Comparative Study of Text Summarization Methods

Nikita Munot
Department of Computer Engineering
PIIT, New Panvel, India

Sharvari S. Govilkar
Department of Computer Engineering
PIIT, New Panvel, India

ABSTRACT

Text summarization is one of application of natural language processing and is becoming more popular for information condensation. Text summarization is a process of reducing the size of original document and producing a summary by retaining important information of original document. This paper gives comparative study of various text summarization methods based on different types of application. The paper discusses in detail two main categories of text summarization methods these are extractive and abstractive summarization methods. The paper also presents taxonomy of summarization systems and statistical and linguistic approaches for summarization.

Keywords

NLP, text summarization, abstractive summary, semantic graph theory, linguistic approach, statistical approach

1. INTRODUCTION

Natural language processing (NLP) is a field of computer science, artificial intelligence and linguistics concerned with the interactions between computers and human language. Natural language processing is a process of developing a system that can process and produce language as good as human can produce. The use of World Wide Web has increased and so the problem of information overload also has increased. Hence there is a need of a system that automatically retrieves, categorize and summarize the document as per users need. Document summarization is one possible solution to this problem.

Text summarization is a process to express the content of a document in a condensed form that meets the needs of the user. More and more electronic data is available on the Internet and it is not possible to read everything and hence some form of information condensation is needed. Summarization serves as a tool which helps the user to efficiently find useful information from immense amount of information.

Text summarization can be used by various applications; for instance researchers need a tool to generate summaries for deciding whether to read the entire document or not and for summarizing information searched by user on Internet. News groups can use multi document summarization to cluster the information from different media and summarize.

The paper presents a detail survey of various summarization techniques and advantages and limitation of each method. Text summarization is defined in section 2. Related work done and past literature is discussed in section 3. Text summarization methods based on statistical and linguistic approaches are discussed in detail in section 4 along with the comparison of each method. Finally, section 5 concludes the paper.

2. TEXT SUMMARIZATION

Text summarization a process of reducing the size of the original document while preserving its information content and its summary is less than half of the main text. Summarization has been viewed as a two step process. The first step is the extraction of important concepts from the source text by building an intermediate representation of some sort. The second step uses this intermediate representation to generate a summary. News blaster is a good example of a text summarizer, that helps users find the news that is of most interest to them. The system automatically collects, cluster, categorizes, and summarizes news from several sites on the web on a daily basis. A summarization machine can be viewed as a system which accepts either a single document or multiple documents or a query as an input and produces a abstract or extract summary.

3. LI TERATURE SURVEY

Past literature that use the various summarization techniques are cited in this section. Most of the researchers concentrate on sentence extraction rather than generation for text summarization. The most widely used method for summarization is based on statistical features of the sentence which produce extractive summaries.

Luhn[4] proposed that the most frequent words represent the most important concept of the text. His idea was to give the score to each sentence based on number of occurrences of the words and then choose the sentence which is having the highest score. Edmunson[16] proposed methods based on location, title and cue words. He stated that initial few sentences of a document or first paragraph contains the topic information and that should be included in summary. One of the limitation of statistical approach is they do not consider semantic relationship among sentences. Goldstein [2] proposed a query-based summarization to generate a summary by extracting relevant sentences from a document based on the query fired. The criterion for extraction is given as a query. The probability of being included in a summary increases according to the number of words co occurred in the query and a sentence. Goldstein[2][1] also studied news article summarization and used statistical and linguistic features to rank sentences in the document.

One of the approach for summarization can be done by sentence extraction and clustering. ZHANG Pei-ying & LI Cun[5] suggested that sentences are clustered based on the semantic distance among sentences and then calculates the accumulative sentence similarity between the clusters and finally chooses the sentences based on extraction rules. The method used to cluster the sentences is k-means algorithm[5].

The concept of lexical chain was first introduced by Morris and Hirst[9][7]. Lexical chains [7] exploit the cohesion among an arbitrary number of related words. Lexical chains are created by grouping set of words that are semantically related. Barzilay and Elhadad[8][6] constructed lexical chain by calculating semantic distance between words using WordNet. Strong lexical chains are selected and the sentences related to these strong chains are chosen as a summary.

H. Gregory Silber and McCoy [10] developed a liner time algorithm for lexical chain computation. The author follows Barzilay and Elhadad [6] for employing the lexical chains to extract important concepts from the source text by building an intermediate representation. The paper [10] discusses an algorithm for creating lexical chain which creates an array of Meta-Chain whose size is the number of nouns senses in the Word Net and in the document. There were some problems with the algorithm like proper nouns and anaphora resolution that were to be addressed.

There is another method for summarization by using graph theory [11]. The author proposed a method based on subject-object-predicate (SOP) triples from individual sentences to create a semantic graph of the original document. The relevant concepts, carrying the meaning, are scattered across clauses. The author [11] suggested that identifying and exploiting links among them could be useful for extracting relevant text.

One of the researchers, Pushpak Bhattacharyya [12] from IIT Bombay introduced a Word Net based approach for summarization. The document is summarized by generating a sub-graph from Word-net. Weights are assigned to nodes of the sub-graph with respect to the synsnet using the Word Net. The most common text summarization techniques use either statistical approach or linguistic approach or a combination of both.

4. TYPES OF SUMMARIZATION TECHNIQUES

Different types of summary might be useful in various applications and summarization systems can be categorized based on these types. In addition to abstract and extract, there are various types of summaries. A full understanding of the major dimensions of variation, and the types of reasoning required to produce each of them, is still a matter of investigation. This makes the study of automated text summarization an exciting area in which to work. Various summarization methods can be compared based on the type of summary and application. Summarization system can be classified into the following categories, they are:

1. Based on approaches

There are two strategies for summarization those are summarization by extraction, which consists of extracting source sentences as it is and adding into a summary, and summarization by abstraction, which involves generating novel sentences for the summary[1]. The need for abstraction is especially high when opinions are diverse.

Summarization by extractive just extracts the sentences from the original document and adds them to summary. Extractive method is usually easy to implement and is based on statistical features not on semantic relation with sentences. Therefore the summary generated by this method tends to be inconsistent.

Summarization by abstraction needs understanding of the original text and then generating the summary which is semantically related. It provides more generalized summary but it is difficult to compute.

2. Based on type of details

Based on type of detail summary can be either informative or indicative[1]. An indicative summary is used for quick view of a lengthy document and it provides only the main idea of the original text. These are usually small and it encourages a user to read the original document. For example while purchasing any novel a buyer reads the summary provided at back side of novel.

Informative summary serves as a substitution to the original document. It provides the concise information about the original document to the user.

3. Based on type of content

This classification is based on the type of content in the original document[1]. Generic summarization is system which can be used by any type of the user and summary does not depend on the subject of the document. All the information is at same level of importance and which is not user specific.

Query-based summarization [1] is question answer type where the summary is the result of query. It provides the users view and cannot be used by any type of user.

4. Based on limitation

Summary can be classified based on limitation of input text[1]. Genre specific systems only accept special type of input like newspaper articles, stories, manuals etc. Limited to the type of input they can accept.

Domain independent system can accept different type of text. They are not dependent on the domain and can be used by any type of user. There are few systems that are domain dependent.

5. Based on number of input documents

Summarization can be classified based on whether a system accepts one or more documents as input[1]. Single document summarization can accept only one document as input. They are usually easier to produce as it involves summarization of a single document.

Multi-document summarization accepts several documents of same topic as an input. It is more difficult to implement as there are multiple documents to summarize.

6. Based on language

Mono lingual system only accepts documents with specific language and output is based on that language only. Multi-lingual systems can accept documents in different languages and produce summary of different languages.

Following tables presents a comparison of all summarization methods based on type of summary.

Table 1. Comparison of summarization methods

Type of summarization methods	**Subtype**	**Concept**	**Advantages**	**Disadvantages**	**Application/Work Done**
1.Approaches Figures	Abstractive	It is the process of reducing a text document in order to create a summary that is semantically related	Good compression ratio. More reduced text and semantically related summary	Difficult to compute	SUMMRIST [14]

	Extractive	It consists of selecting important sentences from original document based on statistical features	Easy to compute because it does not deal with the semantics and more successful	Suffers from inconsistencies, lack of balance, results in lengthy summary	Summ-It applet,designed by Surrey University [15]
2.Details	Indicative	It only presents main idea of text to user. They can be used to quickly decide whether a text is worth reading	Encourages the users to read the main document in depth. Used for quick categorization and easier to produce	Detailed information is not present	Information present on the back of the movie pack or novels Length 5 to 10%
	Informative	Gives concise information of the main text	Serves as a substitution for the main document	Does not provide quick overview	SumUM [3] Length 20 to 30%
3.Content	Generic	Generalized summary irrespective of the type of user. Information is at same level of importance	Can be used by any type of user	It provides an author's view not user specific	SUMMARIST [8]
	Query based	User has to determine the topic of original text in the form of query and system only extract that information	Specific information can be searched. It reflects user’s interest	Not used by any type of user. It is based on type of user	Mitre's WebSumm [17]
4.Limitation	Domain dependent	Summarize the text which their subject can be defined in the fixed domain	They are aware of the special domain on which they are dependent	Limited to the subject of the document	TRESTLE [15]
	Genre specific	Accept only special type of text as input.	Overcomes the problem of summarizing heterogeneous document	Limitation template of the text	Newsblaster
	Domain Independent	Can accept any type of text.	Any type of text input is accepted. It is not domain dependent	Difficult to implement	Copy and Paste system [15]
5.Number of input document	Single document	Can accept only one input document	Less overhead	Cannot summarize multiple documents of related topics	Copy and paste system [15]
	Multi-document	Can accept multiple input documents	Multiple documents of same topic can be summarized to single document	Difficult to implement	SUMMONS Designed by Columbia university [15]
6.Language	Mono-Lingual	Can accept input only with specific language and output is based on that language	Need to work with only one language	Cannot handle different language	FarsiSum [18]
	Multi-Lingual	Can accept documents in different language	Can deal with multiple language	Difficult to implement.	SUMMARIST(English, Japanese,Spanish) [14]

Text summarization methods can be classified mainly into categories these are extractive and abstractive. Text summarization by extraction simply is extracting few sentences from the original document as it adding it to the summary.

4.1 Extractive Summarization

Extractive text summarization works by selecting a subset of existing words, phrases or sentences from the original text to form summary. Extractive summarization uses statistical approach for selecting important sentences or keyword from document. Extractive summarization uses statistical approach for selecting the important sentences or keyword from document. Various statistical methods are discussed in the below section. Extracted sentences tend to be longer than average. Conflicting information may not be presented accurately.

4.2 Abstractive Summarization

Abstractive text summarization method generates a sentence from a semantic representation and then use natural language generation techniques to create a summary that is closer to what a human might generate. Such a summary might contain words not explicitly present in the original. It consists of understanding the original text and re-telling it in fewer words. It uses linguistic approach to understand the original text and then generates summary.

Abstractive summaries are more accurate as compared to the extractive summary but are difficult to generate because it needs deep understanding of the NLP tasks. Abstractive and extractive summarization uses either statistical or linguistics approaches or combination of both to generate summary.

4.3 Statistical Approaches

Statistical approaches [1] can summarize a document using statistical features of the sentence like title, location, term frequency, assigning weights to the keywords and then calculating the score of the sentence and selecting the highest scored sentence into the summary. Importance of a sentence can be decided by several methods such as:

4.3.1 Title method[16]

This method [16] [4] states that sentences that appear in the title are considered to be more important and are more likely to be included in the summary. The score of the sentences is calculated as how many words are commonly used between a sentence and a title. Title method cannot be effective if the document does not include any title information.

4.3.2 Location method[16]

Weights are assigned to text based on location whether it appears in lead, medial or final position in a paragraph or in appears in the prominent section of the document such as conclusion or introduction. Leading several sentences of a document or last few sentences or conclusion are considered to be more important and included in summary. Hovy & Lin [14] and Edmundson [16] used this method. The location method relies on the following intuition headings, sentences in the beginning and end of the text, text formatted in bold, contain important information to the summary.

4.3.3 tf-idf method[7]

The term frequency-inverse document frequency is a numerical statistic which reflects how important a word is to a document. It is often used as a weighting factor in information retrieval and text mining. tf-idf is used majorly for stop words filtering in text summarization and categorization application. The tf-idf value increases proportionally to the number of times a word appears in the document. tf–idf weighting scheme are often used by search engines as a central tool in scoring and ranking a document's relevance given a user query.

The term frequency f(t,d) means the raw frequency of a term in a document, that i the number of times that term t occurs in document d. The inverse document frequency is a measure of whether the term is common or rare across all documents. It is obtained by dividing the total number of documents by the number of documents containing the term.

4.3.4 Cue word method[16]

Weight is assigned to text based on its significance like positive weights "verified, significant, best, this paper" and negative weights like "hardly, impossible". Cue phrases are usually genre dependent. The sentence consisting such cue phrases can be included in summary. The cue phrase method is based on the assumption that such phrases provide a "rhetorical" context for identifying important sentences. The source abstraction in this case is a set of cue phrases and the sentences that contain them. Above all statistical features are used by extractive text summarization.

4.4 Linguistic Approaches

Linguistic is a scientific study of language which includes study of semantics and pragmatics. Study of semantics means how meaning is inferred from words and concepts and study of pragmatics includes how meaning is inferred from context. Linguistic approaches are based on considering the connection between the words and trying to find the main concept by analyzing the words. Abstractive text summarization is based on linguistic method which involves the semantic processing for summarization.

Linguistic approaches have some difficulties in using high quality linguistic analysis tools (a discourse parser, etc.) and linguistic resources (Word Net, Lexical Chain, Context Vector Space, etc.). Barzilay and Elhadad[6], Miller et al proposed and developed strong concepts with the help of linguistic features but they require much memory for saving the linguistic information like Word Net and processor capacity because of additional linguistic knowledge and complex linguistic processing.

4.4.1 Lexical chain[6][10]

The concept of lexical chains was first introduced by Morris and Hirst[9]. 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 that are semantically related. Identities, synonyms, and hypernyms/hyponyms are the relations among words that might cause them to be grouped into the same lexical chain. Lexical chains are used for IR and grammatical error corrections [6] [10]. In computing lexical chains, the noun instances must be grouped according to the above relations, but each noun instance must belong to exactly one lexical chain. There are several difficulties in determining which lexical chain a particular word instance should join. Words must be grouped such that it creates a strongest and longest lexical chain.

4.4.2 Word Net[19]

Word Net is a on-line lexical database available for English language. It groups the English words into sets of synonyms called sys-nets. Word Net also provides a short meaning of each sys-net and semantic relation between each sys-net. Word-net also serves as a thesaurus and a on-line dictionary which is used by many systems for determining relationship between words. Thesaurus is reference work that contains a list of words grouped together according to the similarity of meaning. Semantic relations between the words are represented by synonyms sets, hyponym trees. Word-net are used for building lexical chains according to these relations. Word Net contains more than 118,000 different word forms[19]. LexSum is a summarization system which uses Word Net for generating the lexical chain.

4.4.3 Graph theory[11]

Graph theory [11] can be applied for representing the structure of the text as well as the relationship between sentences of the document. Sentences in the document are represented as nodes. The edges between nodes are considered as connections between sentences. These connections are related by similarity relation. By developing different similarity criteria, the similarity between two sentences is calculated and each sentence is scored. Whenever a summary is to be processed all the sentences with the highest scored are chosen for the summary. In graph ranking algorithms, the importance of a vertex within the graph is iteratively computed from the entire graph.

TextRank algorithm is a graph based algorithm which is applies in summarization. A graph is constructed by adding a vertex for each sentence in the text. Edges between vertices's are established using sentence inter-connections.

These connections are defined using a similarity relation, where similarity is measured as a function of content overlap. The overlap of two sentences can be determined as the number of common tokens between lexical representations of two sentences. The iterative part of algorithm is consequently applied on the graph of sentences. When its processing is finished, vertices's (sentences) are sorted by their scores. The top ranked sentences are included in the result.

Extracting summary by semantic graph generation[11] is a method which uses subject–object–predicate (SOP) triples from individual sentences to create a semantic graph of the original document. Using the Support Vector Machines learning algorithm, it trains a classifier to identify SOP triples from the document semantic graph that belong to the summary. Usually main functional elements of sentences and clauses are Subjects, Objects, and Predicates, thus identifying and exploiting links among them could facilitate the extraction of relevant text. A method that creates a semantic graph of a document, based on logical form triples subject– predicate–object (SPO), and learns a relevant sub-graph that could be used for creating summaries.

4.4.4 Clustering[5]

Clustering is used to summarize a document by grouping and clustering the similar data or sentences. The method states that summarization result not only depends on the sentence features, but also depends on the sentence similarity measure. MultiGen is a multi-document system in the news domain. One of the sentence clustering method developed by ZHANG Pei-ying and LI Cun-he[5] is discussed in the paper[5]. Algorithm used for determining the number of the clusters is K-means method. It helps to cluster the sentences of the document, and extracts the topic sentences to generate the extractive summary for the document. In this way sentences are clustered and selected for summarization. Linguistic approaches are harder to implement whereas statistical approaches are more successful but has few limitations.

5. CONCLUSION

As natural language understanding improves, computers will be able to learn from the information on-line and apply what they learned in the real world. Combined with natural language generation, computers will become more and more capable of receiving and giving instructions.

Due to rapid growth of technology and use of Internet, there is information overload. This problem can be solved if there are strong text summarizers which produces a summary of document to help user. Hence there is a need to develop system where a user can efficiently retrieve and get a summarized document. One possible solution is to summarize a document using either extractive or abstractive methods. Text summarization by extractive is easier to build. But text summarization by abstractive technique is stronger because they produce summary which is semantically related but difficult to produce. This paper discussed different types of summarization methods used for summarizing a document and advantages and disadvantages of each method.

6. REFERENCES

[1] Saeedeh Gholamrezazadeh, Mohsen Amini Salehi, Bahareh Gholamzadeh.A Comprehensive Survey on Text Summarization Systems. 2009 In proceeding of: Computer Science and its Applications, 2nd International Conference.

[2] Goldstein, J., Kantrowitz, M., Mittal, V., Carbonell, J., 1999.Summarizing text documents: Sentence selection and evaluation metrics. In: Proc. ACM-SIGIR'99, pp. 121–128.

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

[4] Luhn, H.P., 1959. The automatic creation of literature abstracts. IBM J.Res. Develop., 159–165.

[5] ZHANG Pei-ying, LI Cun-he. Automatic text summarization based on sentences clustering and extraction.

[6] 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.

[7] Youngjoong Koa, Jungyun Seo 2008.An effective sentence-extraction technique using contextual information and statistical approaches for text summarization.

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

[9] Morris, J., and Hirst, G. 1991. Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics 17(1):21–43.

[10] 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.

[11] Kedar Bellare, Anish Das Sharma, Atish Das Sharma, Navneet Loiwal and Pushpak Bhattacharyya.Generic Text Summarization Using Word net. Language Resources Engineering Conference (LREC 2004), Barcelona, May, 2004.

[12] J. Leskovec, M. Grobelnik, N. Milic-Frayling.Extracting Summary Sentences Based on the Document Semantic Graph. Microsoft Research, 2005.

[13] D. Radev, E. Hovy, K. McKeown, "Introduction to the Special Issue on Summarization", Computational Linguistics, Vol. 28, No. 4, pp. 399-408, 2002.

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

[15] Http://Web.science.mq.edu.au/swan/summarization/projects_full.html.

[16] Edmundson, H.P., 1968.New methods in automatic extraction. J. ACM16 (2), 264–285. S.

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

[18] Martin Hassel, Nima Mazdak, "A Persian text summarizer", International Conference on Computational Linguistics, 2004.

[19] William P. Doran, Nicola Stokes, John Dunnion, and Joe Carthy, "Comparing lexical chain-based summarization approaches using an extrinsic evaluation," In Proc. Global Word net Conference (GWC 2004), 2004.