

Automatic Arabic Text Summarization Approaches

Samira Lagrini

Labged Laboratory, Computer
Science Department,
University Badji Mokhtar, PO
BOX 12, Annaba
Annaba, 23000, Algeria

Mohammed Redjimi

Computer science Department
University 20 Aout 1955
Skikda, 21000, Algeria

Nabiha Azizi

Labged Laboratory, Computer
Science Department,
University Badji Mokhtar, PO
BOX 12, Annaba
Annaba, 23000, Algeria

ABSTRACT

In recent years, automatic text summarization has seen renewed interest, and has been experiencing an increasing number of researches and products especially in English language. However, in Arabic language, little works and limited researches have been done in this field. This paper exposes a literature review of recent research works on Arabic text summarization. Current approaches used in this field are presented followed by a discussion about their limitations and the main challenges faced when dealing with such application. As a final point, a proposed approach to improve the quality of Arabic text summarization system is presented.

General Terms

Natural Language Processing (NLP); Automatic summarization approaches.

Keywords

Arabic text summarization, clustering, RST, machine learning, graph theory, text entailment, extractive approaches.

1. INTRODUCTION

Text summarization is one of the most important applications of Natural Language Processing (NLP). It is an essential tool for assisting and interpreting text information. The goal of automatic text summarization is to abbreviate one or more texts into shorter version conserving their information contents and overall meanings [1]. This will help the reader to decide if a document covers desired information with minimum effort and time loss.

There is no doubt about the importance of such application. It could be used, for example, in search engine web pages as an informative tool for the user to find the pertinent and required information [2].

A summary can be defined as "a text which is produced from one or more texts, that conveys important information in the original texts and that is no longer than half of the original text(s) and usually significantly less than that" [3].

Summarization systems can be classified according to many features: input, output, generality, language, method (see fig.1). This allows summaries to be characterized by a broad range of properties [4]. For instance, a summary can be generated for a single document or a group of documents. The first is called single document summarization while the second deals with multi-document summarization. According to its generality, a summary can be classified into generic summary and query driven summary. Generic summary attempts to represent all pertinent topics of the input document. While, Query driven depends on a specification of a user information needs. Concerning the style of the output, a large distinction is made between indicative and informative

summary. Indicative summary is used to indicate what topics are addressed in the source text. This kind of summary helps reader getting a general idea of a text without taking into account further details. The informative summary is intended to cover all topics in the source text.

Furthermore, one can talk about monolingual and multilingual summarizer. Monolingual summarization systems are designed to work with only one language and have the input document and the output summary in the same language, unlike multilingual summarizers, which cover more than one language. Finally, text summarization methods can be classified into extractive and abstractive summarization. The extractive summarization method consists in selecting important sentences from the original document based on statistical and linguistic features, and concatenating them into shorter form [1]. While the abstractive summarization method differs mainly from the extractive ones by providing summaries having some degree of inference about background knowledges not necessary presented in the original document [5]. It uses linguistic methods to examine the text, and then find new concepts and expressions to best describing it.

The goal of this research is to present an investigation on extractive Arabic text summarization approaches. This research work has three main contributions: (i) to investigate state-of-the-art of text summarization approaches proposed in Arabic language. (ii) to analyze the limitation of current approaches and the peculiarity of Arabic language, which have posed challenge to the task of Arabic text summarization. Therefore, (iii) to propose the main lines of a new approach, which combines semantic information extracted from Arabic WordNet and rhetorical structure theory (RST), one of the most widely, used discourse theories in natural language processing.

The remainder of this article is organized as follows. Section 2 investigates the main approaches proposed for extractive Arabic text summarization. Then, section 3 explains the limitation of these approaches. Section 4, and section 5 describe the proposed approach and its main steps, finally, Section 6 concludes this article.

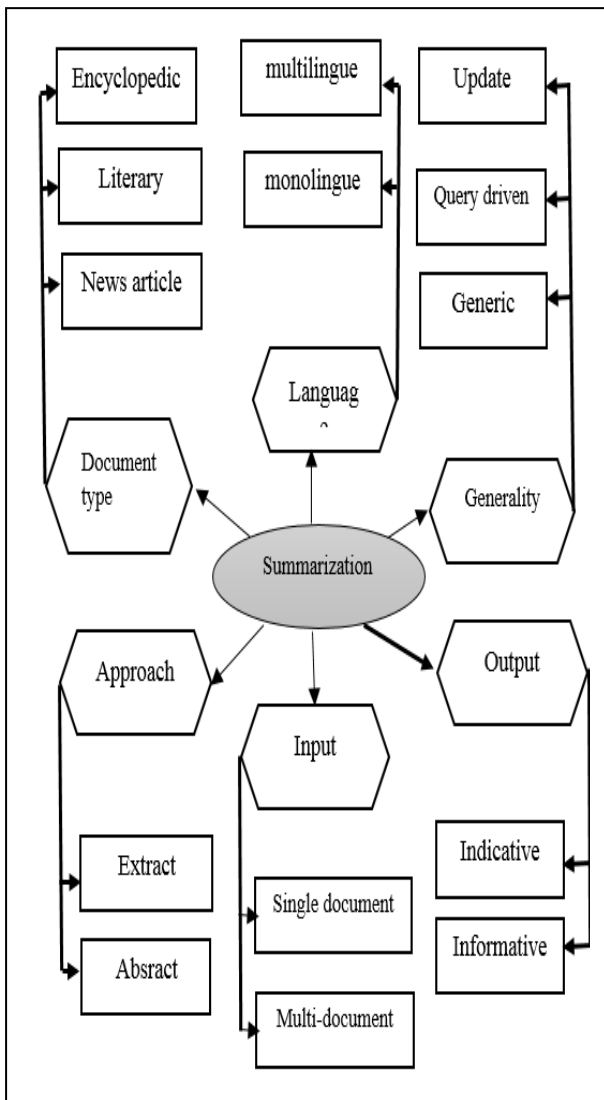


Fig 1: Summary taxonomy

2. ARABIC TEXT SUMMARIZATION APPROACHES

2.1 Discourse theory

1) Rhetorical structure theory (RST)

The Rhetorical Structure Theory (RST) [6] is a descriptive theory of discourse organization using the discourse relations that hold between text segments. According to RST, a coherent text can be represented as a discourse tree where the leaf nodes are non-overlapping text segments called elementary discourse units (EDUs), and internal nodes are the concatenation of continuous EDUs. Adjacent nodes are linked through particular discourse relations to form a discourse sub tree, which can then be related to other adjacent nodes in the tree structure.

In [7], author has addressed the first employment of this theory on Arabic text summarization. The authors attempted to develop an infrastructure for Arabic text summarization based on the RST. Therefore, he suggested different techniques, algorithms and design patterns to be considered when dealing with such application.

Recently, the rhetorical structure theory is also used for classifying Arabic security documents [8]. The proposed technique parses each paragraph in the document and builds

the rhetorical tree that represents its structure. Then, it determines the importance of each paragraph by examining the promotion of the tree root. If the importance of the paragraph conforms to the user instruction, the classifier labels it with the required classification.

In [9], the authors propose a two-pass algorithm, where a primary summary is generated using rhetorical structure theory in the first pass. In the second pass, to each sentence within the primary summary is awarded a score based on the frequency of the word and the title keyword in the sentence. For the final summary, sentences are selected with an objective of maximizing the overall score of the summary whose size should not exceed the user selected limit.

Other approaches provide a hybrid model like in [10]. The proposed model combines RST and vector space model (VSM). RST is used to discover the most important paragraphs based on semantic criteria, and VSM is used for ranking these paragraphs based on the cosine similarity features. The author shows that the hybrid model combining RST and VSM is able to take the advantages of both, as it improve the average precision of the summary over using RST only.

Similarly, in [11] author proposes a hybrid approach that combines the RST and support vector machine (SVM) algorithm. In their system, the authors use rhetorical structure analysis to identify the minimal units of the text and determine the rhetorical relations between them. SVM algorithm is used to decide whether a sentence having a rhetorical relation "Others" is relevant or not for the final extract. This relation is assigned when no other rhetorical relation is determined. Evaluation results show that the system achieves good results, which confirms that applying a hybrid approach could enhance the summarization system performance.

2) Segmented Discourse Representation Theory (SDRT)

The Segmented Discourse Representation Theory (SDRT)[12], is a theory of discourse interpretation that represent the discourse relations holding between elementary discourse units (EDUs), which are clauses or complex discourse units that are built up from linked EDUs in recursive way. Unlike RST, SDRT allows attachment between non-adjacent discourse units and for multiple attachments to a given discourse unit, which means that the discourse structures created are not always trees but rather directed acyclic graphs.

For the best of our knowledge, [13] addressed the first employment of this theory on Arabic text summarization. The authors proposed a semantically driven approach to analyze Arabic discourse following the SDRT framework. They, also show how the discourse analysis can be used to produce indicative summaries of Arabic documents. For this purpose, they design several content selection algorithms that take as input the document discourse structure and produce as output a subset of EDUs, which better summarizes the original document. The selection process is guided by three discursive criteria: the semantics of discourse relations, their nature (coordinating vs. subordinating) and the document discourse structure (tree vs. graph). To measure the impact of discourse structure on producing indicative summaries, the authors evaluate their algorithms by comparing their performances against the gold standard summaries manually generated from two different corpora, the first one annotated according to the SDRT framework and the second annotated according to the RST. Results show that all discourse information are useful

for the content selection task and in turn improve the quality of Arabic text summarization systems.

2.3 Cluster based approach

Many Arabic text summarization systems use clustering to generate a summary. For instance, [14] proposed a new Arabic single and multi-document summarization method based on automatic sentence clustering and an adapted discriminate analysis method. Their system used a clustering algorithm to group similar sentences into cluster. The proposed method takes advantage of term's discriminate power to score sentences.

Similarly, [15] proposed a new model for single and multi-document summarization based on document clustering and key phrase extraction. The model use a hybrid clustering (partitioning and k-means) to group Arabic documents into several clusters, then it extract important key-phrases from each cluster; second it find similar sentence based on cosine similarity algorithms. Only sentences that have a greater similarity than the predefined threshold are selected to represent the summary. The model reached good results for single and multi-document summarization but no comparison with other systems is achieved.

Unlike the previous presented systems, [16] use clustering to group words with the same root in the same cluster. The number of words in that cluster determines the weight of each word in the cluster. Then the model calculates the score of each sentence based on several features. Sentences with the highest score are selected to represent the final summary. Finally, in [17].The authors explore how clustering can be applied to Arabic multi-document summarization as well as for redundancy elimination. Two experiments are conducted. In the first one, K-means algorithm is used to cluster sentences. More precisely, a number of sentences are selected randomly as the initial centroids, and then all sentences are iteratively assigned to the closest cluster based on their cosine similarity measure. To produce the summary, two selection methods are used: In the first methods, the first sentence of each cluster is selected, while in the second one, all sentences in the biggest cluster are selected and ranked according to their similarity to the centroid. For the second experiment, sentences are selected before applying the clustering. From each document, only the first sentence and the most similar sentence to the first one are selected. Then, all the subsequent steps are similar to the first experiment. For evaluation, DUC-2002 dataset and an Arabic parallel translation version are used. The generated summaries are evaluated using ROUGE, precision and recall metrics. Results are compared with the top five systems in the DUC 2002 competition. Comparing the ROUGE-1 results. The proposed summarizers achieve higher scores than the top systems reported at DUC-2002.

2.4 Machine Learning based Approach

In machine learning based approach, the summarization process is modeled as a classification problem: given a set of training document and their extractive summaries, each sentence is classified as a summary sentence or non-summary sentence based on statistical features.

Some Arabic summarization systems have been adopting machine learning and statistical techniques. For instance, in [18] the authors integrate Bayesian and genetic programming classification methods in an optimized way to extract the summary sentence using reduced feature set. The system requires training and uses manually labeled corpora. Experiments show that, Bayesian classifier tends to have large

recall unlike GP classifier, which tends to have large precision. By integrating both classifiers, the author found that using the union for integration increased the recall and the summary size, while using the intersection for integration increased the precision and decreased the size of the summary.

Later, in [19] the authors investigate the use of genetic algorithm (GA), mathematical regression (MR), feed forward neural network (FFNN), probabilistic neural network (PNN) and Gaussian mixture model (GMM) for automatic text summarization task. Ten features are used in combination to train the over mentioned methods on a manually created corpus. The system was also tested using DUC 2001 corpus. Numerous experiments were performed. In addition, results indicated that GMM model is the best.

Recently, [20] proposed a supervised approach using Adaboost to produce Arabic extracts. The authors use a set of statistic features such as overlap with word title, sentence position, the number of keywords in the sentence, and sentence length.

The authors use a manually created corpus composed of 20 Arabic news articles with their summaries. Performance evaluation in term of F-measure are compared to those obtained using multilayer perceptron (MLP) and j48 decision trees. Results indicate that the proposed model outperform multilayer perceptron and j48 decision trees.

2.5 Graph based Approach

With graph-based approach, the document is represented in the form of undirected graph. There is a node for every sentence. An edge between two nodes is drawn if there is a relation between these two nodes. A relation can be a cosine similarity above a threshold, or any other type of relationships.

After drawing a graph, it is possible to view the sub-graphs of connected nodes as a cluster of distinct topics covered in the document. This yields two results: For query-specific summaries, sentences may be selected only from the pertinent sub-graph, while for generic summaries sentences would be selected from each sub graph for best coverage.

Recently, in [21] the authors represent each document by weighted directed graph whose nodes represent sentences and edges weights represent similarity between sentences. This similarity is determined by ranking the sentences according to some statistical features. The cosine similarity measure is chosen based on term weighting scheme, which is the TF-IDF (Term Frequency-Inverse Document Frequency). The summary is extracted by finding the shortest path between the first and the last nodes in the graph considering the user compression ratio. The proposed approach is evaluated using EASC corpus, and intrinsic evaluation method.

2.6 Ontology based Approach

Arabic Word Net (AWN) is a lexical database for Arabic language. It groups Arabic words into sets of synonyms called *synsets*, and recording the various semantic relations between these synonym sets. The goal is to produce a combination of dictionary that support automatic text analysis applications.

Recently, some Arabic researchers tend to use this lexical database in their systems. For instance [22] presented a new query based Arabic text summarization system (OSSAD) using Arabic word Net and an extracted knowledge base. The user's query is expanded once by using the Arabic Word Net thesaurus, and then by adding the domain specific knowledge

base to the expansion. For summarization, decision tree algorithm is used, which was trained by a set of features extracted from the original documents. The OSSAD summary was evaluated against other Arabic summarization systems working on the same data. The results show that OSSAD performance overcomes other systems and reach a good level of performance.

2.7 Textual entailment based approach

Textual entailment has been proposed recently as a generic framework for modeling semantic variability in many natural language processing applications. An entailment relation consists in determining whether another one can infer the meaning of one sentence [4]. The summary obtained by using entailment inferences only include sentences that are not entailed by any of the sentences in the previously accumulated summary.

Very little research has been done to combine Arabic text summarization and text entailment to produce extracts. Only, in [23] where the authors tackle the problem of developing Arabic text summarization system (LCEAS), that produces extract without redundancy. Lexical cohesion and entailment-based segmentation were utilized. Lexical cohesion is applied to distinguish the important sentences from the unimportant sentences in the text. As a result, poor information is removed from the text before applying text entailment algorithm. In the next stage, cosine directional similarity method is applied to decide which sentences are not redundant. The text entailment algorithm suggested in [24] is enhanced to make it suitable for Arabic language. The enhancement include using root and semantic relation between the senses of the word to extract the common words and specifying new threshold value to suit Arabic language. Performances evaluation of LCEAS are compared with previous Arabic text summarization systems. Results indicate that LCEAS is the best.

3. LIMITATION OF CURRENT APPROACHES AND MAIN CHALLENGES

As we can see, all the summarization approaches described in the previous section are extractive. This means that sentences are selected from the input document to produce a summary. Unless a background repository is being used, the system is limited only to the words mentioned in the input document [5]. In machine learning based approaches [18][19][20], other limitations appear: one limitation is ignoring pertinent words that appear in abundance in the test set but not in the training set, thus the system lack the ability to analyze such words and it will treat them as irrelevant words. Another limitation is the lack of detection for the implicit relationships between words in the input document. The ability to detect such relationships requires an external knowledge base. Most of the Arabic text summarization approaches are affected by a similar limitation in the detection of concepts and the relatedness between them. Probably, this is due to the lack of linguistic resources for Arabic language.

For discourse-based approach, other challenges appear, for instance, identifying elementary discourse units' boundaries for Arabic texts is a difficult task. Punctuation marks are not widely used in Arabic texts. We can find an entire paragraph without any punctuation. Hence, relying on punctuation mark alone is not sufficient for Arabic texts segmentation. In addition, Arabic does not have capital letters, which affect the recognition of named entities in the process of annotation.

Furthermore, Arabic discourse connectives are highly ambiguous. In the Arabic rhetoric system, the meaning of the discourse connective plays a great role of understanding consecutive sentences and in turn determines the sentence boundaries. It is realized that the connector "و" is the most ambiguous connector due to its mostly rhetorical use. In [25] the author classified the six meaning of connective "و" into two classes: the first class called "Fasl" which means that the connector is a good indicator to begin a segment. This class contains: (1) "واو القسم" that means testimony, (2) "رب" that means few or someone and (3) "واو الاستئناف" that simply joins two unrelated sentences. The second class called "Wasl" include the different states where the connector has no effect on the segmentation. This class contains: (1) "واو الحال" that introduces a state, (2) "واو المعية", which means the accompaniment and (3) "واو العطف" that means the conjunction of related words or sentences.

Finally, the major challenge in summarization task lies in distinguishing the most informative parts of a document from the less informative ones. According to [26], a good summarization technology aims to combine the central topics with completeness, readability and conciseness. Determining the effective features that extract the main ideas from the input document and that cover all important themes is a greater challenge in extractive text summarization especially for Arabic language which have a complex nature and rich semantic.

4. PROPOSED APPROACH

In our research work, we propose to combine both linguistic and semantic feature selection methods in order to improve a quality of Arabic text summarization. The proposed approach uses the RST in the first phase to generate a primary summary, this is followed by the second phase where we assign a score to each sentences in the primary summary based on its semantic similarity with the original text title and sub titles. The main steps of the proposed approach are presented in the next section.

5. PROPOSED APPROACH MAIN STEPS

There are two main phases in our approach. Each phase can be divided into sequence steps. Fig.2 represents the suggested Text summarization steps.

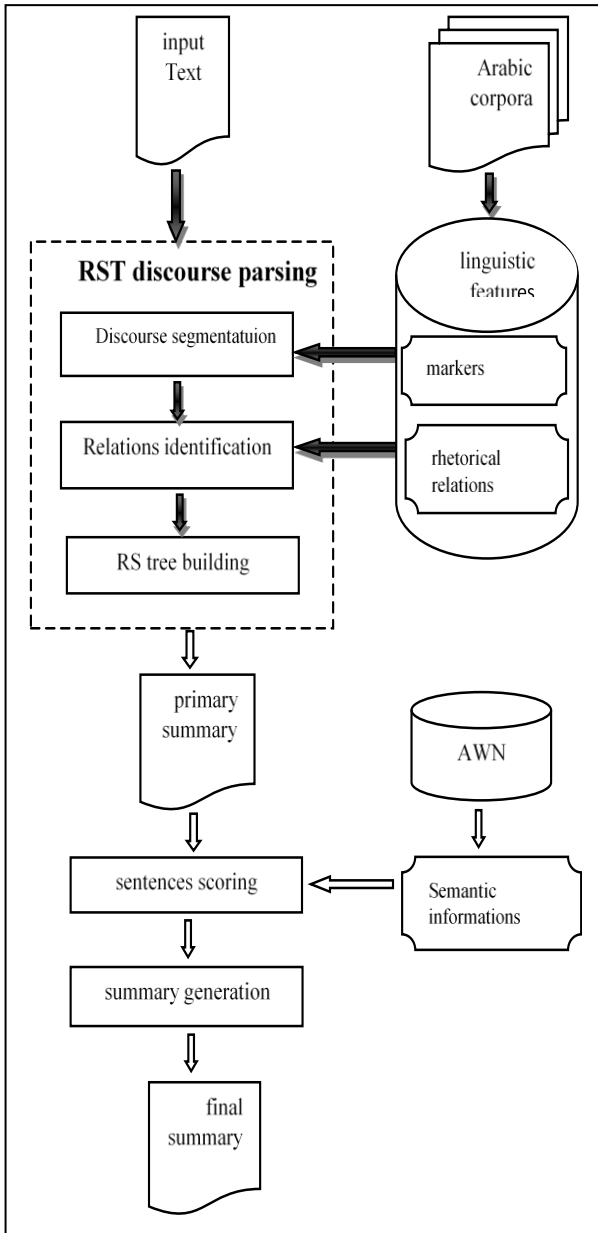


Fig 2: Proposed approach main steps

1) Discourse segmentation: This step consists of splitting an input text into non-overlapping elementary discourse units EDUs. Conventionally, the task of automatic discourse segmentation is formulated as follow: given a sentence, the segmentation model identifies the boundaries of the composite EDUs by predicting whether a boundary should be inserted before each particular token in the sentence [27]. To achieve this task we can use either a machine learning approach or a rule-based approach. In machine learning approach, each word in the sentence is considered independently. The segmentation model scans the sentence word by word, and a binary classifier is used to predict whether it is suitable to insert a boundary before the word being examined.

Fig.3 shows an example of sentence segmentation into three elementary discourse unit.

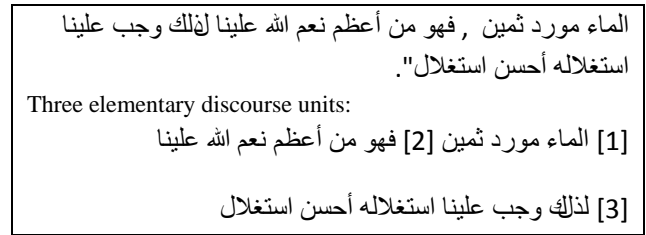


Fig 3: An example of sentence segmentation

Identifying elementary discourse unite boundaries is an important step in RST discourse parsing, therefore, its performance is crucial to the overall accuracy.

2) Rhetorical Relations identification: In order to identify the discourse relations that hold between text spans we can use a multi class classifier that takes as input two text spans (feature vectors) and generate as output the relation label that can link this segments. TABLE 1. Present Some Arabic discourse relations and discourse connectors that can signal them.

TABLE 1. Some Arabic Rhetorical Relations

Rhetorical relations		Discourse connectors
Condition	شرط	لو - إذا
Explanation	التفسير	أي -بمعنى أن -
Concession	الاستدراك	لكن -غير أن -رغم
Reason	التعليل	لأن
Rectification	اضراب	بل
Result	نتيجة	لذا --نتيجة لهذا --لذلك
Exemplification	تمثيل	مثلا -على سبيل المثال
Attribution	استدلال	أوضح- أفاد - أعلن - قال
Exception	استثناء	عدا -خلا

3) RST- tree building: Two kind of RS-tree-building strategies can be adopted: greedy strategies, which consider only one possible solution at each parsing step, and non-greedy strategies, which consider multiple or even all possible solutions at each parsing step, and choose the best one among them. For our approach, we will adopt the greedy strategy since it is straightforward and quite efficient in practice.

As an example, if we want to build the RS-tree of the previous text in (Fig.3) we will end up with the diagram shown in (Fig.4).

As we can see, each node has associated a type T that represent the rhetorical relation that hold between the two linked EDUs a status is either nucleus represented by solid circle or satellite represented by hollow circle. The values within circles are the elementary discourse units' numbers, which denote the most important units of the textual span that is dominated by that node.

Once the RS-tree that represent the input text is built, the primary summary will be generated based on several features.

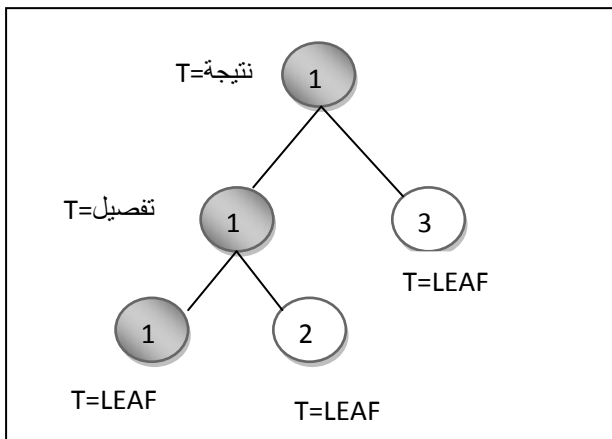


Fig 4: RS-tree diagram

4) Sentence scoring and summary generation: After generating the primary summary using RST, each sentence in this summary will be awarded a score based on its semantic similarity with original text title and sub titles. Since text, title and sub titles are the most important parts in the text, which can be viewed as a very short summary of the original text. To compute this similarity, we rely on semantic information extracted from Arabic Word Net, for the final summary, sentence with highest score will be selected taking into account the user's compression ratio.

6. CONCLUSION

This survey paper focuses on extractive Arabic text summarization approaches. The most recent progresses and researches raised in this field through the last years are presented here

At first, some basic notions related to automatic text summarization are described, and then the main approaches proposed in this field are presented. After that, the limitations of these approaches and the major challenges are discussed. Finally, an approach aiming to enhance the quality of automatic Arabic text summarization system is presented.

As a conclusion, we can say that Arabic text summarization is still in its initial stage compared to the work done in English and other languages; this is partially due to the lack of resources (NLP tools) and the rich semantic and complex morphology of Arabic language.

Further works concern the implementation of the proposed approach. Thanks to the progress of software engineering, there is now a wide range of tools (design methods, development platforms, programming environments, etc.) that enable modeling, simulation, development and implementation of complex applications.

7. REFERENCES

[1] Gupta, V. 2010. A survey of text summarization extractive techniques. *J. Emerging Technologies in Web Intelligence*, 2(3), 258-268.

[2] El-Shishtawy, T., El-Ghannam, F. 2012. Keyphrase based Arabic summarizer (KPAS). In *informatics and systems (INFOS)*, 8th International Conference, IEEE.

[3] Radev, D.R. , Hovy, E., McKeown, K. 2002. Introduction to the special issue on Summarization. *Computational linguistics*, 28(4), 399-408.

[4] Lloret, E., Ferrandez, O., Munos, R. 2008. A Text summarization approach under the influence of textual entailment. In *NLPCS*, 22-31.

[5] Bawakid, A. 2011. Automatic documents summarization using ontology based methodologies. Doctoral Thesis. University of Birmingham.

[6] Mann, W.C. , Thompson, S. 1988. Rhetorical structure theory: toward a functional theory of text organization, *Text-Interdisciplinary J. for the Study of Discourse*, 8(3), 243-281.

[7] AlSanie, W. 2005. Towards an infrastructure for Arabic text summarization using rhetorical structure theory. Master's Thesis. King Saud University, Riyadh, Saudi Arabia.

[8] Mathkour, H. 2009. A Novel rhetorical structure approach for classifying Arabic security documents. *J. Computer Theory and engineering*, 1(3), 195.

[9] Azmi, A., Al-Thanyyan, S. 2012. A Text summarizer for Arabic. *J. Computer Speech and Language*, 26(4), 260-273.

[10] Ibrahim, A., Elghazaly, T., Gheith, M. 2013. A Novel Arabic text summarization model based on rhetorical structure theory and vector space model. *J. Computational Linguistique and Natural Language Processing*, 2(8) , 480-485.

[11] Maaloul, M.H. 2012. Approche Hybride pour le Résumé Automatique de Textes Application à la langue arabe. Doctoral Thesis. Université de Provence-Aix-Marseille I, France.

[12] Asher, N., Lascarides, A. 2003. *Logics of conversation*, Cambridge University Press .

[13] Keskes, I. 2015. Discourse analysis of arabic documents and application to automatic summarization. Doctoral Thesis. University of Toulouse III-Paul Sabatier.

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

[15] Fejer, H.N., Omar, N. 2014. Automatic Arabic text summarization using clustering and keyphrase extraction. In *Information Technology and Multimedia (ICIMU)*, 293-298.

[16] Haboush, A., Al-zoubi, M. 2012. Arabic text summarization model using clustering techniques. *J. Word of Computer Science and Information Technology (WCISIT)*, 2221-0741.

[17] El-Haj, M., Kruschwitz, U., Fox, C. 2011. Exploring clustering for multi-document arabic summarisation. In *Asia Information Retrieval Symposium*, Springer Berlin Heidelberg, 550-561.

[18] Sobh, I.M.A.H. 2009. An optimized dual classification system for Arabic extractive generic text summarization, Doctoral Thesis. Faculty of engineering, Cairo university, Giza, Egypt.

[19] Abdel Fattah, M., Ren, F. 2009. GA, MR, FFNN, PNN and GMM based models for automatic text

- summarization. *Computer Speech & Language* , 23(1), 126-144.
- [20] Belkebir, R., Guessoum, A. 2015. A supervised approach to Arabic text summarization using adaboost. *New contributions in information systems and technologies*, Springer International Publishing, 227–236.
- [21] Al-Taani, A.T. , Maha, M. 2014 .An extractive graph-based Arabic text summarization approach.
- [22] Imam, I., Nounou, N., Hamouda, A., Abdul Khalek, H. 2013. An ontology-based summarization system for arabic documents (ossad). *J. computer application(IJCA)*, 74(17).
- [23] Al-Khawaldeh, F., Samawi, V.W. 2015. Lexical cohesion and entailment based segmentation for Arabic text summarization(LCEAS) .*J. the Word of Computer Science and Information Technology*, 5(3), 51-60.
- [24] Tatar, D., Mihis, A., Lupsa, D. 2009. Entailment based linear segmentation in summarization. *J. Software Engineering and Knowledge Engineering*, 19(08), 1023-1038.
- [25] Khalifa, I., Feki, Z. and Farawila, A. 2011. Arabic Discourse Segmentation Based on Rhetorical Methods. *Int. J. Electric Comput. Sci*, 11(1), 10-15.
- [26] Lloret, E., Palomar, M. 2012. Text summarisation in progress: a literature review. *Artificial Intelligence Review*, 37(1),1-41.
- [27] Feng, V.W. 2015 *RST-Style Discourse Parsing and Its Applications in Discourse Analysis*. Doctoral Thesis. University of Toronto, Canada.