Journal of Asian Scientific Research
ISSN(e): 2223-1331/ISSN(p): 2226-5724

journal homepage: http://www.aessweb.com/journals/5003

AUTOMATIC TEXTS SUMMARIZATION: CURRENT STATE OF THE ART

Nabil ALAMI[1]† --- Mohammed MEKNASSI[2] --- Noureddine RAIS[3]

[1,2,3]Laboratory of Computer and Modeling (LIM), University Sidi Mohamed Ben Abdellah (USMBA), Fez, Morocco

ABSTRACT

To facilitate the task of reading and searching information, it became necessary to find a way to reduce the size of documents without affecting the content. The solution is in Automatic text summarization system, it allows, from an input text to produce another smaller and more condensed without losing relevant data and meaning conveyed by the original text. The research works carried out on this area have experienced lately strong progress especially in English language. However, researches in Arabic text summarization are very few and are still in their beginning. In this paper we expose a literature review of recent techniques and works on automatic text summarization field research, and then we focus our discussion on some works concerning automatic text summarization in some languages. We will discuss also some of the main problems that affect the quality of automatic text summarization systems.

Keywords: Automatic text summarization, Clustering, RST, Graph theory, Latent semantic analysis, Fuzzy logic, Machine learning, Topic identification.

Contribution/ Originality

This study is one of very few studies which have investigated on automatic text summarization field. The paper's primary contribution is analyzing some of the main problems that affect the quality of automatic text summarization systems in different languages studied in this paper. The purpose of this analyze is to find a good approach that can be applied to Arabic text.

1. INTRODUCTION

With the advent of the Internet, and the multiplicity of media mass storage, the amount of electronic documents and textual data became huge. The human, unable to manually handle large text volumes, must find automatic analysis methods adapted to automatic processing of personal data. These methods fall into the field of natural language processing (NLP). Among the most

† Corresponding author

DOI: 10.18488/journal.2/2015.5.1/2.1.1.15
ISSN(e): 2223-1331/ISSN(p): 2226-5724

popular applications include machine translation, automatic summarizer, information retrieval, text mining, spell correction, speech synthesis, speech recognition or handwriting recognition. Our study focuses on automatic text summarization which allows user to decide whether the document is of interest or not, without looking at the whole document by extracting brief information without losing the meaning and important information in the original text. The first attempt at automatic text summarization is started in the late fifties with Luhn [1]. According to Mani and Maybury [2], text summarization is the process of distilling the most important information from a source to produce an abridged version for a particular user and task. Text summarization has experienced a great development in recent years, and a wide range of techniques and paradigms have been proposed to increase researches in this field witch become a new challenges because the new emerging technologies. So it is essential to analyze its progress and present the most important techniques and to investigate recent researches works made on this field. Text Summarization approaches can be classified into extractive and abstractive summarization. An extractive summarization method consists of extraction important sentences from the source document and presented it in the same order as a summary. The importance of sentences is decided by the weight of each sentences based on statistical and linguistic features. An Abstractive summarization method consists of understanding the main concepts in original document and presents it in a shorter text. It requires human knowledge, statistical methods and linguistic methods. In this paper, we present an investigation on automatic text summarization research works and approaches proposed in some language, focusing especially on the last six years and the new types of summarization methods that have appeared in recent years, such as Graph Theory, Latent Semantic Analysis (LSA), Fuzzy logic and other methods and techniques.

2. AUTOMATIC SUMMARIZATION OF ENGLISH TEXTS

In this part, we will outline new and existing approaches to automatic summarization of English texts. Note also that most of these approaches are generic and apply to other languages especially European language.

2.1. Cluster Based Approach

Some automatic summarization systems use clusters to generate a significant summary approaching the various topics of the document. The documents are represented using term frequency-inverse document frequency (TF-IDF). In this context the term frequency (TF) is the average number of occurrence (by document) in the cluster. The topic is represented by words of which the value TF-IDF is higher in the cluster. The selection of the relevant sentences is based on the similarity of the sentences with the topic of cluster *Ci*. In Zhang and Cun-He [3] the measurement of similarity between the sentences is calculated according to the similarity of the words between two sentences and the semantic similarity of words. Then, the K-means method is used to gather the sentences of the document in the clusters. Yulia, et al. [4] propose a method based on the following steps: terms selection, terms weighting and sentences selection. In the first step, one of the three models of the text is extracted: Bag-of-words model, n-grams model and Maximal frequent Sequence (MFS) model. In the second stage, the terms are weighted by using the

Boolean method, TF, IDF or TF-IDF. In the third steps, the Expectation-maximization algorithm (EM) is used to form similar groups of sentences in order to obtain a sentence representing each cluster to be included in the summary.

2.2. Topic-Based Approaches

Moreover Teng, et al. [5] propose an approach which combines the automatic topics identification technique with the terms frequency method. This methodology consists of calculating initially the similarity between the sentences, then carry out the identification of the subject covered by gathering similar sentences in clusters. In a second stage, and based on terms frequency, the projecting sentences are selected starting from the local topics already identified. Kuo and Chen [6] use not only the frequency of terms to detect relevant information in a text, the authors also use informative words and event-driven. This type of words indicates concepts and the important relations which can be used to detect important sentences in the text.

2.3. Approaches Based on Lexical Chains

The automatic text summarization by lexical chains was introduced in Barzilay and Elhadad [7]. This method uses the WordNet database knowledge to determine the relations of cohesion between terms then composes chains based on these terms. Scores are given based on the number and type of relation in the chains. The final summary contains the sentences where the strongest chains are very concentrated. A similar method with a graphs using the knowledge bases of WordNet and Wikipedia was presented in Pourvali and Abadeh Mohammad [8]**.** This method consists initially in finding the exact meaning of each word in the text using WordNet, then builds the lexical chains and removes those which have a weak score compared with the others. The structure of the lexical cohesion of the text can be exploited to determine the importance of a sentence.

2.4. Discourse Based Approaches

New techniques were born to solve the problem of automatic summarization; these techniques are based on the analysis of discourse and its structure. Among these techniques we quote the *Rhetorical Structure Theory (RST)*. Moreover, Khan, et al. [9] combines the RST with a generic summarizer to add linguistic knowledge to the process of automatic summarization. But this mixed approach could not improve the results obtained by the generic summarizer. In other words, the disadvantage of this approach is found at the analyzer level which could not detect all RST relations, in fact, a good analysis and languages knowledge could have improved the output of the summary system. In the paper published by Li Chengcheng [10], the system extracts the rhetorical structure of the text and the components of the rhetorical relations between the sentences, then calculates the weight of each sentence of the text according to its utility and removes the least important parts of the structure having a weak weight.

2.5. Graphs Based Approaches

LexRank and TextRank are the most important algorithms used in automatic summarization system based on graph method. In the same context, [11] proposed a method based on graphs algorithm for automatic texts summarization. This method consists in building a graph from the text. Nodes of the graph are represented by the text sentences, for each sentence there is a node. The edge of the graph represent connection (lexical or semantic) between the sentences, this connection is evaluated by calculating the similarity between the sentences. The weight of each node is calculated by using the function COS. After that the summary is made up by taking the shortest way which starts with the first sentence of the original text and finishes with the last sentence. In addition, SUMGRAPH [12] and Time stamped Graph [13] are two automatic summarization systems based on graphs.

2.6. Latent Semantic Analysis (LSA) Based Approaches

LSA is an algebraic-statistical method that extracts and represents semantic knowledge of the text based on the observation of the co-occurrence of words. This technique aims to builds a semantic space with very large dimension from the statistical analysis of the whole co-occurrences in a corpus of texts. The starting point of LSA consists of a lexical table which contains the number of occurrences of each word in each document. Gong and Liu [14] proposed an automatic summarization system of news text with the use of LSA as a way to identify the important topics in the documents without using lexical resources like WordNet. In this way, the SVD is applied to matrix A to decompose into three new matrices as follows: $A = UWV^T$. The suggested that the row of the matrix V^T can be considered as various topics covered in the original text, while each column represents a sentence in the document. And finally, in order to produce an extractive summary, they consider each row of matrix V^T consecutively, and select the sentence with the highest value. In Yeh, et al. [15] another method using LSA was proposed. It is a mixed approach between graphs based method and LSA based method. After using the SVD on a matrix of words per sentence and reduction of these dimensions, the corresponding matrix $A'=U'\Sigma'V'^T$ is built. Each column of A' denotes the sentence semantic representation which is used, instead of an occurrence frequency vector of keyword, in order to represent document as a graph of relations between sentences. A ranking algorithm is then applied to the resulting graph. In the same context, a Non-negative Matrix Factorization (NMF) algorithm was proposed in Mashechkin, et al. [16], instead of the SVD, to reduce the dimensions of the matrix. The idea is that from the matrix A whose columns are the n sentences of text and rows are the m terms, and since the elements of A are non-negative, so NMF can then decompose the matrix A into two positive matrices W_k and H_k. in order to approximate the matrix A in the decomposition form $A_k \approx W_kH_k$. Matrices W_k correspond to the mapping of space of k topics and the space of m terms, and H_k correspond to the representation of the sentences in the space of topics. Subsequently, we can find out what the terms of the text best characterize each topics associated with the columns of the matrix W_k. After this decomposition, and based on this representation.

2.7. Approach Based on Fuzzy Logic

In Farshad, et al. [17], another approach to automatic summarization has been proposed, this time it is based on a fuzzy logic. This method takes into account every feature of the text such as word frequency, similarity to keywords, similarity to the title words, sentences position, statistics of co-occurrence of lexical chain, indicative expression etc. After extracting these features and depending on the results, a value of 0-1 is assigned to each sentence of the text according to the characteristics of sentences and rules available in the knowledge base. The value obtained at the output determines the degree of importance of the sentence in the final summary. In Esther Hannah, et al. [18], different characteristics of each sentence were taken into account, such as title words, sentence length, term weight, Sentence to sentence similarity, etc. the values of these features are used by the inference engine to generate the score of each sentence of the text.

3. AUTOMATIC SUMMARIZATION OF CHINESE TEXTS

The Chinese language has experienced lately strong growth in the field of NLP. The research works in this area in number and quality have contributed to considerable advances with remarkable results. We present in this section the latest works in the automatic summarization of Chinese texts. In Xiaojun and Yuxin [19], sentences are classified according to their weight calculated on the basis of frequency words and sentences position, and then a few relevant sentences are reserved as candidate sentences. Weight based on word frequency is calculated based on the sum weight of words in the sentence according to TF.IDF method. The sentences already marked as candidates are reviewed by the EMD-MMR method (EMD: earth mover's distance, MMR: maximum marginal relevance) in order to eliminate redundancy. The idea outlined in Changwei, et al. [20] is that a text consists of a sequence of phrases where few key phrases usually cover the important content of the original text. The first step is to calculate the similarity between the sentences of the text and build clusters from similar sentences. The second step is to apply the "*Affinity Propagation Cluster*" (APC) [21] algorithm on the resulting clusters from the first step, to identify summary sentences, and then compose the summary according to the sequence in the text. The proposal in Jiang [22] was based on the assumption that the statistical methods and algorithms cannot solve the problem of not understanding the content of a document. The quality of an automatic summarization based on keywords will be enhanced if these keywords are known in advance. Thus, the author proposed a method to extract keywords based on lexical chain to generate the automatic summarization and reduce redundancy. The author uses the HowNet database to determine the relationship between the Chinese words for building the lexical chain. To improve the accuracy of extracting keywords, the author chose all the names, verbs and adjectives that appear in HowNet, and the unknown new terms that may be candidate words. After the construction of lexical chains with their weights, an algorithm is applied to the lexical chains obtained as result to extract relevant keywords. The paper published by Wang [23] proposes a strategy for texts summarization of Chinese news based on the “veins theory”. The veins theory was initially proposed for Western languages. In this work, the author tested the applicability of this theory in the Chinese language by taking into account the news text as a specific domain. This method can produce a summary of the original text based on discourse structure without requiring its semantic interpretation. In Yang, et al. [24], the author used a mixed approach between APC and LSA. After calculating the similarity with LSA, the APC algorithm is applied to group sentences into clusters. And finally to build the summary, sentences are selected from each cluster in an orderly way until the desired size of the summary is reached.

4. AUTOMATIC SUMMARIZATION OF PERSIAN TEXTS

Unlike English, the automatic summarization of texts written in Persian language presents a new line of research that has grown significantly in recent times. Here we describe some systems and works in this field. The oldest automatic summarization system of Persian texts is *FarsiSum* [25]. This is an application http client/server programmed in Perl language and designed for Persian newspapers documents in text/html form. It uses a list of stop words in Unicode format and a set of heuristic rules. The summarization process has three phases: tokenization, Scoring and extracting keywords. *Automatic Persian Text Summarizer* as described in Zohre and Mehrnoush [26] uses a hybrid approach to automatically summarize the Persian texts. In this system, the following techniques are used to select the sentences that should be included in the final summary: lexical chains, graphs based approaches, the selection of important sentences based on keywords, the number of similar sentences, similarity between sentences, and the similarity with the topic and user query. In Azadeh, et al. [27], a *Hybrid Farsi text summarization* uses a technique based on the co-occurrence of terms and conceptual property of the text has been defined. In this study, for each pair of words (two words), the degree of co-occurrence is calculated. After that, the lexical chain is created and the *n* highest ranked words are selected. Then, a graph is created; words are the nodes of the graph. Graph edges are determined based on the degree of co-occurrence between words. The score of each sentence is then calculated by adding the weight gain of all his words, and finally the *n* highest ranked sentences are selected to build summary. *PARSUMIST* is another system proposed by Shamsfard, et al. [28]. It is based on the lexical chains with an improvement in representation level and conceptual and semantic understanding of the text using all synonyms and applying the redundancy check. *PARSUMIST* architecture consists of three main parts: preprocessing, analysis and selection. The main resources used in this system are stop words (empty words), keywords and all the synonyms of the Persian words *ARSUMIST* also checks the redundancy to avoid repeating similar sentences in the summary. *Azom* [29] is another automatic summarizer system of Persian texts. It combines statistical, conceptual and structural features of a text to make the summary. The proposed approach was used for the Persian language, but it can easily be applied to other languages. After the preprocessing phase, *Azom* proceeds to the construction of the corresponding document fractal tree by extracting the text structure composed of chapters, sections, paragraphs and sentences. Then, each word is looked up in the lexical database from the Persian language to extract relations between words. In Shakeri, et al. [30], the study made by authors is based on graphs algorithm. Graphs, as explained above, are used to represent the structure of the text and understand the connections and relationships between. In this work, a graph is built, the nodes of the graph are represented by the sentences, for each sentence there is a node. The edges of the graph represent the similarity between sentences; this similarity is evaluated by considering the following features: a) Number of common words between sentences;

b) Number of keywords shared between sentences; c) Existence of words that have the same explanation; d) Existence of two words in the same paragraph.

5. AUTOMATIC SUMMARIZATION OF ARABIC TEXTS

Most systems of automatic text summarization are made to handle the most popular languages such as English, French, etc. In the other hand, there are few systems and little researches on Arabic language. Works in this area are very limited. Therefore, there is a growing need to develop systems that process and summarize electronic Arabic texts. El-Shishtawy and El-Ghannam [31] proposed an Arabic automatic text summarization by extraction-based approach. The Stemming and lemmatization of Arabic words were used in this work to calculate text features. The key-phrases are used to evaluate the importance of a sentence. Instead of using only the statistical information such as terms frequency and distance of terms, the extractor is also provided with linguistic knowledge to improve its effectiveness. *Ikhtasir* [32] is an automatic summarization system for Arabic texts proposed by *Azmi* and *Al-thanyyan*. This system incorporates a *RST* method with some features for calculation scores in order to determine the importance of a sentence in the text. It Calculate the frequency of each word in the text based on its root and use a rhetorical analysis to generate a rhetorical tree of the text in order to obtain a primary summary of the text from level six of the generated tree. Sobh, et al. [33] developed a system based on machine learning and uses a manually tagged corpus. It includes methods of Bayesian classification and Genetic Programming (GP) in an optimized way to get better results using reduced features of each sentence. The RST was used in AlSanie [34]. This system; and after the rhetorical analysis of the text; generates all possible representations of the text in the form of a rhetorical tree. Then, the summary is extracted from the highest level of the generated trees. The Lakhas system Douzidia and Lapalme [35] is based on normalization by replacing the different variants of characters by a single one, removal of stop words and lemmatization. The weight of each sentence is calculated according to the words frequency, indicative expressions, sentence position and TF-IDF value of each word in the sentence. The system described in Haboush and Al-Zoubi [36] is oriented towards the determination of the root of each word in a sentence. Based on roots found in the text, the words can be grouped into separate clusters. The authors assume that the important words in the text appear several times. Thus, the main feature considered for Arabic text summarization is a words frequency and indicative expressions to increase the importance of a sentence.

6. THE MAIN PROBLEMS IN AUTOMATIC TEXT SUMMARIZATION

After analyzing different works cited in this paper, two main problems has been detected in automatic summarization field: Problems related to natural language processing and problems related to the application of different approach and methods used in automatic text summarization field. These problems can affect negatively the quality of the resulted summarize.

Before applying any automatic summarization technique, text to be summarized must beforehand be analyzed automatically. This is because we need an appropriate representation for the textual units in statistical or semantic form depending on the chosen approach. This analysis differs from one language to another depending on the complexity and specificity of each language and advances made in this area. One of the most challenging problems in the field of Information Technology is how to do text summarization and how to employ efficient algorithm on the mass of information. In this section, we will discuss some of aspects and components which impact significantly the quality of text summarization and all issues encountered in different language studied in this paper.

6.1. Chinese

In the field of automatic text summarization, the elimination of stop words is an important issue, since their existence, results are more efficient. Stop words refer to those words that appear too frequently in the text and with no significant value. There is no need to index or use those words in research. Compared with English, no Chinese stop word list has been commonly accepted yet. Existing works on stop words identification for European languages are based on the two most important characteristics of stop words, which are the length and the frequency. In Chinese language, extracting stop words is a difficult task because the lack of spaces or other word delimiters and little diversity in the length of words. Most current researches on Chinese text summarization make use of manual stop word lists which are based on the authors' experiences.

The second problem that we are remarked in the process of summarization Chinese text is the Word segmentation. Word segmentation, as a required reprocessing for Chinese texts, is the process of identifying the boundaries between the words in natural language texts. It also used to divide a text into paragraph, sentences and words. For English and other western languages, the segmentation of texts is simple and trivial. In those languages, text can be segmented into words by using spaces and punctuations as word delimiters. However, many Asian languages like Chinese do not delimit the words by spaces. In addition to this, in western languages, there are only a small number of characters. However, the Chinese language does not have a fixed number of characters. As a key problem in Chinese information processing, the process of text segmentation make other problem very difficult, such as word-sense disambiguation, unknown Chinese word and named-entity recognition. To resolve this problem, different Chinese segmentation algorithms have been proposed in the past, but none of them has been commonly accepted as a standard.

Another problem is identified in the development of automatic Chinese summarization systems, it is the keywords identification. Keyword is one of the most important elements in the research of text summarization systems, since it would greatly affect their performances. As it is known, keyword is the key for understanding the text content and given information about it. With keywords, we could get a brief summary of the text content, and decide what will interest us. Thus, automatically extraction or identification of these keywords has been the focus of this field in the recent decades. There is no blank to mark word boundaries in Chinese text. As a result, identifying words is difficult, because of segmentation ambiguities and occurrences of unknown words. In Chinese and some Asian languages, it is difficult to identify keyword in a text since all Chinese characters can either be a morpheme or a word and there are no blank to mark word boundaries. Researches in keywords identification for Chinese language have been quite a hot topic for decades and research in this field is therefore very active. There are several kinds of classification of keywords identification or extraction methods up till now.

6.2. Persian

Several problems are identified in developing Persian text summarization systems. The existence of various written prescriptions, spaces between or in the words, structural ambiguities, recognizing, morphological changes, pos tagger, multi word expressions and non-written Ezafe construction are among this set. Ezafe marker is a short vowel added between preposition, nouns or adjectives in phrase. It is pronounced but usually not written, so it cause problems in syntactic and semantic processing and make ambiguities in tokenization and stemming. In the preprocessing stage, unlike English, and because of the complex morphology of Persian language due to the variations in word forms which have similar semantic interpretations, the Persian stemming algorithm is applied to reduce infected words to their stem, base or root form. Like Arabic language, most words in Persian are derived from a root which is usually consists of three letters.

Other problems in Persian language processing are presented in the normalization and tokenization tasks. In Persian language, tokenization (Word segmentation) has more problems compared to English language because the different in writing styles, but compared with Chinese and the existence of space, less problems occur. Space is not a delimiter and boundary sign. It may appear in different places within a word or between words or may be absent between sequential words. In addition, there are many words which can be written in different formats with space or no space. The optional nature of the white space such as adding space within a word or omitting spaces between words are the main problem in processing of Persian texts. Verb detection is another problem which rises in tokenization and affects syntax parsing. So recognizing the verb in Persian texts is a challenging problem and some work has been done in this field. STeP-1 [37] present a complete work to solve the normalization and tokenization problems. STeP-1 converts Persian texts into standard ones and tokenizes texts besides doing morphological analysis on obtained words. Some other works have been done recently in this field

On the other hand, there is a lack of language resources such as semantic lexicons, lists of stop words and cue-words, computational dictionaries, corpora, terminological Ontologies and thesaurus. Semantic lexicons and lexical Ontologies, like WordNet in English, are essential resource for natural language processing and they can make summarization of text more efficient. The lack on such resources and even on language processing tools makes text summarization a hard task for Persian language. Although there have been some efforts in creation some of the essential resources. One of the richest lexicons available for Persian is made by Eslami [38]. In addition, FarsiNet, the Persian WordNet, is the only available resource for lexical semantic.

Researches in automatic text summarization in Persian language have improved their work by using semantic features and representing a conceptual meaning of the text using synonym sets, checking eliminate redundancy. Other resources have been used such as stop-words, cue words and synonym sets of Persian words. Other works in Persian text summarization uses Persian thesaurus as a helpful knowledge to obtain the real frequencies of words in the corpus and the experimental results show a significant improvement in the case of employing Persian thesaurus rather common methods.

7. COMPARISON AMONG THE TECHNIQUES USED IN AUTOMATIC TEXT SUMMARIZATION

In this section we give a comparison of different methods and techniques used in automatic text summarization field applied for Chinese, Persian and Arabic languages.

Table7-1. Comparison among the Techniques of Text Summarization in different languages

Research work	**Language**	**Technique**	**% average Recall**	**% average Precision**	**% average F-measure**

Chinese Text Automatic Summarization Based on Affinity Propagation Cluster [20]	*Chinese*	*Clustering;*
		Sentences Similarity Measure;
		Affinity Propagation
			63,5	67,8	65,57

Chinese Automatic Text Summarization Based on Keyword Extraction [22]	*Chinese*	*keyword extraction;* *lexical chain ;* semantic similarity between terms ;	69	58	63
[38] ATSS-LS: Automatic Text Summarization Based on Lexical Chains and Structural Features	*Chinese*	*Lexical chain ;* *Structural features*	77	74	75,47

Parsumist: A Persian text summarizer [28]	*Persian*	Lexical chains;
		Graphs theory;
		synonym sets using lexical ontology;
			65	65	65

AZOM: A Persian Structured Text Summarizer [29]	*Persian*	Fractal Theory; Statistical Weighting, Structural Weighting, Conceptual Weighting	76	81	78,42
A New Graph-Based Algorithm for Persian Text Summarization [30]	*Persian*	Graph theory;	67	52	58;
Ikhtasir — A user selected compression ratio Arabic text summarization system [32]	*Arabic*	Word frequency; RST	57	72	60

An Optimized Dual Classification System for Arabic Extractive Generic Text Summarization [33]	*Arabic*	Statistical features extraction;
		Bayesian classifier;
		Genetic Programming classifier
			72,5	49	59

Arabic text summarization using Rhetorical Structure Theory [39]	*Arabic*	RST	26	34	29

As we can see, the analysis of the above table of comparison shows that in Chinese language, the performance of the system *ATSS-LS* [38] are better on average than other systems. This proves that Chinese text need semantic analysis to improve the quality of text summarization. The semantic analysis consists to determine the relationship between words by their related concept in the HowNet knowledge database. The authors use also Structural features for extracting importance sentence from the source text (keyword, cue words, title …etc). In preprocessing step, the system segments Chinese words, then it filters stop words with very little semantic contents such as empty words and some high-frequency words.

In Persian language, the good results are obtained by Zamanifar and Kashefi [29]. This is because the system use Structural features combined with Conceptual property of the text. This system uses a lexical database in order to determine the relationship between words as conceptual feature of the text. It is like ATSS-LS, the better system used for Chinese text. The difference is that Azadeh, et al. [27] uses, in addition, the statistical features of the text. The system described in Zamanifar and Kashefi [29] is developed for Persian language but easily can be applied to Arabic language.

8. CHALLENGES IN AUTOMATIC SUMMARIZATION OF ARABIC TEXT

Arabic as an important language in the world has not been studied enough, and the numbers of researches still few in Arabic natural language processing. That is because the complex nature of Arabic language. Some of those reasons are, first the different ways that certain combinations of characters can be written. Second, the wide range of derivations and inflection of functional words makes the task of morphology analysis very complex. Third, Arabic words are often ambiguous due to the tri-literal root system. In this section we will discuss some problems and challenges that can affect the quality of automatic summarization of Arabic texts.

8.1. Morphology Analysis

Morphology is the branch of linguistics that deals with the internal structure of words. It studies word formation, including affixation behavior, roots, and pattern properties. It’s consists to identify and analyzes the internal structure of words and other linguistic units, such as stem, root, affixes, part of speech...etc. Word morphology is very helpful in the process of acquiring linguistic information. It also has an important role to play in the disambiguation of word sense.

There are some morphology analysis systems in Arabic that address this issue. We can quote ARAMORPH which is limited only to analysis of words appearing in Arabic dictionaries. MORPH-2 is another morphology analysis systems based on a lexicon containing all the words (3266 root words) with their associated characteristics.

8.2. Part-Of-Speech Tagging

It consists to assign the grammatical category, such as noun, verb, adjective, adverbs, etc., to each word in the text depending to the context which it appears. Some Arabic Part-of-speech tagging systems was proposed using a combination of both statistical and rule-based techniques since hybrid taggers seem to produce the highest accuracy rates, but the most commonly used is based on a numerical approach.

8.3. Arabic Word Stemming

One of the most challenging issues in Arabic language is the word stemming. Arabic words can have different form by adding affixes (prefixes and suffixes) to the original words (root). Stemming an Arabic word consist to find the appropriate root for the giving word by removing the attached affixes. The main characteristic feature of Arabic is that most words are built up from roots by following certain fixed patterns and adding infixes, prefixes and suffixes. Arabic is highly inflectional and derivational, and words can have many different forms which makes morphology is a very complex task. In addition, written character in different ways depends on the position of letter in the word, which can add a complexity to Arabic words analysis. Therefore extracting lemma, stem or root is a hard problem for Arabic language. A good representation of Arabic text may impact positively on quality and accuracy of automatic text summarization task. One of The strengths of Arabic language is the root of words. Arabic words are generally based on a root, which mean that the root can be a base of different words with informative related meaning and with adding suffix on the root we can build a set of derivations. These derivations represent a same area. Finding a root of Arabic word (stemming); helps in mapping grammatical variations of word to instances of the same term. For example the root لعب “laaeba” is used for many words relating to “playing”, including لاعب ” , “ laaeb”, “player”, ملعب “ malaab” . Based on this consideration, multi derivations of the wording structures in Arabic language allow a semantic representation of the text which is being closer to the semantic foundations.

In our research work we can improve a quality of Arabic text summarization by using not only a statistical feature selection method but also structural and conceptual (semantic) ones. In addition, because words sharing a root are semantically related, feature selection techniques based on the root can improves a technique of clustering Arabic text which can be used as a basic method of Arabic text summarization. Our system can also be outperformed by using a knowledge database like Arabic WordNet.

9. CONCLUSIONS

In this paper, we presented an overview of the most recent advances and challenges of automatic summarization raised in the last years. At first we explained the new approaches proposed in automatic summarization of texts in different languages and especially in English texts. Then we exposed some new methods and works done to summarize Chinese, Persian and Farsi texts. In the last section, we explained some work and recent advances on automatic summarization of Arabic texts. Arabic is spoken in over 22 countries, and it is the official language of over 250 million peoples and the second for 40 million. Therefore, this language, which is very rich in terms of words categorization, deserves much more interest by scientist due to the lack of works in natural language processing field in general and in automatic summarization of Arabic texts in particular. For this, we propose in our future work to develop and implement a new method and a new system for automatic summarization designed for texts in Arabic language.

REFERENCES

[1] H. P. Luhn, "The automatic creation of literature abstracts," *IBM Journal of Research and Development*, vol. 2, pp. 159–165, 1958.

[2] I. Mani and M. T. Maybury, *Advances in automatic summarization*. Cambridege, MA: MIT Press, 1999.

[3] P.-Y. Zhang and L. Cun-He, "Automatic text summarization based on sentences clustering and extraction," Computer Science and Information Technology, 2009. ICCSIT 2009. 2nd IEEE International Conference. [Accessed 8-11 Aug], 2009.

[4] L. Yulia, H. René García, S. Romyna Montiel, R. Rafael Cruz, and G. Alexander, "EM clustering algorithm for automatic text summarization," in *Proceedings of the 10th Mexican international Conference on Advances in Artificial Intelligence - Volume Part I (MICAI'11), Vol. Part I. Springer-Verlag, Berlin,* Heidelberg, 2011, pp. 305-315.

[5] Z. Teng, Y. Liu, F. Ren, S. Tsuchiya, and F. Ren, "Single document summarization based on local topic identification and word frequency," in *MICAI '08: Proceedings of the 2008 Seventh Mexican International Conference on Artificial Intelligence*, 2008, pp. 37–41. Available: http://dx.doi.org/10.1109/MICAI.2008.12.

[6] J. Kuo and H. Chen, "Multidocument summary generation: Using informative and event words," *ACM Trans Asian Lang Inf Process (TALIP),* vol. 7, pp. 1–23, 2008.

[7] R. Barzilay and M. Elhadad, "Using lexical chains for text summarization," in *Proceedings of the ACL/EACL'97 Workshop on Intelligent Scalable Text Summarization*, Madrid, Spain, 1997, pp. 10–17.

[8] M. Pourvali and S. Abadeh Mohammad, "Automated text summarization base on lexicales chain and graph using of word net and wikipedia knowledge base," *IJCSI International Journal of Computer Science Issues, No. 3,* vol. 9, 2012.

[9] A. Khan, S. Khan, and W. Mahmood, "MRST: A new technique for information summarization," presented at the The Second World Enformatika Conference, WEC'05, 2005.

[10] Li Chengcheng, "Automatic text summarization based on rhetorical structure theory," Computer Application and System Modeling (ICCASM), 2010 International Conference. [Accessed 22-24 Oct], 2010.

[11] S. T. Khushboo, R. V. D. Dharaskar, and M. B. Chandak, "Graph-based algorithms for text summarization," presented at the Third International Conference on Emerging Trends in Engineering and Technology, 2010.

[12] K. Patil and P. Brazdil, "Sumgraph: Text summarization using centrality in the pathfinder network," *IADIS Int J Comput Sci Info Sys.,* vol. 2, pp. 18–32, 2007.

[13] Z. Lin, "Graph-based methods for automatic text summarization," Ph.D. Thesis, School of Computing National University of Singapore 2006–07, 2006–07.

[14] Y. Gong and X. Liu, "Generic text summarization using relevance measure and latent semantic analysis," in *Proceedings of the Annual International ACM SIGIR Conference on Research and Development in Information Retrieval,* 2001, pp. 19–25.

[15] J.-Y. Yeh, H.-R. Ke, W.-P. Yang, and I.-H. Meng, "Text summarization using a trainable summarizer and latent semantic analysis," *In Special Issue of Information Processing and Management on An Asian Digital Libraries Perspective,* vol. 41, pp. 75–95, 2005.

[16] I. V. Mashechkin, M. I. Petrovskiy, D. S. Popov, and D. V. Tsarev, "Automatic text summarization using latent semantic analysis," *In Programming and Computer Software,* vol. 37, pp. 299–305, 2011.

[17] K. Farshad, K. Hamid, E. Esfandiar, and K. D. Pooya, "Optimizing text summarization based on fuzzy logic," in *Proceedings of Seventh IEEE/ACIS International Conference on Computer and Information Science, IEEE, University of Shahid Bahonar Kerman*, UK, 2008, pp. 347-352.

[18] M. Esther Hannah, T. V. Geetha, and M. Saswati, "Automatic extractive text summarization based on fuzzy logic: A sentence oriented approach," in *Proceedings of the Second international conference on Swarm, Evolutionary, and Memetic Computing - Volume Part I (SEMCCO'11), Bijaya Ketan Panigrahi, Ponnuthurai Nagaratnam Suganthan, Swagatam Das, and Suresh Chandra Satapathy (Eds.), Vol. Part I. Springer-Verlag, Berlin, Heidelberg. DOI=10.1007/978-3-642-27172-4_63*, 2011, pp. 530-538. Available: http://dx.doi.org/10.1007/978-3-642-27172-4_63.

[19] W. Xiaojun and P. Yuxin, "A new re-ranking method for generic Chinese text summarization and its evaluation," in *Proceeding of: Digital Libraries: Implementing Strategies and Sharing Experiences, 8th International Conference on Asian Digital Libraries, ICADL 2005, Bangkok, Thailand. Proceedings.* Springer Berlin Heidelberg, 2005, pp. 171-175. [Accessd December 12-15].

[20] Z. Changwei, P. Qinke, and S. Suhuan, "Chinese text automatic summarization based on affinity propagation cluster," in *Fuzzy Systems and Knowledge Discovery, 2009. FSKD '09. Sixth International Conference*, 2009, pp. 425,429. [Accessed 14-16 Aug].

[21] B. J. Frey and D. Dueck, "Clustering by passing messages between data points," *Science*, vol. 315, pp. 972-976, 2007.

[22] X.-Y. Jiang, "Chinese automatic text summarization based on keyword extraction," in *Database Technology and Applications, 2009 First International Workshop*, 2009, pp. 225-228. [Accessed 25-26 April].

[23] D. Wang, "A summarization strategy of Chinese news discourse," in *Proceedings of the 2011 2nd International Congress on Computer Applications and Computational Science, (AISC) Advances in Intelligent and Soft Computing 145*, Springer Berlin Heidelberg, 2012, pp. 389-394.

[24] R. Yang, Z. Bu, and Z. Xia, "Automatic summarization for Chinese text using affinity propagation clustering and latent semantic analysis," *Web Information Systems and Mining. Lecture Notes in Computer Science*, vol. 7529, pp. 543-550, 2012.

[25] N. Mazdak, "Farsi sum-a Persian text summarizer," Master Thesis, Department of Linguistics, Stockholm University, 2004.

[26] K. Zohre and S. Mehrnoush, "A system for automatic persian text summarization," presented at the 12th International CSI Computer Conference, (In Persian), 2007.

[27] Z. Azadeh, M.-B. Behrouz, and S. Mohsen, "A new hybrid farsi text summarization technique based on term co-occurrence and conceptual property of the text," presented at the 9th ACIS International Conference on Software Engineering, Artificial Intelligence, Networking and Parallel/Distributed Computing, 2008.

[28] M. Shamsfard, T. Akhavan, and M. E. Jourabchi, "Parsumist: A Persian text summarizer," Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference.[Accessed 24-27 Sept], 2009.

[29] A. Zamanifar and O. Kashefi, "AZOM: A Persian structured text summarizer," in *Proceedings of the 16th International Conference on Natural Language Processing and Information Systems (NLDB'11). Springer-Verlag Berlin*, Heidelberg, 2011, pp. 234-237.

[30] H. Shakeri, S. Gholamrezazadeh, M. A. Salehi, and F. A. Ghadamyari, "New graph-based algorithm for Persian text summarization," in *Lecture Notes in Electrical Engineering, Computer Science and*

its Applications International Conference; 3rd, Computer Science and its Applications, Springer, 2012.

[31] T. El-Shishtawy and F. El-Ghannam, "Keyphrase based Arabic summarizer (KPAS)," presented at the Informatics and Systems (INFOS), 2012 8th International Conference. NLP-7, NLP-14. [Accessed 14-16 May], 2012.

[32] A. Azmi and S. Al-thanyyan, "Ikhtasir — a user selected compression ratio Arabic text summarization system," Natural Language Processing and Knowledge Engineering, 2009. NLP-KE 2009. International Conference, 2009, pp. 1-7. [Accessed 24-27 Sept].

[33] I. Sobh, N. Darwish, and M. Fayek, "An optimized dual classification system for arabic extractive generic text summarization." Available: http://www.rdi-eg.com/rdi/technologies/papers.htm, 2006.

[34] W. AlSanie, "Towards an infrastructure for Arabic text summarization using rhetorical structure theory," M.Sc Thesis, Dept. of Computer Science, King Saud University, Riyadh, Saudi Arabia, 2005.

[35] F. S. Douzidia and G. Lapalme, "Lakhas, an Arabic summarization system," in *Proceedings of 2004 Document Understanding Conference (DUC2004)*, Boston, MA, 2004.

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

[37] M. Shamsfard, H. S. Jafari, and M. Ilbeygi, "STeP-1: A set of fundamental tools for Persian text processing," in *LREC 2010-8th Language Resources and Evaluation Conference,* Malta, 2010.

[38] M. Eslami, "Persian generative lexicon," in *Proceedings of The 1st Workshop on Persian Language and Computer, May 26-27 2004, University of Tehran, with the Cooperation of the Research Center of Intelligent Signal Processing (RCISP)*, Tehran, Iran, 2004.

[39] L. Yu, J. Ma, F. Ren, and S. Kuroiwa, "Automatic text summarization based on lexical chains and structural features," in *Proceedings of the Eighth International IEEE ACIS Conference*, 2007, pp. 574-578.

Views and opinions expressed in this article are the views and opinions of the authors, Journal of Asian Scientific Research shall not be responsible or answerable for any loss, damage or liability etc. caused in relation to/arising out of the use of the content.