language Processing*, pages 79-84. Springer, Berlin, Heidelberg.

Riadh Belkebir, Ahmed Guessoum. 2015. A supervised approach to Arabic text summarization using adaboost. In *New contributions in information systems and technologies*, pages 227-236. Springer International Publishing

Al-Radaideh, Q. A., Bataineh, D. Q. 2018. A Hybrid Approach for Arabic Text Summarization Using Domain Knowledge and Genetic Algorithms. *Cognitive Computation*, pages 1-19.

Qaroush, A., Farha, I. A., Ghanem, W., Washaha, M., & Maali, E. 2019. An efficient single document Arabic text summarization using a combination of statistical and semantic features. *Journal of King Saud University- Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2019.03.010.*

Mahmoud El-Haj, Udo Kruschwitz, and Chris Fox. 2011. Exploring clustering for multi-document Arabic summarisation. In *Asia Information Retrieval Symposium*, pages 550-561. Springer, Berlin, Heidelberg.

Oufaida, H., Nouali, O., Blache, P. 2014. Minimum redundancy and maximum relevance for single and multi-document Arabic text summarization. *Journal of King Saud University-Computer and Information Sciences*, *26*(4): 450-461.

Hamzah Noori, Fejer, Nazlia, Omar. 2014. Automatic Arabic text summarization using clustering and keyphrase extraction. *In proceeding of International Conference on Information Technology and Multimedia (ICIMU)*, IEEE. *Putrajaya, Malaysia*, pages 293-298.

Waheeb, S. A., Khan, N. A., Chen, B., and Shang, X. 2020. Multi-document Arabic text summarization based on clustering and Word2Vec to reduce redundancy. *Information*, *11*(2), 59.

Alqaisi, R., Ghanem, W., and Qaroush, A. 2020. Extractive Multi-Document Arabic Text Summarization Using Evolutionary Multi-Objective Optimization With K-Medoid Clustering. *IEEE Access*, *8* : 228206-228224.

Alami, N., Meknassi, M., En-nahnahi, N., El Adlouni, Y., & Ammor, O. 2021. Unsupervised neural networks for automatic Arabic text summarization using document clustering and topic modeling. *Expert Systems with Applications*, *172*: 114652.

Nabil, ALAMI, Yassine El Adlouni, Noureddine En-nahnahi, Mohammed Meknassi. 2017. Using Statistical and Semantic Analysis for Arabic Text Summarization. In proceeding of *International Conference on Information Technology and Communication Systems*, Springer, Cham, pages 35-50.

Kumar, Y. J., Goh, O. S., Halizah, B., Ngo, H. C., Puspalata, C. 2016. A review on automatic text summarization approaches. Journal of Computer Science, *12*(4):178-190.

Azmi, A. M., Al-Thanyyan, S. 2012. A text summarizer for Arabic. *Computer Speech & Language*, *26*(4), pp. 260-273.

Al Khawaldeh, F., Samawi, V. 2015. Lexical cohesion and entailment based segmentation for Arabic text summarization (LCEAS). *The World of Computer*

Science and Information Technology Journal (WSCIT), 5(3):51-60.

Azmi, A. M., and Altmami, N. I. 2018. An abstractive Arabic text summarizer with user controlled granularity. *Information Processing & Management, 54*(6): 903-921.

Mann, W. C., Thompson, S. A. 1988. Rhetorical structure theory: Toward a functional theory of text organization. *Text-Interdisciplinary Journal for the Study of Discourse, 8*(3):243-281.

Taboada, M., and Mann, W. C. 2006. Rhetorical structure theory: Looking back and moving ahead. *Discourse studies, 8*(3): 423-459.

Lynn Carlson, Daniel Marcu, Mary Ellen Okurowski. 2003. Building a discourse-tagged corpus in the framework of rhetorical structure theory. In: van Kuppevelt J., Smith R.W. (eds) Current and New Directions in Discourse and Dialogue. Text, Speech and Language Technology, volume 22, page 85-112, Springer, Dordrecht. https://doi.org/10.1007/978-94-010-0019-2_5

Lagrini, S., Azizi, N., Redjimi, M., and Dwairi, M. A. 2019. Automatic identification of rhetorical relations among intra-sentence discourse segments in Arabic. *International Journal of Intelligent Systems Technologies and Applications, 18*(3): 281-302.

Lagrini, S., Azizi, N., Redjimi, M., and Dwairi, M. A. 2019a. Toward an automatic summarisation of Arabic text depending on rhetorical relations. *International Journal of Reasoning-based Intelligent Systems, 11*(3): 203-214.

Attia, Mohammed. A. 2007. Arabic tokenization system. In *Proceedings of the 2007 workshop on computational approaches to semitic languages: Common issues and resources*, pages 65-72.

Maha Althobaiti, Udo Kruschwitz, Massimo Poesio. 2014. AraNLP: A Java-based library for the processing of Arabic text. *In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC), pages 4134-4138.*

El-Khair, I. A. (2006). Effects of stop words elimination for Arabic information retrieval: a comparative study. *International Journal of Computing & Information Sciences, 4*(3): 119-133.

Nabil Alami; Mohammed Meknassi; Said Alaoui Ouatik; NourEddine Ennahnahi. 2016. Impact of stemming on Arabic text summarization. In *2016 4th IEEE International Colloquium on Information Science and Technology (CiSt), IEEE, pages 338-343.*

Shereen Khoja and Roger Garside. (1999). Stemming Arabic text. Lancaster, UK, Computing Department, Lancaster University.

Al-Abdallah, R. Z., Al-Taani, A. T. 2017. Arabic Single-Document Text Summarization Using Particle Swarm Optimization Algorithm. *Procedia Computer Science, 117*: 30-37.

El Haj, M., Kruschwitz, U., Fox, C. (2010).Using Mechanical Turk to Create a Corpus of Arabic Summaries. *Editors & Workshop Chairs*, pages 36-39.

Chin-Yew, Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. *Proceedings of Workshop on Text Summarization Branches Out,* Barcelona, Spain, pages: 74-81.

Appendix. Sample Arabic text

والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد ، اذ وقبل اندلاع الأعمال الفدائية بزمن، كان هناك مناضلون فلسطينيون ينطلقون من ذلك القطاع وعبره للقيام بعمليات قاسية ضد قوات الاحتلال الاسرائيلية.وكان الوضع يصل الى لحظات توتر قصوى، عند بدايات 1955، حيث اندلعت أعمال عنف ضد القوات الاسرائيلية وكذلك ضد المنشآت التابعة للامم المتحدة. ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبدالناصر الذي، كان يحاول ان يهدئ الأوضاع، مراعاة لخاطر الاميركيين من جهة، وتأجيلاً للانفجار المحتمل بين مصر واسرائيل من جهة ثانية . ومن هنا حين احتلت القوات الاسرائيلية غزة في العام 1956 خيل للكثيرين انها لن تنسحب منها بعد ذلك، على رغم الضغوط الدولية . ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه، ان تطرح الثقة في الحكومة، لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون التي لم تبد أي اهتمام حتى بالتظاهرات التي نظمتها المعارضة في الشارع داعية الى الابقاء على احتلال قطاع غزة.

In fact, Gaza Strip has been a long time ago disturbing the Israeli authorities, as long before the outbreak of guerrilla actions, there were Palestinian militants who launched from and through that strip to carry out harsh operations against the Israeli occupation forces. The situation reached moments of extreme tension, at the beginning of 1955, Where violence erupted against the Israeli forces as well as against United Nations Establishments. This tension prompted Israel to launch deadly raids on the Gaza Strip, which generated a harsh reaction from the Egyptian authorities led by Gamal Abdel Nasser, who was trying to calm the situation, taking into account the dangers of the Americans on the one hand, and postponing the possible explosion between Egypt and Israel on the other. Hence, when the Israeli forces occupied Gaza in 1956, many imagined that they would not withdraw from it after that, despite international pressures. Many parliamentary groups, including groups from the ruling Labor Party, have tried to raise confidence in the government, but all this did not weaken the resolve of the Ben-Gurion government, which did not show any interest even in the demonstrations organized by the opposition in the street calling for maintaining the occupation of the Gaza Strip.

The final summary followed by its translation is:

والحقيقة ان قطاع غزة كان يقض مضاجع السلطات الاسرائيلية منذ زمن بعيد. ولقد دفع ذلك التوتر اسرائيل الى شن غارات قاتلة على القطاع، مما ولد رد فعل قاس لدى السلطات المصرية بزعامة جمال عبد الناصر الذي كان يحاول ان يهدئ الأوضاع . ولقد حاولت فئات نيابية كثيرة، ومنها مجموعات من حزب العمل الحاكم، نفسه، ان تطرح الثقة في الحكومة، لكن هذا كله لم يوهن من عزيمة حكومة بن غوريون.

In fact, the Gaza Strip has long been a sleeper of the Israeli authorities. This tension prompted Israel to launch deadly raids on the Gaza Strip, which generated a harsh reaction from the Egyptian authorities led by Gamal Abdel Nasser, who was trying to calm the situation. Many parliamentary groups, including groups from the ruling Labor Party itself, tried to put confidence in the government, but all this did not weaken the resolve of the Ben-Gurion government.