Information Science and Technology
ISSN: 0976-917X & ISSN: 0976-9188, Volume 3, Issue 1, 2014, pp.-043-050.
Available online at http://www.bioinfopublication.org/jouarchive.php?opt=&jouid=BIJ0000002

A TAXONOMY FOR TEXT SUMMARIZATION

OTHMAN B.M.M.[1]*, HAGGAG M.[2] AND BELAL M.[2]

[1]Institute of Statistical Studies and Research, Cairo University, Giza, Egypt.
[2]Department of Computer Sciences, Faculty of Computer Sciences and Information Systems, Helwan University, Cairo, Egypt.
*Corresponding Author: Email- b.m.m.othman@gmail.com

Received: November 03, 2012; Accepted: March 20, 2014

Abstract- Text summarization is the branch of NLP where a computer summarizes a text. A text is entered into the computer, a specific technique is applied, and then a summarized text is returned. This summary should be a non-redundant extract from the original text. There are many categories for summarization: single document, multi-document, extractive, abstractive, informative, indicative, user- focused, generic, statistical, linguistic, and machine learning approach based.

However, most of the surveys that concerned with text summarization was covering a specific perspective of the field and didn't clearly illustrate the whole picture of the state- of- the art they covered; the purpose of this survey is to clearly illustrate the whole picture of the previous work in the field of text summarization introducing a general taxonomy that covers all possible aspects of categorizing the text summarization field with clear comparison based on all aspects and features that the text summarization field could have.

In this paper all approaches for single document and multi-document summarization, extractive and abstractive summary construction methods, and informative and indicative information content will be introduced. Additionally, query-based and generic summary triggers, statistical, linguistic and machine learning methods for choosing the most relevant sentences from documents had been explored. All these approaches will be introduced and discussed.

Keywords- Text summarization, survey, taxonomy

Citation: Othman B.M.M., Haggag M. and Belal M. (2014) A Taxonomy for Text Summarization, Information Science and Technology, ISSN: 0976-917X & ISSN: 0976-9188, Volume 3, Issue 1, pp.-043-050.

Introduction

As the amount of information increases rapidly, systems that can summarize one or more documents become increasingly desirable. Recent research has investigated types of summaries, methods to create them, and methods to evaluate them. Text summarization has been developed and improved in order to help users manage all the information available these days. A summary can be defined as a text that is produced from one or more texts, that contain a significant portion of the information in the original text(s), and that is no longer than half of the original text(s). The main goal of a summary is to present the main ideas in a document in less space. If all sentences in a text document were of equal importance, producing a summary would not be very effective, as any reduction in the size of a document would carry a proportional decrease in its informativeness [1]. As interest in text summarization emerged as early as the fifties [2]; which means it is more than 60 years of progressive development appeared in this field.

As literature agreed, the definition of Natural Language Processing is a field of computer science and linguistics concerned with the interactions between computers and human languages. One of the cutting-edge areas in NLP that currently draws a large amount of research activity is automatic text summarization. Text summarization domain is one of the applications or sub-fields of the NLP. Summarization is the creation of a shortened version of a text by a computer program. The product of this procedure still contains the most important points of the original text. The phenomenon of information overload has meant that access to coherent and correctly-developed summaries is vital. As access to data has increased, the interest in automatic summarization increased as well. An example of the use of summarization technology is search engines such as Google. Technologies that can make a coherent summary, of any kind of text, need to take into account several variables such as length, writing-style and syntax to make a useful summary.

To text summarization domain, there are some related domains; they are a super field, a sub field, or a domain which overlap with text summarization domain; among these fields: text generation, machine learning, and text mining.

Text generation is the field of Natural Language Generation (NLG) which is the natural language processing task of generating natural language from a machine representation system such as a knowledge base or a logical form. In a sense, one can say that an NLG system is like a translator that converts a computer based representation into a natural language representation. However, the methods to produce the final language are very different from those

of a compiler due to the inherent expressivity of natural languages. The most successful NLG applications have been data-to-text systems which generate textual summaries of databases and data sets; these systems usually perform data analysis as well as text generation.

Machine learning is a scientific discipline that is concerned with the design and development of algorithms that allow computers to evolve behaviors based on empirical data, such as from sensor data or databases. A major focus of machine learning research is to automatically learn to recognize complex patterns and make intelligent decisions based on data; the difficulty lies in the fact that the set of all possible behaviors given all possible inputs is too complex to describe generally in programming languages. Artificial intelligence is a closely related field, as also probability theory and statistics, data mining, pattern recognition, adaptive control, and theoretical computer science. Text summarization uses the field of machine learning especially for extractive summaries; as in such a case sentences of each document are modeled as vectors of features extracted from the text. The summarization task can be seen as a two-class classification problem, in which you train a classifier to specify which sentences to be taken in the summary and which sentences to be not.

Text mining, sometimes referred to as text data mining, roughly equivalent to text analytics, refers to the process of deriving high-quality information from text. High-quality information is typically derived through the divining of patterns and trends through means such as statistical pattern learning. Text mining usually involves the process of structuring the input text (usually parsing, along with the addition of some derived linguistic features and the removal of others, and subsequent insertion into a database), deriving patterns within the structured data, and finally evaluation and interpretation of the output. 'High quality' in text mining usually refers to some combination of relevance, novelty, and interestingness. Typical text mining tasks include text categorization, text clustering, concept/ entity extraction, production of granular taxonomies, sentiment analysis, document summarization, and entity relation modeling (i.e., learning relations between named entities).

Related Work

Gupta, et al. [2] proposed a survey that covered deliberately most of the extractive techniques used for text summarization for about 15 years, but, still lacking the deliberate covering of the abstractive text summarization techniques.

Israel, et al. [3] proposed a survey concerned with the state-of-the art that used to summarize documents in a collection with a concentration toward a particular external request (i.e. query, question, topic, etc.), or focus; which means only the user- focused techniques for multi- document summarization are covered. Although this paper not only briefly explores the state-of-the-art in automatic systems techniques, but also a comparison with human summarization activity. But, still generic techniques and single document summarization techniques remain uncovered.

Ontotext, et al [4] showed in their paper an intriguing consideration to query- based approaches used in the last 5 years from its publishing date. It covered approaches based on graphs, linguistics and machine- learning but only for user- focused techniques without even a quick gaze to the previous work in generic techniques.

Das and Martins [5] showed the big effort made to cover the most of the techniques used in summarization field for a long period of time, but it gave more emphasis to extractive techniques more than abstractive techniques which are not covered with the same interest. Also it concerned with statistical approach specifically. Beside the lack of visual aids and only depending on large blocks of text to compare between previous works in each aspect generates a little confusion in understanding it.

Ding [6] and Satoshi Sekine, et al. [7] showed the real creativeness in covering the features of multi- document summarization and gave a really detailed view of some selective systems that used multi document summarization. But, still the single document summarization world out of scope.

As old is gold; Radev, et al. [8] had introduced a comprehensive work to survey the state-of-the art till its publishing date. But only the lack of generalization, visualization and diagrams made a kind of misleading too. While Feldman, et al. [9], showed lots of extraction methods and techniques used which opened the door for new strategies to come over.

Concluded from the aforementioned surveys, each survey handled the topic from its own perspective, none of them compared all features. Instead, each one has a partial view; most of them tend to classify them from the single or multi document view. Others classified them according to extractive or abstractive view. Hence, the purpose of this survey is to propose a new classification way by presenting a general enumeration of all aspects and categories. Especially this survey presents:

- Taxonomy for text summarization that could help as a background for researchers in this field.
- Check matrix that contains all text summarization categories and about the 35 works of the last 15 years. This matrix helps in verifying the limitation or completeness of text summarization approaches. So a single look to this table, one could know which papers to read to understand a specific aspect or a combination of aspects.

Systems on Text Summarization

Text summarization had been implemented in lots of large and effective systems, such as:

- SciSumm system (2011): this system embodies unsupervised approach to multi-document summarization of scientific articles, in this approach the collection of documents is a list of papers cited together within the same source article, otherwise known as a co-citation. At the heart of the approach is a topic based clustering of fragments extracted from each co-cited article and relevance ranking using a query generated from the context surrounding the co-cited list of papers. The system apply this approach to the 2008 ACL Anthology to provide a web based interface for viewing and summarizing research articles in it [10].
- Related Work Summarization (ReWoS) system (2010): this system given multiple articles as input, a related work summarization system creates a topic-biased summary of related work specific to the target paper. It takes in set of keywords arranged in a hierarchical fashion that describes a target paper's topics, to drive the creation of an extractive summary using two different strategies for locating appropriate sentences for general topics as well as detailed ones. The results show an improvement over generic multi-document summarization baselines in a human evaluation [11].

- Citation-Sensitive In-Browser Summarizer (CSIBS) (2009): CSIBS is a research tool that builds a summary of the cited document, bringing together meta-data about the document and a citation-sensitive preview that exploits the citation context to retrieve the sentences from the cited document that are relevant at this point, it developed based on a user requirements analysis. The summary is shown as a pop-up text box within the same browser in which the citing document is being viewed (for example, Adobe Acrobat Reader or a web browser) [12].
- Multiple Alternative Sentence Compressions (MASC) frameworks (2007): the MASC framework uses a sentence compression module to generate multiple compressions of source sentences in combination with a candidate selector to construct a summary from the compressed candidates. The selector uses a combination of static and dynamic features to select candidates that will maximize relevance while minimizing redundancy within the summary [13].
- NEO-CORTEX system (2007): this system uses an approach to topic-oriented multi-document summarization (MDS), builds on the CORTEX system (The COndensation et R´esum´es de TEXtes) work in single-document summarization (SDS). The NEO-CORTEX is proved to be an effective system and achieves good performance on topic-oriented multi-document summarization task, it is sensitive to the sentence segmentation, and its key point is the ability of the system to be language independent [14].
- SumUM system (2002): that is a text summarization system that takes a raw technical text as input and produces an indicative informative summary. The indicative part of the summary identifies the topics of the document, and the informative part elaborates on some of these topics according to the reader's interest [15].
- SUMMARIST system (1997): SUMMARIST is an attempt to create a robust automated text summarization system, based on the 'equation' summarization = topic +identification + interpretation + generation. The goal of SUMMARIST is to provide both extracts and abstracts for arbitrary English input text. It combines symbolic world knowledge (embodied in WordNet, dictionaries, and similar resources) with robust NLP processing (using IR and statistical techniques) [16]. An important aspect to be addressed is the combination of the outputs of various modules in each stage [17].

Taxonomy for Text Summarization

In this section, a taxonomy for text summarization is proposed. It presents a general enumeration for all aspects and categories of text summarization. As shown in [Fig-1], types of summaries according to the number of source inputs could be single document when only one document is the input or multi- document when the input is a cluster of text documents. According to summary construction method depending on the nature of text representation in the summary, text summarization is called extractive, where an extract is a summary consisting of a number of salient text units selected from the input or abstractive, where the abstract is a summary, which represents the subject matter of the article with the text units, which is generated by re-drafting unit outstanding selection of inputs. An abstract may contain some text units, which are not present in to the input text.

We can also categorize the text summarization based on the summary target to user- focused (query-focused) summaries which designed to the requirements of a particular user or group of users, and generic summaries that aimed at a broad community of readers, which offer in a concise manner the main topics of a given text. And based on the information content of the summary, it can be categorized as informative summary which covers all salient information in the document at some level of detail, i.e., it will contain information about all the different aspects such as purpose of the article, scope, approach, results and conclusions etc. and indicative summary presents an indication about the purpose of an article and approach to the user for selecting the article for in-depth reading. For example, an abstract of a medical research article is more informative than its headline.

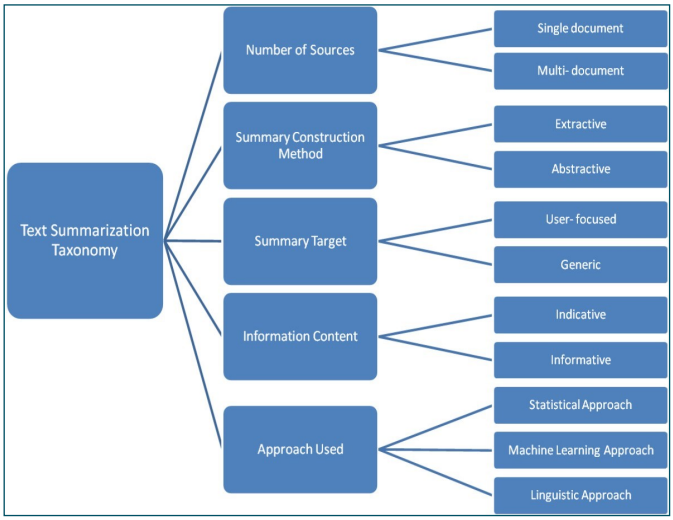

Fig. 1- Shows the text summarization taxonomy

Summarization methods can be roughly grouped into three categories [18]: Statistical approach; summarizes without understanding, but rather depends on the statistical distribution of certain properties, Linguistic approach; summarization based on these method requires knowledge of the language so that the computer can analyze the sentences semantically and then decide what sentences to choose considering the position of the verb, subject, noun and etc. these methods are more difficult than statistical methods, and Machine Learning Approach; each text element, is considered as an N-dimensional vector. It is therefore possible to use some of the metric in this space to measure the similarity between the elements of the text. There are some methods that use the hybrid combination of these approaches.

Text Summarization Techniques

In this section, a brief description of the contributions made in text summarization field is presented according to the taxonomy in [Fig-1], were 39 related works covered and presented in the comparative study in [Table-1], covering the state of the art for the last 15 years.

Agarwal, et al. [10] presented the system: SciSumm to provide a web interface to displaying, viewing, and summarizing research articles in the ACL Anthology corpus (2008). On the other hand, in Kowsalya [19] model; an extractive summary produced for a given set of documents on the basis of the sequence of word models by extracting the most frequent sequences of a particular text.

Negi, et al. [20] provided -somehow- a new technique to extract as

they used pattern recognition techniques to improve the performance of the retrieval of relevant information. And the design and implementation of the proposed systems have the means to summarize the information from the retrieved set of documents or corpuses, and measured by the quality system by how much is useful for typical users of the system. In the basic approach, a query is generated from "ideal" document that meet the need of information. The system's function then is to estimate the likelihood of each document in the group for being a perfect document and classified accordingly.

Kiyoumarsi, et al. [21], Khosraviyan, et al. [22], and Fattah and Ren [23] used the same way to train a summarizer to improve the way of extraction to serve their different approaches. The three of them used machine learning technique to train and test the same summarizer although Kiyoumarsi, et al. [21] used Cellular Automata, Khosraviyan, et al. [22] used Genetic Programming, and Fattah and Ren [23] used Mathematical Regression to obtain a suitable combination of feature weights. While Galgani, et al. [11] used a combination of all available approaches to build and evaluate their summarizer.

Hariharan [24] provided studies on the combination of many parameters that affect the extraction algorithm, and it showed that the frequency term approach along with the weight position gives better results, while adding the node with the weight, produces results that were much better than the former approach. While Hoang and Kan [25] proposed the (ReWoS) system as mentioned in section 3; given numerous articles as input, a related work summarizing system creates a biased summary of the topic related work specified in the target paper.

Nishikawa, et al. [26] proposed algorithm to summarize the opinion that takes into account the content and coherence. It has achieved this with an integer linear programming formulation (ILP), which is a powerful combination of the problem of maximum coverage and the traveling salesman problem, which are applicable on a large scale to generate the text and summarize it. On the other hand, Wan, et al. [12] presented the (CSIBS) research tool as mentioned in section 3, which builds summary of the document bringing together the collection of metadata about the document, and sensitive citation preview.

Haghighi and Vanderwende [27] made an exploration of probabilistic generative models for multi-document summary. Starting from a simple model based on word frequency, they are building a series of models each pumping more structure to represent the group content of the document. Their model, HIERSUM, used for hierarchical LDA-style model to represent content privacy as the hierarchy of the vocabulary of the distribution of the subject. So this paper showed that the use of structured topic models can benefit the quality of summarization standards as measured by automatic and manual metric.

Gillick, et al. [28] in this paper provided an integer linear program (ILP) to derive an accurate model according to the maximum coverage to summarize automatically. Compare this model, which operates in the sub-sentence, or "concept" level, to level governance model, with previously resolved to leave. While Sarkar [29] combines several features of the selected domain with some other known features such as the term frequency and position to improve the performance of summarization in the medical field. And Kyoomarsi, et al. [18] presented an approach to the design of automatic text summarizer that generates summary using fuzzy logic to obtain better results compared with previous methods.

Nastase [30] made a model that allows users to determine a request for information in the form of a set of one or more sentences or questions. To "understand" it, it expands the query using encyclopedic knowledge in Wikipedia. The expanded query associated with the documents attached to it by activating the publication in the graph that represents the words and grammatical links in these documents. The subject expanded words and activated nodes in the graph are used for the production of extractive summary

Wong, et al. [31] investigated combined features of sentence to extractive summarization. To determine the weights of different features, they used under the supervision of learning a framework to determine the how likely a sentence is important. They then used the seeded semi-supervision of learning to combine the data described and data is called instead of the one overseen by a summary in the context of supervised because the supervised one is time consuming and expensive. On the other hand, Wang, et al. [32] proposed a new framework based on semantic analysis on the sentence level (SLSS) to capture the relationships between sentences in a semantic similarity matrix and the building of governance, which is based on it to perform the proposed symmetric nonnegative matrix factors (SNMF) to the group of sentences.

Wan and Yang [33] proposed the cluster-based Markov Conditional random walk model (ClusterCMRW) and the model of cluster - based HITS (ClusterHITS) to get the full information collected on the level of cluster as the Markov random walk stochastic model exploited recently for multiple- documents summarizing, by taking advantage of the link relations between sentences in a document, under the assumption that all the provisions cannot be distinguished from each other. While Hennig [34] proposed a new method based on probabilistic latent semantic analysis, which allows him to represent sentences and queries as probability distributions over the underlying issues. This approach combines query- focused and the features of the calculation and objectivity in the space subject to the discretion of the underlying summary, and the importance of the provisions.

Elena Lloret, et al. [1] in their paper explored the possibility of using Textual Entailment to help in the task of summarizing the text. Proved that, if the integration of textual entailment and features in the systems summarized to generate summaries of partial or final, this can lead to good improvements. Additionally, Madnani, et al. [13] proposed the (MASC) framework as we mentioned in section 3, born multi-document summary by generating compressed versions of the sentences concise and source candidates to use the features of these candidates are likely to build summaries. It combines the analysis approach and trim with a new technique for the production of pressure on the alternative multi-source provisions. It also describes experiments using the words on the basis of a new feature to examine the repetition.

Yih, et al. [35] used a simple procedure based on maximizing the number of words of useful content. They appointment a score on each term in the cluster of documents, using only the frequency and position information, and then find a range of sentences in the document that maximizes the sum of these grades, subject to the limitations of length. While, Boudin and Manuel [14] designed the NEOCORTEX system as we mentioned in section 3, it proved to be an effective system and achieves good performance in the subject-oriented multi-document summary task, it is sensitive to the retail wholesale, and its main point is the ability of the system to be lan-

guage independent.

Nenkova, et al. [36] studies the contribution to the summarization of three factors related to frequency: content word frequency, the selection and composition functions, and the sensitivity of context. Besides Bollegala, et al. [37] presented a bottom-up approach to arrange the strings for the extraction and multi-document summary. To capture the associations and the order of text of two parts, they defined four criteria, in chronological order, local convergence, precedence, and the succession. While, Conroy, et al. [38] presented an "Oracle" score, based on the probability distribution of unit-grams in the summaries of humans, and then indicates that result with the oracle score, one can generate extracts that are recorded, on average, better than summaries of humans. In addition, it introduces an approximation to the point that oracle produces a system with the performance the best known for the document understanding conference 2005 (DUC) evaluation.

Hachey and Grover [39] reported a set of tests for the classification of provisions of sentences that is whether they should be part of the summary extraction or not. The task of extracting rule is part of the system automatically summarized in the legal field. And Seki [40] in his doctoral abstract proposed a new method for automatically summarization with a focus on the type of document – document genre - and the text structure -functional aspects of the text and the text is divided into units or components of the sentence, according to their roles and function - and to verify its effectiveness.

Sripada, et al. [41] presented three techniques to generate summaries based on extracting the list including the formulation of a graph. The first method uses the degree of importance calculator judgment on the basis of attributes tag is different, and the degree of similarity semantic, and the second method of rule sets based on the degree of similarity Semantic and chooses one representative from each group to be included in the summary that was created, and the third method is the problem formulation based on the drawing graphic which is created on the basis of summaries found cliques in the constructed graph.

Elhadad [42], in his work examined two types of user tailoring: Single, facts specific and of interest to the reader, and class-based, any degree of expertise of the reader. This summarizer is trying to provide all types of users with tailored combinations of findings in the report in clinical studies. On the other hand, Jaoua and Hamadou [43] presented a method which deals with the summary as of the minimum unit for the extraction and uses the two steps; Generation: combines text sentences to produce a population of extracts. And Category: assesses each extract using the global standards in order to select the best one.

Saggion and Lapalme [15] investigated the SumUM system as mentioned in section 3, which takes raw text as input and produces technical brief explanatory information. And, Schiffman, et al. [44] trained a summarizer that uses several new strategies to determine the sentences interesting and informative, including an innovative importance derived from the analysis of a large corpus. System also computes concept frequencies rather than word frequency as an additional measure of importance. It integrates these strategies with a number of heuristics summarize familiar with the provisions to rank sentences. While Nomoto, et al. [45] showed how to build a generic document summarizer for a single input. The summarizer consists of the following two operations: Search for Diversity: The Search for different areas on the subject of the text. And reduce Redundancy, from every subject area to identify the most important sentence, and take that sentence as a representative of the region. A summary then is a set of sentences generated by the repetition limit.

Barzilay, et al. [46] described two naive techniques of ordering (the Majority Ordering "MO", and the Chronological Order "CO". They have conducted additional tests to determine the Cohesion constraint; they provide a practical mean to ensure the coherence of the output summary. Request algorithm with the constraint of cohesion, and compare it to the naive algorithms. On the other side, Jain, et al. [47] used the k means clustering algorithm to produce a coherent and readable summary. While Saggion and Lapalme [15] described a way to summarize the text, which produces indicative-informative summaries of technical papers. And Goldstein [48] discussed the approach of text extraction for multiple documents builds on methods of summarizing a single document by using additional, available information on the document group as a whole, and the relationships between these documents. Other researchers used different techniques for text summarization like the lexical chains as Berker and Gungor [49].

Mitra, et al. [50] tried to assess the domain- independent techniques summarized automatically by extracting paragraphs by comparing these automatic extracts to those generated by humans. On the other hand, Hovy and Lin [17] made the SUMMARIST system as we mentioned in section 3, it provides all of the extracts and summaries of arbitrary text input in English.

A Comparative Study for Text Summarization Techniques

The features that any text summarization technique may use are covered in [Table-1], where each feature presented in a specific column. Those features are: single- document summarization, multi - document summarization, extractive summaries, abstractive summaries, user- focused for query- based summaries, generic summaries, indicative summaries, informative summaries, statistical approach, machine learning approach, linguistic approach.

As explained in section 5, the research works covered in this comparative study are arranged from most recent to oldest ones. In this matrix a classification for each work according to its category is presented.

Concluding Remarks

This paper presented taxonomy for text summarization, covering all its aspects, categories, and approaches implemented in different applications in the last 15 years were presented in a check matrix.

Our goal was to present different use cases of single- document and multi- document, extractive and abstractive, query-based and generic, indicative and informative, statistical approach based, machine learning approach based, or linguistics approach based text summarization.

We have chosen to include a brief discussion on some methods that we found relevant to future research, even if they focus only on small details related to a general summarization process and not on building an entire summarization system.

Based on the check matrix [Table-1], it's clear that nearly half of the works based on single document summarization and the other half based on multi-document summarization as seen in [Fig-2a]. The percentage of single and multi- document text summarization is nearly the same. And should be noted that, most of people who used single document summarization pretend to make their work applicable for multi- document input in the future.

Table 1- Shows the state- of- the art Check Matrix

Research Work	Number of source inputs		Summary construction method		Summary target		Information content		Approach used		
	Single	Multi	Extract	Abstract	User- focused	Generic	Indicative	Informative	Statistical	ML	Linguistic
A Compositional Context Sensitive Multidocument Summarizer [36]		√	√			√		√	√		√
A New Approach to Unsupervised Text Summarization [45]	√		√			√		√			√
A Scalable Global Model for Summarization [28]	√		√			√		√	√		
A Text Summarization Approach under the Influence of Textual Entailment [1]	√		√			√		√			√
Automated Text Summarization in SUMMARIST [17]	√		√	√		√	√	√			√
Automatic Legal Text Summarization [39]	√		√		√			√		√	
Automatic Summarization Focusing on Document Genre [40]		√	√		√		√	√			√
Automatic Text Summarization by Paragraph Extraction [50]	√		√			√		√	√		√
Automatic Text Summarization of Scientific Articles [43]	√		√			√	√				√
Automatic Text Summarization [23]	√		√			√		√		√	
Bottom-up Approach to Sentence Ordering for Multi-document Summarization [37]		√	√			√		√			√
Combining Different Summarization Techniques for Legal Text [11]		√	√			√	√	√	√	√	√
Concept Identification and Presentation in the Context of Technical Text Summarization [16]	√			√		√	√	√			√
Experiments in Multidocument Summarization [44]	√		√			√		√			√
Exploring Content Models [27]		√		√		√	√		√		
Extractive Summarization Using Supervised and Semi-supervised Learning [31]	√		√			√		√		√	
Generating Indicative-Informative Summaries with SumUM [15]	√			√		√	√	√			√
Multi document extraction based Summarization [41]		√	√			√		√			√
Multi Document Extractive Summarization Based On Word Sequences [19]		√	√			√		√		√	
Multi Document Summarization by Combinational Approach [24]		√	√			√		√			√
Multi-Document Summarization by Maximizing Informative Content Words [35]		√	√			√		√		√	√
Multi-Document Summarization By Sentence Extraction [48]		√	√			√	√	√	√		
Multi-Document Summarization Using Cluster-Based Link Analysis [33]		√	√			√		√			√
Multi-Document Summarization via Sentence-Level [32]		√	√		√			√			√
Multiple Alternative Sentence Compressions [13]		√	√	√		√	√	√			√
NEO-CORTEX [14]		√	√		√			√	√		√
Opinion Summarization [26]	√		√			√		√	√		
Optimizing Machine Learning Approach [18]	√		√		√			√		√	
Producing Summaries Tailored to the Citation Context [12]	√		√	√	√		√	√			√
Sentence Ordering in Multidocument Summarization [46]		√									√
Text Summarization Based on Cellular Automata [21]	√		√			√		√		√	
Text Summarization Based on Genetic Programming [22]	√		√			√		√		√	
Text Summarization for Information Retrieval [20]		√	√	√	√		√	√	√		
Text Summarization using Term Weights [38]	√		√			√		√	√		
Topic-based Multi-Document Summarization with Probabilistic [34]		√	√		√			√	√		
Topic-Driven Multi-Document Summarization [30]		√	√		√			√			
Topic-Focused Multi-document Summarization Using an Approximate Oracle Score [38]		√	√		√			√	√		
Towards Automated Related Work Summarization [25]		√	√		√			√			√
Towards Multi-Document Summarization of Scientific Articles [10]		√	√		√		√	√			√
User-Sensitive Text Summarization [42]	√		√		√		√	√			√
Using Domain Knowledge for Text Summarization [29]	√		√		√			√			√
Using Genetic Algorithms With Lexical Chains For Automatic Text Summarization [49]		√	√			√	√	√			√

Regarding the summary construction method, a very high percentage of the works used extractive summary, rare works based on abstractive summary, and a very few number of works used extractive- abstractive summary as presented in [Fig-2b]. Thus, most of researchers prefer extractive summaries more than abstractive. The reason behind it is that abstraction is quite hard, and the most successful systems tested at the Text Analysis Conference (TAC) and Document Understanding Conference (DUC), were extractive.

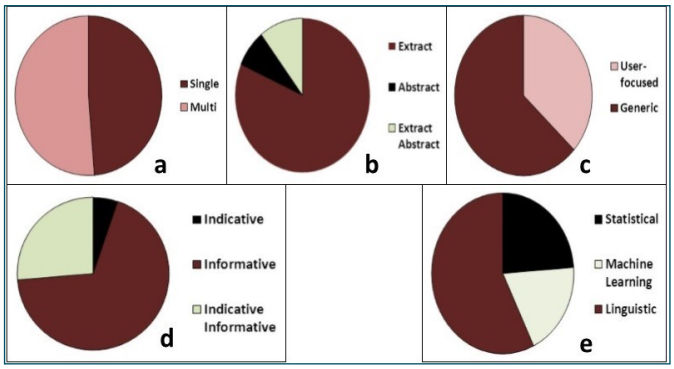

Fig. 2- Shows the analysis of text summarization categories according to [Table-1]

In particular, sentence extraction represents a reasonable trade-off between linguistic quality, guaranteed by longer textual units, and summary content, often improved with shorter units. In addition, extractive summaries are much more accurate than abstraction. Extraction is depending mainly on sentences that already contained in the original input.

Compared to creating an extract, generation of abstract is relatively harder since the latter requires: representation of text units (sentences or paragraphs) semantically in the text, reformulation of two or more text units, and rendering the new representation in natural language.

With respect to summary target, there are a moderate number of works used user- focused technique, and more than half of the works used generic techniques as shown in [Fig-2c]. The summary target methods experienced a change over time, as the user- focused summary target hadn't paid the attention of researchers except recently and it needs more effort to be done.

And for information content, indicative-based summary is rare, a very high percentage of the works is informative- based summary, and few works based on indicative informative summary as clear in [Fig-2d]. As derived from check matrix [Table-1], it can be noticed that informative- based summaries has strong relationship with extraction techniques, so as the usage of extractive techniques is too high, also the informative approach is used a lot too. About indicative information content, it seems clear that it has limited usage to abstraction summarization methods.

As shown in [Fig-2e], more than half of the works used the linguistic approach, few number of works used machine learning approach, and a moderate number of works used the statistical approach. Most of researchers used the linguistics approach because of its accuracy in checking sentence similarity and relativeness; systems depends on this approach are more efficient and reliable. In addition, most of researchers avoided approaches that based on machine learning because of its complications.

Conflicts of Interest: None declared.

References

[1] Lloret E., Ferrández O., Munoz R. & Palomar M. (2008) *NLPCS* 22-31.

[2] Gupta V. & Lehal G.S. (2010) *Journal of Emerging Technologies in Web Intelligence*, 2(3), 258-268.

[3] Israel Q.L., Han H. & Song I.Y. (2010) *Journal of Computing Sciences in Colleges*, 25(5), 10-20.

[4] Damova M. & Koychev I. (2010) *Proc. of Int. Conference S3T'10 Track Intelligent Content and Semantic*, Varna, 11-12.

[5] Das D. & Martins A.F. (2007) *Literature Survey for the Language and Statistics*, II Course at CMU, 4, 192-195.

[6] Ding Y. (2004) *A Survey on Multi-Document Summarization*. Department of Computer and Information Science University of Pennsylvania.

[7] Sekine S. & Nobata C. (2003) *Proceedings of the HLT-NAACL* 03 *on Text Summarization Workshop*, Association for Computational Linguistics, 5, 65-72.

[8] Radev D.R., Hovy E. & McKeown K. (2002) *Computational linguistics*, 28(4), 399-408.

[9] Feldman R., Aumann Y., Finkelstein-Landau M., Hurvitz E., Regev Y. & Yaroshevich A. (2002) *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, 349 -359.

[10] Agarwal N., Reddy R.S., Gvr K. & Rosé C.P. (2011) *Association for Computational Linguistics: Human Language Technologies*, 8.

[11] Galgani F., Compton P. & Hoffmann A. (2012) *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data*, Association for Computational Linguistics, 115-123.

[12] Wan S., Paris C. & Dale R. (2009) *Proceedings of the 9th ACM/ IEEE-CS Joint Conference on Digital Libraries*, 59-68.

[13] Madnani N., Zajic D., Dorr B., Ayan N.F. & Lin J. (2007) Proceedings of Document Understanding Conference.

[14] Boudin F. & Moreno J.M.T. (2007) *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, 551 -562.

[15] Saggion H. & Lapalme G. (2002) *Computational Linguistics*, 28 (4), 497-526.

[16] Saggion H. & Lapalme G. (2000) *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, Association for Computational Linguistics, 1-10.

[17] Hovy E. & Lin C.Y. (1997) *Automated Text Summarization in SUMMARIST*.

[18] Kyoomarsi F., Khosravi H., Eslami E. & Khosravyan P. (2009) *International Journal of Hybrid Information Technology*, 2(2).

[19] Kowsalya R., Priya R. & Nithiya P. (2011) *International Journal of Computer Science Issues*, 8(2).

[20] Negi P.S., Rauthan M.M.S. & Dhami H.S. (2011) *International Journal of Computer Applications*, 21(10), 20-24.

[21] Kiyoumarsi F., Esfahani F.R. & Dehkordi P.K. (2011) *International Conference on Information Communication and Management*, 16.

[22] Dehkordi P.K., Khosravi H. & Kumarci F. (2009) *International*

Journal of Computing and ICT Research, 3(1), 57-64.

[23]Fattah M.A. & Ren F. (2008) *Proceedings of World Academy of Science, Engineering and Technology*, 27, 192-195.

[24]Hariharan S. (2010) *International Journal of Computational Cognition*, 8(4).

[25]Hoang C.D.V. & Kan M.Y. (2010) *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 427-435.

[26]Nishikawa H., Hasegawa T., Matsuo Y. & Kikui G. (2010) *Proceedings of the 23rd International Conference on Computational Linguistics*, Association for Computational Linguistics, 910-918.

[27]Haghighi A. & Vanderwende L. (2009) *Proceedings of Human Language Technologies: Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 362-370.

[28]Gillick D. & Favre B. (2009) *Proceedings of the Workshop on Integer Linear Programming for Natural Langauge Processing,* Association for Computational Linguistics, 10-18.

[29]Sarkar K. (2009) *International Journal of Recent Trends in Engineering*, 1(1).

[30]Nastase V. (2008) *Proceedings of the Conference on Empirical Methods in Natural Language Processing,* Association for Computational Linguistics, 763-772.

[31]Wong K.F., Wu M. & Li W. (2008) *Proceedings of the 22nd International Conference on Computational Linguistics,* Association for Computational Linguistics, 1, 985-992.

[32]Wang D., Li T., Zhu S. & Ding C. (2008) *Proceedings of the 31st annual international ACM SIGIR Conference on Research and Development in Information Retrieval,* 307-314.

[33]Wan X. & Yang J. (2008) *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 299-306.

[34]Hennig L. & Labor D.A.I. (2009) *Recent Advances in Natural Language Processing*.

[35]Yih W.T., Goodman J., Vanderwende L. & Suzuki H. (2007) *International Joint Conference on Artificial Intelligence*, 20.

[36]Nenkova A., Vanderwende L. & McKeown K. (2006) P*roceedings of the 29th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 573-580.

[37]Bollegala D., Okazaki N. & Ishizuka M. (2006) *Proceedings of the 21st International Conference on Computational Linguistics* and 44*th Annual Meeting of the ACL*, 385-392, Sydney.

[38]Conroy J.M., Schlesinger J.D. & O'Leary D.P. (2006) *Proceedings of the COLING/ACL*, Association for Computational Linguistics, 152-159.

[39]Hachey B. & Grover C. (2005) *Tenth International Conference on Artificial Intelligence and Law*, Bologna, Italy.

[40]Seki Y. (2005) *ACM SIGIR Forum,* 39(1), 65-67.

[41]Sripada S., Kasturi V.G. & Parai G.K. (2005) *Multi-document Extraction based Summarization*, CS 224N, Final Project.

[42]Elhadad N. (2004) *Proceedings of the 19th National Conference on Artifical Intelligence,* 987-988.

[43]Jaoua M. & Hamadou A.B. (2003) *Computational Linguistics and Intelligent Text Processing*, Springer Berlin Heidelberg, 623 -634.

[44]Schiffman B., Nenkova A. & McKeown K. (2002) *Proceedings of the Second International Conference on Human Language Technology Research,* 52-58.

[45]Nomoto T. & Matsumoto Y. (2001) *Proceedings of the 24th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, 26-34.

[46]Barzilay R., Elhadad N. & McKeown K.R. (2001) *Proceedings of the First International Conference on Human Language Technology Research*, Association for Computational Linguistics, 1-7.

[47]Jain H.J., Bewoor M.S. & Patil S.H. (2012) *International Journal of Soft Computing and Engineering*, 2(2), 301-304.

[48]Goldstein J., Mittal V., Carbonell J. & Kantrowitz M. (2000) *Proceedings of the NAACL-ANLP Workshop on Automatic Summarization*, Association for Computational Linguistics, 4, 40-48.

[49]Berker M. (2011) *Using Genetic Algorithms with Lexical Chains for Automatic Text Summarization,* Doctoral dissertation, Bogaziçi University.

[50]Mitra M., Singhalz A. & Buckleyyy C. (1997) *Compare*, 22215, 26.

[51]Balabantaray R.C., Sahoo D.K., Sahoo B. & Swain M. (2012) *International Journal of Computer Applications*, 38(1), 10-14.