A Comprehensive Survey on Text Summarization Systems

Saeedeh Gholamrezazadeh
Islamic Azad University of Mashhad
saeedeh.motlagh@gmail.com

Mohsen Amini Salehi
Islamic Azad University of Mashhad
amini@mshdiau.ac.ir

Bahareh Gholamzadeh
Sadjad Institute of Higher Education
bahareh_gholamzadeh@sadjad.ac.ir

Abstract

Text summarization systems are among the most attractive research areas nowadays. Summarization systems offers the possibility of finding the main points of texts and so the user will spend less time on reading the whole document. Different types of summary might be useful in various applications and summarization systems can be categorized based on these types. This paper presents a taxonomy of summarization systems and defines the most important criteria for a summary which can be generated by a system. Additionally, different methods of text summarization as well as main steps for summarization process is discussed. we also go through main criteria for evaluating a text summarization.

1. Introduction

The world of texts is a vast and widespread world. In most of data networks, like internet, the great deals of important information are still in the text format. Nowadays, with the massive increase in the text information that we receive every day, text summarization system could be helpful in finding the most important contents of the text in a short time. Text summarization system can be used in different situations; for instance, as a summarizer in a search engine to give a summarized information of each page to a user [1], and for summarizing the letters and other document in offices. Moreover, newsgroups can use multi-documents summarization system to merge the most important information of documents which are discussing one topic. Summarization systems are also popular in areas that we want to decrease the amount of transferred information. For example, users who check their emails by cell phones prefer to have less transferred data while connecting to the internet. Achieving this goal, web sites may use these systems to decrease the amount of transferred data which results in to access to the information more quickly.

In the following sections, at first, popular classifications for summarization systems will be discussed and then we will go through the main steps of text summarization process. Moreover, different approaches for sentence selection are presented in order to generate a summary from a text and finally, well-known criteria for evaluating text summarization systems are discussed.

2. Text summarization definition

What is the key information in a document? Different users have different attitudes toward the significance of information in a document and these attitudes are generally based on their needs. Therefore, a text might be summarized in many different ways.

First of all, we should define "what does summary mean?" There are various definitions for summary. *Edward Hovy et.al* [2] defines the summary as a text which is based on one or more texts; it has the most important information of the main texts and its content is less than half of the main texts. Mani [3], describes the text summarization as a process of finding the main source of information, finding the main important contents and presenting them as a concise text in the predefined template.

3. Different Types of Summary

There is not a unit classification for summarized texts and summaries could be categorized based on different criteria. Different papers discussed some of these criteria. However, in this paper we tried to cover the most comprehensive criteria for classifying summaries. These categories which are presented in figure 1, are as following:

3.1. Approaches

Based on the different approaches of analyzing the texts and generation of the summary, text summarization systems are divided to *extract* and

978-1-4244-4946-0/09/$25.00 ©2009 IEEE

abstract systems. The *extract* summary is formed by reusing the portions of the main text like words and sentences. In this type of summary The most important information of the main text which is usually the first sentence of each paragraph, special names, italic or bold phrases are copied to the final summary. Unfortunately, extracts suffer from inconsistencies, lack of balance, and lack of cohesion [4]. One example of a system which use extract summary is Summ-It applet [1] which is designed by Surrey University.

In an *abstract* summary, the summarized text is an interpretation of an original text. The process of producing involves rewriting the original text in a shorter version by replacing wordy concept with shorter ones [5]. At first, the system analyses the main text and then it presents its comprehension from the text in a human understandable form. For example SUMMARIST [2] includes modules to perform topic interpretation and summary generation which enable it to produce abstract summaries.

3.2. Details

This category is based on the details which are important in the desired summary. Two different groups could be considered: *indicative* and *informative*.

Indicative summarization systems only present the main idea of the text to the user. The typical length of this type of summarization is 5 to 10 percent of the main text. *Indicative* summaries might be used for encouraging the readers to read the main documents [4]. The information on the back of the movie packs is an example of this kind of summarization.

On the other hand, the *informative* summarization systems give concise information of the main text and it can be considered as a substitution for the main document. The length of informative summary is 20 to 30 percent of the main text [5].

One example for detailed based summarization system is SumUM [6], a text summarization system that takes a raw technical text as input and produces an indicative informative summary.

3.3. Content

Another classification, which is based on the importance of the content in the original text, is *generic* versus *query-based* summaries.

In *generic* summarization, the system does not depend on the subject of the document. In such systems, the user does not have any previous understanding of the text and it is assumed that the summary may be used by different types of users, so all the information are in the same level of importance [4].

On the other hand, in *query-based* summarization, the user has to determine the topic of original text in form of a query, before the summarization process. In these types, the user has general information about the text and searches for specific information, which is usually an answer to a question. So, user asks that special information in form of a query and the system only extract that information from the text and presents it as a summary [4].

Mitre's WebSumm [7] is a type of *Query-based* summarizer which performs sentence extraction over single or multiple documents in conjunction with a search engine. The resulting summary is an extract of sentences based on a users query.

An example of generic summarizer system is SUMMARIST [2] which produces summaries of web documents.

3.4. Limitation

Another classification is based on limitation on input text. Three different summary can be defined, *domain dependant, genre specific*, and *independent*.

Genre specific systems only accept special type of texts as their input and there is a limitation on the template of the text. There are different templates like newspaper articles, scientific papers, stories, manuals and etc. The system uses the structure of these templates for generating the summary [4]. On the other hand, the independent systems do not have any predefined limitation and can accept different types of texts. Moreover, there are some systems that only summarize the texts which their subject can be defined in the domain of the system; these systems are called *domain dependant*. These kinds of systems exert some limitations on the subject of documents. Such systems know everything about a special domain and use this information for summarization [4]. There are several systems which apply these limitations in their summarization process. Some examples of these systems are as follows:

The Copy and Paste system is a type of *domain independent* summarizer. It is designed to take the results of a sentence extraction summarizer and extract key concepts from these sentences. These concepts are then combined to form new sentences [1].

A research group in Sheffield University designs a summarizer system “TRESTLE” [1] which produces summaries in the news domain.

There is also a system for *genre-specific* summarization of documents which overcomes the problem of summarizing heterogeneous document collections by taking the genre, or type, of document into account when selecting summary sentences [7].

3.5. Number of input documents

A summarization system can accept one or more documents as input. Some systems are *single-document* and only accept one document as input while others, which get several documents as input, termed *multi-document*. These documents must have topic relation with each other and the summarizer generates a summary based on them [4].

An example of multi-document summary system is SUMMONS which is designed in Columbia University [1]. On the other hand, Copy and Paste system is a single document summarizer which has mentioned earlier [1].

3.6. Language

Based on the language of the input text and the generated summary we can categorize summarization systems in two main groups, *mono lingual* and *multi lingual*. *Mono lingual* systems only accept documents with specific language, like English, and the generated summary is based on that language too. On the contrary, *multi lingual* systems can accept documents in different languages and the user can choose the language of the output summary [4].

For instance, FarsiSum [8] is a type of *mono lingual* text summarization systems which only summarizes Farsi texts. However, SUMMARIST [2] is a *multi lingual* summarizer which is available for English, Japanese, Spanish, Arabic, Indonesian, and Korean.

4. Main steps for text summarization

Lin, and Hovy presented three main steps for summarizing documents [9]. These steps are *topic identification, interpretation* and *summary generation* .

4.1. Topic Identification

In topic identification step, the most prominent information in the text is identified. Usually the system assign different precedence to different parts of the text like sentence, words, and phrases; then a fuser module mix the scores of each part in order to find the total score for a part. At last, the system presents the N highest score parts in accordance with predefined length [2].

Several techniques for topic identification, including methods based on Position, Cue Phrases, word frequency, and Discourse Segmentation have been reported in the literature [2].

Methods which are based on the position of phrases are the most useful methods for topic identification. The orderly structure of the texts results in to extraction of the main information based on their position. Because there are lots of differences in various texts with different domains and structures, the position based methods should be defined in accordance with the template as well as text dominant.

Phrases like, “the most”, “in conclusion”, “the best”, “in summary”, and etc are good indicators which implies the most important parts of the texts. Moreover, the most frequent words can be considered as the most important part of the text unless they are functional words such as determiner, and prepositions.

4.2. Interpretation

Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content [10]. Fusing topics into one or more characterizing concepts is the most difficult step of automated text summarization [2]. Nowadays, all these systems should have symbolic word knowledge for doing the interpretation, but gathering enough information in a system is very difficult and more work should be done in this area.

4.3. Summary Generation

In this step, the system uses text generation method which itself is still an open research topic that has lots of similarities with text summarization. This step includes a range of various generation methods from very simple word or phrase printing to more sophisticated phrase merging and sentence generation [10]. In text generation, the computer generates the natural language from the processed information of the previous steps.

5. Text summarization Methods

In order to generate a high quality summary different natural language processing techniques like language analysis, information inference, and etc must be used. Nowadays, most of summarization systems produce summary based on key sentence selection. There are three different approaches for scoring and selecting sentences which are discussed below and presented in figure 2.

5.1. Statistical approaches

In Statistical methods, sentence selection is done based on word frequency, indicator phrases and other features regardless of the meaning of the words [11]. These methods are based on the idea that text surface cues are the most obvious indication of the text

contents. There are several methods for determining the key sentences such as, The Title Method [12], The Location Method [13], The Aggregation Similarity Method [5], The Frequency Method [14], TF- Based Query Method [12], and Latent Semantic Analysis [15, 14]. Below are some examples of different systems which use these methods for generating a summary.

- IN XIGHT's [1] Summary Server is an application that creates extraction-based summaries offline. It uses statistical extraction techniques based on features such as sentence position, sentence length and keywords.
- The Text Summarization project which proposed in Ottawa University planned to use surface level statistics such as frequency analysis and surface level linguistic features such as sentence position [1].
- University of southern California developed the SUMMARIST [1] system which produces summaries of web documents. It first identifies the main topics of the document using statistical techniques based on features such as position, and word counts. Current research is underway to use cue phrases and discourse structure.

5.2. Linguistic approaches

Linguistic approaches are based on considering the connections between words and trying to find the main concept by analyzing the words. Different techniques used in this approach are discussed here.

5.2.1. Lexical chain. Lexical chain produces a presentation of text contiguous structures. Basically, lexical chains exploit the cohesion among an arbitrary number of related words. Lexical chains can be computed in a source document by grouping (chaining) sets of words which are semantically related. Identities, synonyms, and hyponyms are the possible relations among words that might make them to be grouped into the same lexical chain [16]. Lexical chain is used for information retrieval and grammatical error corrections. In [17] the authors has proposed a new technique to produce a summary of an original text without requiring its full semantic interpretation, but instead relying on a model of the topic progression in the text derived from lexical chains.

5.2.2. WordNet. WordNet [18] is a thesaurus that is used for determining relationship between words. Semantic relations between words in the WordNet database are represented relationally by synonym sets, hyponym/hyponym, metonym trees, and etc. WordNet is used for building lexical chain according to these relations.

One example of a summarization system which uses WordNet for generating the lexical chain is LexSum.

5.2.3. Graph theory. Graph can be applied for representing the structure of the text as well as the relationship between sentences of the document. Sentences in documents are presented as nodes. The Edges between nodes illustrates connections between sentences. These connections are introduced by similarity relation and this relation is measured as a function of likelihood between contents. By deploying different similarity criteria, the similarity between two sentences is calculated and each sentence is scored. All the scores for one sentence are combined to form a final score for each sentence. When the graph processed, the sentence will be categorized by their scores and sentences in higher orders are chosen for final summary [4].

In [19] the authors introduced a stochastic graph-based method for computing relative importance of textual units for natural language processing. They also test this technique on the problem of text summarization. In this method, a connectivity matrix based on intra-sentences cosine similarity is used as the adjacency matrix of the graph representation of sentences.

5.2.4. Clustering. Clustering is used to decrease the information by categorizing and grouping the similar data. There are two major types of clustering; hierarchy clustering and partitioning.
In hierarchy clustering smaller clusters are mixed with each other to form bigger clusters or sometimes the clusters are divided by half.
On the other hand, partitioning tries to decompose the data collection to distinct clusters.

MultiGen [1] is a multi-document system in the news domain. It extracts sentence fragments that represent key pieces of information in the set of related documents.

Linguistic approaches need great amount of memory for saving supplementary linguistic information like WordNet. Moreover, complex linguistic processing needs powerful processors [12].

5.3. Rhetorical approaches

Rhetorical structure theory (RST) is based on the Rhetorical connections between different parts of the text. In this theory the Rhetoric behind the decomposed text is extracted. In summarization systems, Rhetorical structure (RS) presents the logical connections between different parts of the text and interprets these

connections. These information represent the discourse structure and features of the main document [20].

After identifying text units and rhetorical connections between them, the RS tree is formed based on these information. In [20] the formulation of forming the RS tree has been described.

6. Evaluating the summarization systems

Evaluation methods are useful in evaluating the usefulness and trustfulness of the summary. In summary, evaluating the qualities like comprehensibility, coherence, and readability is really difficult. System evaluation might be performed manually by experts who compare different summaries and choose the best one. A problem with this approach is that the individuals who perform the evaluation task normally have very different ideas on what a good summary should contain. In a test, Hassel (2003) found that at best there was a 70% agreement between summaries created by two individuals [10]. A further problem with manually performed evaluation is that it is an extremely time consuming task [10]. Automatic system evaluation is another way for evaluating summarization systems [21] which is still an open research topic. Since there is not a base standard for evaluating systems, different criteria are being used for evaluation. In the following paragraph two major and most practical methods are discussed.

Two main criterion for evaluating the proficiency of a system is *precision* and *recall* which are used for specifying the similarity between the summary which is generated by a system versus the one generated by human. These terms are defined by following equations [22]:

$$Precision = Correct/(Correct + Wrong) \quad (1)$$

$$Recall = Correct/(Correct + Missed) \quad (2)$$

where, *Correct* is the number of sentences that are the same in both summary which are produced by human and system.; *Wrong* is the number of sentences presented in summary and produced by system but is not included in human generated summary; *Missed* is the number of sentences which are not appeared in system generated summary but presented in the summary produced by human. Therefore, *Precision* specifies the number of suitable sentences which are extracted by system and *Recall* specify the number of suitable sentences that the summarization system missed.

There are also two other criteria for evaluating system which are *compression ratio* and *retention ratio* And defined as follows [23]:

$$Compression\ Ratio: CR = \frac{Length\ S}{Length\ T} \quad (3)$$

$$Retention\ Ratio: RR = \frac{Information\ in\ S}{Information\ in\ T} \quad (4)$$

Where S is the summarized text and T is the main text. So we can conclude that a good summary is the one with low *CR* and high *PR* [24].

7. Conclusion

Currently, text summarization is one of the hot areas of research and attracts lots of attentions from different fields. Text summarization systems can be categorized in to various groups based on different approaches were presented in this paper. As discussed earlier, there are three main steps for producing a summary from an input text (topic identification, interpretation, and summary generation). Most of summarization systems follow these steps in order to generate a summary.

In this paper, different types of summarization methods, which might be used in a system for generating a summary, were also presented. We also discussed the most important issues in evaluating a summary and present common criterion for evaluating a summarization system.

8. References

[1] http://web.science.mq.edu.au/~swan/summarization/projects_full.htm

[2] Eduard Hovy and Chin Yew Lin, *Automated text summarization in SUMMARIST*, MIT Press, 1999, pages 81–94.

[3] Mani, I., *Automatic Summarization,* 2001, John Benjamin's Publishing Co. pp.1-22.

[4] Karel Jezek, and Josef Steinberger, *Automatic Text Summarization (the state of the art 2007 and new challenges),* Znalosti 2008, pp. 1-12.

[5] Waleed al-sanie, "Towards an infrastructure for Arabic text summarization using rhetorical structure theory", Master Thesis, Department of computer science. King Saud university, Riyadh, Kingdom of Saudi Arabia, 2005.

[6] S. Horacio, L. Guy, "Generating indicative-informative summaries with SumUM : Summarization", *Computational linguistics - Association for Computational Linguistics,* 2002, vol. 28, pp. 497-526.

[7] Kupiec, Julian M, Schuetze, Hinrich, "System for genre-specific summarization of documents", Xerox Corporation, 2004.

[8] Martin Hassel, Nima Mazdak, "A Persian text summarizer", *International Conference on Computational Linguistics*, 2004.
[9] Lin, C.Y.and Hovy, "Identify Topic by Position", in *Proc. 5th Conference on Applied Natural Language Processing*, March. 1997.
[10] Nima Mazdak, "A Persian text summarizer", Master Thesis, Department of linguistics, Stockholm university., January 2004.
[11] Elhadad, M. , "Using Argumentation to Control Lexical Choice: A Functional Unification-Based Approach", Ph.D. dissertation, Columbia University. 1992.
[12] Youngkoong Ko, Jungyun Seo, "An Effective Sentence-Extraction Technique Using Contextual Information and Statistical Approaches for Text Summarization", Pattern Recognition Letters. doi:10.1016/j.patrec. 2008.02.008.
[13] Wasson, M., "Using leading text for news summaries: Evaluation results and implications for commercial summarization applications", in *Proc. 17th International Conference on Computational Linguistics and 36th Annual Meeting of the ACL,* 1998, pp.1364-1368.
[14] Salton, G., *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer,* Addison-Wesley Publishing Company, 1989.
[15] Bellegarda, J., "Exploiting latent semantic information in statistical language modeling," in *Proc. IEEE,* August 2000.Vol. 88, No. 8,pp: 1279-1296.
[16] Silber G.H., Kathleen F. McCoy, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," *Computational Linguistics 28(4)*: 487-496, 2002.
[17] Barzilay, R., Elhadad, M, "Using Lexical Chains for Text Summarization," in *Proc. ACL/EACL'97 Workshop on Intelligent Scalable Text summarization,* Madrid, Spain,1997, pp. 10–17.
[18] William P. Doran, Nicola Stokes, John Dunnion, and Joe Carthy. "Comparing lexical chain-based summarisation approaches using an extrinsic evaluation," In *Proc. Global Wordnet Conference (GWC 2004)*, 2004.
[19] Gunes Erkan, Dragomir R. Radev, "Efficiently Computed Lexical Chains as an Intermediate Representation for Automatic Text Summarization," *Journal of Artificial Intelligence Research 22* (2004) 457-479
[20] Mann, W.C., A. Thompson, and S. , *Rhetorical Structure Theory: Toward a functional theory of text organization*, 1998, pp. 243-281.
[21] Dalianis, H., Hassel, M., de Smedt, K., Liseth, A., Lech, T.C. and Wedekind, J. "Porting and evaluation of automatic summarization," *In Holmboe, H. (ed.) Nordisk. 2003.*
[22] Eduard Hovy, *The Oxford Handbook of Computational Linguistics*, Oxford University Press, Oxford, 2003, chapter 32.
[23] E. Hovy and Chin-Yew Lin, "Automated Text Summarization and the SUMMARIST system," TIPSTER Text Program Phase III final report, October 1998.
[24] Grishman, R.Hobbs, J.Hovy, E. Sanfilippo, A. Wilks, Y., "Cross-lingual Information Extraction and Automated Text Summarization," (ed. Hovy). Report to US NSF & EU Commission, April 1999

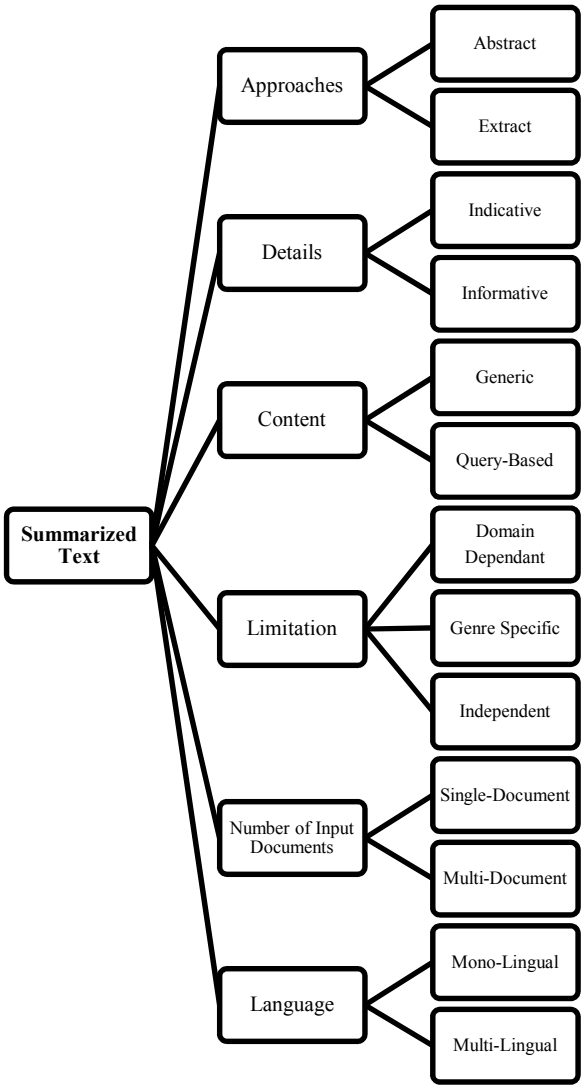

Figure 1. Type of summary

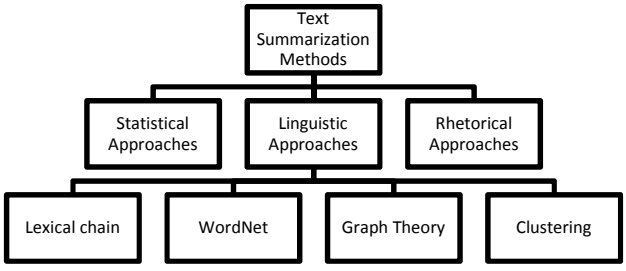

Figure 2. Text summarization methods