OPEN ACCESS

International
Journal
Of Modern Engineering Research (IJMER)

Query Answering Approach Based on Document Summarization

Hesham Ahmed Hassan[1], Mohamed Yehia Dahab[2], Khaled Bahnassy[3], Amira M. Idrees[4], Fatma Gamal[5]

[1]*Faculty of Computer and Information Computer Science Department, Cairo University*
[2]*Faculty of Computer and Information, Computer Science Department, King Abdulaziz University*
[3]*Faculty of Computer and Information, Computer Science Department, Ain Shams University*
[4]*Faculty of Computer and Information, Information Systems Department, Fayoum University*
[5]*Faculty of Computer and Information, Computer Science Department, Cairo University*

Abstract: *The growing of online information obliged the availability of a thorough research in the domain of automatic text summarization within the Natural Language Processing (NLP) community.The aim of this paper is to propose a novel approach for a language independent automatic summarization approach that combines three main approaches. The Rhetorical Structure Theory (RST), the query processing approach, and the Network Representationapproach (NRA). RST, as a theory of major aspect for the structure of natural text, is used to extract the semantic relation behind the text.Query processing approachclassifies the question type and finds the answer in a way that suits the user's needs. The NRA is used to create a graph representing the extracted semantic relation. The output is an answer, which not only responses to the question, but also gives the user an opportunity to find additional information that is related to the question.We implemented the proposed approach. As a case study, the implemented* approach*is applied on Arabic text in the agriculture field. The implemented approach succeeded in summarizing extension documents according to user's query. The approach results have been evaluated using Recall, Precision and F-score measures.*

Keywords: *Information Extraction, Text Summarization, Natural Language Processing.*

I. Introduction

Summarization is "a brief restatement within the document (usually at the end) of its salient findings and conclusions, and is intended to complete the orientation of a reader who has studied the preceding text" while an abstract is, according to the same standard, a "Short representation of the content of a document without interpretation or criticism". [MARTIN, 2008]. According to Mikael, in [Mikael, 2014], Automatic text summarization approaches can be classified into vector based approach, Fuzzy based approach, Genetic algorithm based approach, and Neural Network based approach.

Semantic pattern can be defined, according to [Mohamed, 2008] as "a generic format for natural language expression, to declare a specific meaning". The distinguishing of these semantic patterns are not straightforward since natural languages may have different lexical items that can be used to make reference to the same situation as well as different syntactic realization of the same arguments.

The semantic patterns elements are:

- Abstract ontological class. These classes were imported from the Agrovoc thesaurus and the publications of CLEASE as we will discuss later.
- Verb group. These groups were extracted from different lexicons like the Wordnet.
- Text constant expression.

All these elements are non-terminal element except the third element, it is a terminal element.We refer to abstract ontological class as a word between "<>" signs.

In this proposed approachwe will be working with single document summarization as an experimental study and our aim will be to produce a short summary that best suits both, the user's criteria and the writer's point of view.We will introduce some common terms in the summarization dialect: extraction is the process of detecting important segments of content and generating a new verbatim of these segments; abstraction targets to construct significant information in a new, non-verbatim way; fusion merges extracted segments coherently; and compression objects to discard unimportant segments of text [Radev et al., 2004]. Initial studieson summarizing documents proposed models for extracting weighty sentences from text using features like word or phrase frequency [Luhn, 1958], position in the text[Baxendale, 1958] and key phrases[Edmundson, 1969]. Our summary will be extractive summarization that takes benefit of three approaches, Rhetorical Structure Theory, query processing approach, and the Network Representation approach.

II. Related Work

Several automatic text summarization techniques have been proposed. These summarization techniques are classified according to [Mikael, 2014] into four categories, they are Heuristic techniques, Semantics-based techniques, Query-oriented techniques, and Cluster-based techniques. Based on these different techniques, we'll review the work done on text summarization in the last few years.

Barzilay and Elhadad in 1997 present an algorithm that computes lexical chains in a text. Based on lexical chains, they identify the nominal groups of sentences and the algorithm for segmentation by using different sources such as the part of speech tagger. Other sources may be used such as the wordnet thesaurus [Barzilay, 1997]. According to [Mikael, 2014] this method shows improvement over commercially available summarizer systems but still has two limitations. First: Sentence granularity- there is a high probability of selecting long sentences to be included in the summary. Second: the sentences that are selected and included in the summary may contain anaphor links to other parts of the text which may not be included in the summary.

In 2006 Wang and Yang [Wang, 2006] suggested a "fractal summarization technique". It uses a fractal approach for controlling the information viewed [Dolores, 2008].The fractal theory converts the text document into a tree hierarchy [Mohsen, 2012]. Their technique proposed A fractal theory to produce a summary by determining the salient features for the text and its hierarchical structure. The proposed technique used a statistical approach; therefore it can be used for multilingual text documents with minor modification of the system due to the difference of each language's features [Mikael, 2014]. Wang and Yang technique enhances the convergence of information analysis of a summary as user can control the compression ratio, and the system produces a summary that expands the information coverage and reduces the dissimilarity from the source document. Fractal theory that considers both the abstraction level of document and statistical property of the text their result shows the superior result compared to flat methods and the other structured summarization method in literature [Mohsen, 2012].

In 2007, Steinberger et al. proposed a new method for using anaphoric information in Latent Semantic Analysis (LSA) and consider its product to develop an LSA-based summarizer [Steinberger, 2013]. This method was able to attain improved performance more than the methods that don't use anaphoric information, and it also had an improvement performance by the rouge measures than all except the ones of the single-document summarizers participating in DUC-2002. The LSA has some limitation; first of which is that the word order has no effect on the syntactic relations or logic, or of morphology. Strangely, despite of this limitation the system succeeds to extract accurate reflections of segment and word implication quite well; nonetheless there would be few errors on some occasions [phiên, 2008]. Another limitation is in the resulting dimensions. The resulting dimension is not easy to interpret. This leads to results which can be acceptable on the mathematical level, but have no meaning in natural language [Steinberger, 2007]. According to [phiên, 2008] another limitation is that, LSA cannot acquire polysemy (A polysemy is different words or phrases with the same meaning). LSA treats each word as if it has the same meaning despite its context, which results to a drawback in the output, as the output will depend on the occurrence average of the words. This method of producing a summary may lead to a difficulty in performing the text comparison [Steinberger, 2007]. Therefore, another approach is introduced namely "Probabilistic latent semantic analysis", this approach based on multinomial model and it had better results than LSA [Thomas, 2007].

Based on the literature survey, the problems and challenges in the area of summarization are identified, providing the basis for the work to be carried out.

III. Proposed Approach

The aim of our proposed approach is to compose an extractive summary for a document using the previously mentioned approaches. Toreach this goal,we combined three main approaches. RST, query processing, and network representation. In our proposed approach we used RST to determine the sentence type to be either nuclear or Satellitethen we were able to decide sentence priority since Nuclear sentences have more importance to the writer; therefore the nuclear sentences must have higher priority than satellite sentences. In the final summary both nuclear and satellite sentences may be included. If a satellite sentence is having a high priority, it will still be included in the final summary and its nuclear will also be dragged to the summary to emphasize the meaning to the reader. Many relations are considered in the proposed approach, these relations are imported from many sources. These sources are:

1. Pen Discourse Tree corpus - Marcu, D., Romera, M. and Amorrortu, E. (1999b) and the relation analysis done by their PDTB Research Group. http://www.seas.upenn.edu/~pdtb/
2. Mann and Thompson paper includes 24 relations [Mann, 2006], called "Classical RST".
3. Alsanie[Al-Sanie, et. al, 2005] defined 11 Arabic relations.
4. The Leeds Arabic Discourse Treebank and the LADTB –discourse annotated Arabic Treebankhttp://www.arabicdiscourse.net/

In our proposed approach, we investigate a graph-based, language-independent approach to extractive text summarization inspired by recent developments in the area of networks. We argue that if two sentences are connected in this network they probably convey complementary information about related topics, possibly about the same topic. As our goal is to construct informative extracts, the concept of complementary sentences is crucial for the development of our summarization techniques. The document under consideration is mapped into a network representation according to the adjacency and weight matrices of order N X N (where N is the number of nodes/sentences). Table 1 is an N X N Matrix for the example in Figure 1.

Table 1: N X N Matrix

	S1	***S2***	***S3***	***Sum***
S1	-	1	1	2
S2	-	-	-	-
S3	-	1	-	1

According to this matrix, it is clear that the sentence with the highest priority is S1 followed by S3 then S2. According to RST, S1has more importance to the writer while S2 and S3 are used just to describe S1, and S2 is used just to describe S3. In all summarization approaches S1 can't be excluded, while S2 and S3 can't be included without sentence S1. In our proposed approach, S1 will be included in the final summary if S1 has a high similarity to the user's query. If S2 or S3 has a high similarity to the user's query, i.e. the answer to the user's query was in a satellite sentence, in this particular case, including the satellite sentence alone will be meaningless to the user (satellite sentences are only used by the writer to emphasize the nuclear ones). So in such situations we are not going to ignore the satellite sentences after all, but we'll have to include it's nuclear sentences as well.

IV. Proposed Framework

The proposed approach represented in Figure 1 is composed of three main phases namely document summarization phase, query processing phase and Generating Final Summary phase. The objective of the first phase - document summarization phase - is to process the user's document, to define the document rhetorical relations and to rank the document's sentences. The objective of the second phase - query processing phase - is to measure the semantic similarity between the user query and the document. The objective of the last phase - generating Final Summary phase – is to generate the final summary by selecting the sentences that are mostly related to both the user's query and the document's writer.

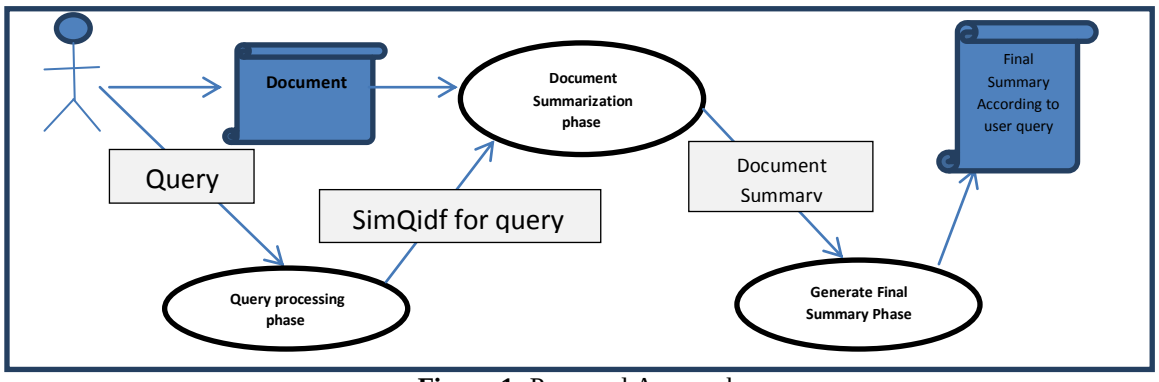

Figure 1: Proposed Approach

Query Processing Phase responds automatically to a user's query, this includes determining the relevance ranking of sentences in the document to the user's query according to a keyword search. To calculate the sentence score, the ranking is performed using "document similarities theory" [Suganya, 2014], according to [Sylvia, 2014] by comparing the deviation of angles between the document sentences' vectors and the query vector, where the query and the sentences of the document are represented as vectors, we can find the similarity between the sentences of the document and the query. This representation leads to the effectiveness in calculating the similarity between the query vector and each sentence vector in the document.

The document summarization performs the main objective which is producing the summary of the document. The first step is classifying the document submitted by the user; the classes that we used were imported from the Agrovocthesaurus [Boris, 2006]. The classification will help in the sentence ranking

component. According to the document class the sentence ranking component will determine the keywords that will raise the sentence's ranking. The second step is determining the relations between each sentence and the proceeding sentence within the same paragraph. The document is then transformed into a directed graph where an edge is drawn between each two sentences if a relation exists between these sentences. According to these relations, the sentence type is determined and a weight is given to the relation. The third step is transforming the document submitted by the user into an RST discourse graph which represents the relation between sentences in the document. The final phase namely, summary generator phase, selects the sentences that are mostly related to both the user's query and the document's writer. a rank for each sentence is measured depending on its importance to the user,elimination the similar sentences is performed, then the summary is generated according to the following rules

1- Select the highest priority sentences, the number of sentences included in summary is determined by the user.
2- If the sentence type is body then the sentence header must be included in the summary,
3- Ifthe sentence type is head and there is no body sentences underneath it are of high priority, then it is excluded.
4- If the sentence is nuclear, then it is included in the summary
5- If the RST type of the sentence is satellite then its nuclear sentence will be included even though its priority is low.

V. Results Analysis

We present a comprehensive evaluation of the automatic text summarization methods based on rhetorical structure theory (RST), claimed to be among the best ones. We also propose a new approach and compare our results to the results of our expert. To the best of our knowledge, most of our results are new in the area and reveal very interesting conclusions.We have applied the proposed system on 15 experiments; each experiment consists of a query and a document. The results of the test cases of the experiments are listed in table 12. According to the test results the system makes accurate results in most cases. 10 cases retrieved correct results, all sentences were relevant and none of the retrieved sentences were irrelevant. These cases had an f score of 1. The cases where some relevant sentences were not retrieved, was found to be due to insufficient data exported from the Agrovoc like the diseases names, the crop names etc., in case of diseases names for example, the disease اللفحه (blight disease) was not included in the Agrovoc, which made it impossible for the engine to identify it. This situation should be handled using a defined methodology.

In other cases where some irrelevant sentences were retrieved by the engine, this was due to the problem of synonyms and antonyms that we discussed earlier; this kind of error is better handled when using the RST in combination with the Semantic patterns.The RST showed a better progress in the retrieval process, but still there were many problems. First, some relations are inclusive, not exclusive, which means that there will be no keywords to recognize them, these relations are out of the scope of our research. Secondly, the relations are some times between non coherent sentences, when we tried to handle the relation between each sentence and all other sentences within the same paragraph, the results were mostly wrong, so we considered the relation between each sentence and the proceeding sentence, but this was not completely successful, cause mainly, in Arabic sentences the relations are usually between more than two sentences, besides sometimes the relation is between two sentences that are not even coherent, sometimes the relation can be between a sentence in the beginning of the paragraph and another sentence at the end of the paragraph. These problems resulted in not being able to recognize all of the relations between the sentences and therefore the priority of the sentences was not completely accurate. The lexical chain approach can be used to solve this problem.

We faced another problem when dealing with the keywords imported from the different corpora. First we thought that stemming a word would give a better result in the similarity checking process, but unfortunately that was completely misleading. We found that not stemming the words would reduce the problem of synonyms and antonyms to a great deal. When using the semantic patterns of the terms and combining them with all their senses, the result was much more accurate. The problem of Ambiguity is varies in different languages, it is increased in the Arabic language due to the overlooking rules that combine words with clitics and affixes [grammar-lexis specifications]. Another source of confusion is that the Arabic verbs can inflect for the imperative mood and the passive voice. One final problem was the problem of readability of the summary. Mostly the summary readability was fine, only in rare situations of having a hidden pronoun, the sentences are misunderstood, the RST minimized this problem to a great deal, as the hidden pronoun in most cases is between two coherent sentences, but it was still present in a very rare situations.

Table 12: The results of the test cases

ID	relevant and retrieved (A)	relevant and not retrieved (B)	Irrelevant retrieved (c)	precision	recall	F
1	4	2	0	1	0.666666667	0.8
2	6	0	0	1	1	1
3	5	0	1	0.833333333	1	0.909090909
4	6	0	0	1	1	1
5	8	0	0	1	1	1
6	5	1	1	0.833333333	0.833333333	0.833333333
7	5	2	0	1	0.714285714	0.833333333
8	5	0	0	1	1	1
9	7	0	0	1	1	1
10	4	2	0	1	0.666666667	0.8
11	6	1	1	0.857142857	0.857142857	0.857142857
12	8	0	0	1	1	1
13	5	0	0	1	1	1
14	4	3	1	0.8	0.571428571	0.666666667
15	6	0	0	1	1	1

VI. Conclusion

This paper shows how question answering systems—which aim at finding precise answers to questions—can be improved by exploiting summarization techniques to extract more than just the answer from the document in which the answer resides. This is done using a graph search algorithm which searches for relevant sentences in the discourse structure, which is represented as a graph. The Rhetorical Structure Theory (RST) is used to create a graph representation of a text document. The output is an extensive answer, which not only answers the question, but also gives the user an opportunity to assess the accuracy of the answer (is this what I am looking for?), and to find additional information that is related to the question, and which may satisfy an information need. This has been implemented in a working multimodal question answering system where it operates with two independently developed question answering modules. The classification process has two phases the first of which is done offline while the second one is done online.

We presented a system for document summarization satisfying user query based on RST. We proposed an approach and applied it on twenty experiments. The experiment results showed success in most cases and it triggered some problems. They are, insufficiency of data imported from different corpora, in addition to the irrelevant sentences and hidden pronoun included in the summary. The problem of synonyms and anatomies is also one of the crucial problems to be monitored.

REFERENCES

[1] [Barzilay, 1997], Barzilay, R. and M. Elhadad, "Using lexical chains for text summarization". In Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and the 8th European Chapter Meeting of the Association for Baxendale,ГР.Г(1958).ГMachine-madeГindexГforГtechnicalГliteratureГ-ГanГexperiment.ГIBM Journal of Research

[2] [Boris,2006], Agrovoc Web Services-Improved, real-time access to an agricultural thesaurus, 2006, Boris Lauser,. MargheritaSini, Gauri. Salokhe,

[3] [Dolores, 2008] M. Dolores Ruiz, Antonio B. BailónEnterprise, 2008, "Summarizing Structured Documents through a Fractal Technique", springer Information SystemsLecture Notes in Business Information Processing Volume 12, 2008, pp 328-340

[4] [Edmundson, 1969], Edmundson, H.P. 1969. "New methods in automatic abstracting". In: *Journal of the Association for Computing Machinery 16 (2).* 264-285. Reprinted in: Mani, I.; Maybury, M.T. (eds.) *Advances in Automatic Text Summarization.* Cambridge, Massachusetts: MIT Press. 21-42, 1969.

[5] [Luhn, 1958]Luhn Hans Peter, 1958"The Automatic Creation of Literature Abstracts". IBM Journal of Research Development, 2(2):159–165.

[6] [Mann, 2006]MaiteTaboada and William C. Mann, 2006, "Rhetorical Structure Theory: looking back and moving ahead", sage publications London.

[7] [Martin , 2008] Martin Hassel, 2008, "Resource Lean and Portable Automatic Text Summarization", Doctoral Thesis Stockholm, Sweden.

[8] [Mikael , 2014] Mikael Kågebäck, OlofMogren, Nina Tahmasebi, DevdattDubhashi, 2014," Extractive Summarization using Continuous Vector Space Models", published in: 2nd Workshop on Continuous Vector Space Models and their Compositionality CVSC, Gothenburg Sweden.

[9] [Mohamed 2008] Hesham Ahmed Hassan, Hesham Ahmed Hassan, Ahmed Rafea, 2008, "TextOntoEx Automatic ontology construction from natural English text", science direct expert system with application 34 2008 1474-1480

[10] [Mohsen, 2012] Mohsen Tofighy , OmidKashefi , Azadehamanifar, Hamid Haj, SeyyedJavadi, 2012, "Persian Text Summarization Using Fractal Theory" , University of Science and Technology, Tehran, Iran.
[11] [phiên, 2008] ĐâyLàPhiênBảnTàiLiệuĐơnGiảnXemPhiênBảnĐầyĐủCủaTàiLiệu, 2008, "Toward Classification And Clustering In Vietnamese Web Documents", Viet Nam National University, Hanoi College Of Technology.
[12] [Radev, et. al, 2004]Radev D. R., E. Hovy, K. McKeown, 2004, "Introduction to the special issue on summarization". Computational Linguistics. Vol 28(4). 399-408.
[13] [Steinberger , 2013] Josef Steinberger April 2013, "Multilingual Summarization and Sentiment Analysis", Habilitation.
[14] [Suganya1 , 2014] B. Suganya1, 2014, "Analysis on Clustering Techniques based on Similarity of Text Documents", International Journal of Advance Research in Computer Science and Management Studies Research. Available online at: www.ijarcsms.com.
[15] [Sylvia, 2014] Sylvia Poulimenou, Sofia Stamou, SozonPapavlasopoulos, and MariosPoulos, 2014, "Keywords Extraction from Articles' Title for Ontological Purposes", Mathematics and Computers in Science and Engineering Series.
[16] [Thomas, 2007] Thomas Landauer, Susan T. Dumais, 2007"A Solution to Plato's Problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge".
[17] [Wang, 2006] Wang, F.L., & Yang, C.C. 2006. "The impact analysis of language differences on an automatic multilingual text summarization system", Journal of the American Society for Information Science and Technology, 57, 684–696.
[18] http://www.arabicdiscourse.net/
[19] http://www.medar.info/BLARK/unannotated_corpora.php
[20] http://www.elda.org/medar_lri/monolingual_corpora.php
[21] http://www.seas.upenn.edu/~pdtb/
[22] http://www.arabicdiscourse.net/
[23] [FAO] AGROVOC Multilingual agricultural thesaurus,Food and Agriculture Organization of the United Nations (FAO), http://aims.fao.org/standards/agrovoc, URI: http://ring.ciard.net/node/1910