NATURAL LANGUAGE PROCESSING: AUTOMATIC TEXT SUMMARIZATION TECHNIQUES

Roshan Ara
Department of Computer Applications,
Dr. Akhilesh Das Gupta Institute Of Management and Technology, GGSIPU Dwarka Delhi-110078

ABSTRACT

NLP is a way for computers to analyze, understand, and derive meaning from human language in a smart and useful way. By utilizing NLP, developers can organize and structure knowledge to perform many tasks. NLP considers the hierarchical structure of language: several words make a phrase, several phrases make a sentence and, ultimately, sentences convey ideas. NLP systems have long filled useful roles. In NLP, my research area is automatic text summarization. **Automatic summarization** is the process of shortening a text document with software, in order to create a summary with the major points of the original document. Technologies that can make a coherent summary take into account variables such as length, writing style and syntax. Automatic data summarization is part of machine learning and data mining. The main idea of summarization is to find a subset of data which contains the "information" of the entire set. Such techniques are widely used in industry today.

Keywords *: Text summarization, Knowledge bases, Extractive, Abstractive, Natural language processing*

1. INTRODUCTION

First we have to know that what a summary is. A summary is a text in short length that is generated from one or more texts, that provides important information in the original text. Summarization is the method to extract relevant knowledge from one or more informant. It grows the plausibility of locating the position of texts, so the person will give limited time on scanning entire documents. Text summarization to be truly useful within data mining. Few persons make conclusion on the basis of analysis they have identified and they can make useful decision in a very short time period. Summarization is a very important concept to save the time with growing size of information.

When we outline a sample of text, we generally scan it perfectly to expand our perception because automatic text summarization is very formidable, and then write a summary highlighting its main points. Since computers do not understand human skills and language proficiencies, it makes automatic text summarization a very hard and a significant task.

There are many more cases of text summaries that we may come across every day such as headlines from around the world, notes for students, minutes of meeting, previews of movies, reviews of book and movie, digest, biography, weather forecast, stock market reports, histories, information summary for businessman, government officials, researchers online search through search engine to receive the summary of relevant pages found, medical field for tracking patient’s medical history for further treatment.

There are many tools are available for text summarization such as, GreatSummary, Text Compacter, Smmry, Sumplify, Topic marks, Tools4Noobs, FreeSummarizer, Shvoong, ApacheOpenNLP, & Boilerpipe are online summarization tools. Resoomer, Shuca, OpenNLP

Open Text summarizer, Classifier4J, NClassifier, TextTeaser are few widely used open source summarization tools.

Text summarization is used in multiple languages [2]. We can produce summary from one type of language to same type of language which is called mono lingual summarization. Summary can also be generated from multi input documents in multi languages and summary is also generated in these languages, it is termed as multi-lingual summarization. Cross-lingual summarization includes input document to be in one language and summary to be generated in some other language.

Text summarization is categorized into two different groups: indicative and informative. Inductive summarization only shows the main idea of the text to the user. It can summarize the main text from 5 to 10 percent. Any other way, the informative summarization systems provides brief information of the main text .The length of informative summary is 20 to 30 percent of the main text.

Text Summarization methods can be grouped into extractive and abstractive summarization. An extractive summarization approach consists of selecting important sentences, paragraphs etc. from the source document and merging them into shorter form. An Abstractive summarization is an understanding of the main concepts in a document and then shows those concepts in clear natural language.

Summarization techniques can be classified as supervised or unsupervised. Supervised techniques use a collection of documents and human-generated summaries for them to train a classifier for the given text. Position of the sentence, number of words in the sentence etc. are the features of sentences that make them good candidates for inclusion in the summary are learnt. Sentences in an original training document can be labelled as “in summary” or “not in summary”. Unsupervised technique represents the document as a graph and uses an algorithm to find top-scored sentences. The key intuition is the notion of centrality or prestige in social networks i.e. a sentence should be highly ranked if it is recommended by many other highly ranked sentences.

2. MAIN STEPS FOR TEXT SUMMARIZATION

There are three main steps for summarizing documents. Namely, identification, interpretation and summary generation.

2.1 Identification:

The most prominent information in the text is identified .There are different techniques for topic identification are used which are Position, Cue Phrases, word frequency. Methods which are based on the position of phrases are the most useful methods for topic identification.

2.2 Interpretation:

Abstract summaries need to go through interpretation step. In This step, different subjects are fused in order to form a general content.

2.3 Summary Generation:

In this method, the system uses text generation method.

3. TECHNIQUES USED FOR TEXT SUMMARIZATION

Text summarization as discussed is broadly divided into abstractive and extractive. The brief description about each approach is discussed in following section:

3.1 Abstractive Summarization Approach:

Abstractive Summarization covers techniques which can generate summaries by rewriting the content in a given text, rather than simply extracting important sentences from it. But most of the current abstractive summarization techniques still use sentence extraction as a first step for abstract generation. In most cases, extractive summaries reach their limitation primarily because only a part of every sentence selected is informative and other part is redundant. Abstractive techniques try to tackle this issue by either dropping the redundant part altogether or fusing two similar sentences in such a way as to maximize the information content and minimize the sentence lengths. Abstractive summarization techniques are classified into two categories. Structured based approaches and Semantic based approaches.

3.1.1 Structured Based Approach:

Structured primarily based approach encodes most vital data from the document(s) through psychological feature schemas like templates, extraction rules and alternative structures like tree, ontology, lead and body, rule, graph based structure. Completely different ways that are used in this approach area are mentioned below:

(i) Tree Based Method

This technique uses a dependency tree to represent the text/contents of a document. Different algorithms are used for content selection for summary e.g. theme intersection algorithm or algorithm that uses local alignment across pair of parsed sentences. The technique uses either a language generator or an algorithm for generation of summary.

(ii)Template Based Method

Text summarization is the process of putting together meaningful text present in a document in a condensed format. Template based summarization in our system is designed to perform this operation. Here user has the freedom of choosing what should be present in the summary. In other words, user prepares template based on which summary is generated. This prepared template can include POS tags like Noun, Verb, Adverb, etc, and the sequence in which user wants them to appear in the document. User is not confined to only one pattern of this kind but can have as many patterns as possible. Once the POS patterns are finalized, user can decide whether to include named entities and dates in the summary expected. This completes the step of preparing the template. The template based summary module takes into account all these requisites of the user while it generates summary. Template is required for representation of the topic which extracts information from multiple documents containing slots and fillers.

(iii) Ontology Based Method

Ontology is a formal naming and definition of the entity types that are related to particular domain act as a knowledge base. In this method, a knowledge base is used to improve summarization result. Most documents on the internet are related to a particular domain having a limited vocabulary that can be better represented by the ontology. With the help of ontology attributes we can improve the semantic representation of information content and query expansion.

(iv) Lead and Body Phrase Method

This method is based on phrases. In lead and body phrase method main sentences, i.e. sentences which are informative in context and have good length are rephrased by inserting and substituting phrase. This method is good for semantically appropriate revisions for revising a lead sentence. One of the major drawbacks of Lead and body phrase is parsing degrade the performance and no generalized model for summarization. It focuses on rewriting techniques, and lacks a complete model which would include an abstract representation for content selection.

(v) Rule Based Method

In rule based method content selection is done with the help of information extraction rules explicitly specified by the user. Finally, language patterns are used for generating summary sentences. The strong point of this method is it creates summaries with greater information density. The main drawback is that all the rules and patterns are manually written, which is a tedious and time consuming task.

3.1.2 Semantic Based Approach

In Semantic based method, Natural Language Generation (NLG) system accepts the document only in its semantic representation. This method identifies noun and verb phrases by processing linguistic data. Semantic Based Approach methods are given below:

(i) Multimodal semantic model

In Multimodal method semantic model is created to establish the relationship among sentences. The important sentences are scored based on measure and selected sentences are represented as summary. The main point of this technique is to analyze information content. It has three steps, first it creates a semantic model which represents the multimodal document by using an ontology. Second it rates a sentence based on a factor such as completeness of trait, the number of connections with the help of information density matrix. Information density matrix is used to score concept and finally, summary is generated with high score concept.

(ii) Information Item Based Method

In this method, a summary is generated from the conceptual representation of the input document. The piece of information is the smallest part of consistent information in a text. Information item based method gives to the point summaries and less repetitions. It is an enthusiastic structure for conceptual summarization, which focuses at choosing the matter of a summary not from sentences, but from conceptual representation of the input documents. This conceptual representation builds on the concept of Information Items which is defined as the smallest element of consistent information in a text or a sentence. This structure includes information item retrieval, sentence formation, sentence selection and summary formation. Sentences are generated using Simple NLG realize. The Sentence is scored based on document frequency and summary is generated.

(iii) Semantic Graph Based Method

In semantic graph method, the source document is semantically represented using semantic graph. sentences are represented as graph in which nodes represent the noun and verb, edges represent the relationship between noun and verb. It generates to the point, consistent and less repetitious and grammatically accurate sentences. A semantic graph called rich semantic graph to represent the semantic of a source document. Sentence ranking is done based on deriving the average weight of word and sentence. With highest rank sentence Rich Semantic Graph is generated and graph reduction is performed with heuristic rules to generate an abstractive summary.

3.2 Extractive text summarization :

This process involves two steps: Pre Processing step and Processing step. Pre Processing is an organized representation of the initial text. It usually consists of:

a) Sentences boundary identification. In English, sentence boundary is recognized with presence of dot at the end of sentence.

b) Stop-Word Elimination—Common words with no semantics

c) Stemming— The purpose of stemming is to attain the stem or radix of each word, which highlight its semantics.

Processing step features determining and calculating the importance of selected sentences and then weights are assigned to these features using weight learning method. Final score of each sentence is computed using Feature-weight equation. Top scored sentences are selected for output summary. A very important aspect for text summarization is summary evaluation. Generally, intrinsic or extrinsic measures are used to evaluate summaries. While intrinsic methods attempt to measure summary quality using human evaluation and extrinsic methods measure the same through a task-based performance measure such the information retrieval oriented task.

Earlier, extractive summarizers have been mostly based on scoring sentences in the source document. The most common and recent text summarization techniques use either statistical approaches, or linguistic techniques. The high frequency words, standard keyword, Cue Method, Title Method, Location Method are used for weighting the sentences.

3.2.1 Features For Extractive Text Summarization:

Most of the current automated text summarization systems use extraction method to generate a summary report. Generally, sentence extraction techniques are used to generate extraction summaries. One of the method is the compression rate in which suitable sentences is to give some numerical measure of a sentence for the summary called sentence scoring and then select the best sentences to form document summary. Compression rate is an important factor of the extraction method which is used to define the ratio between the length of the summary and the source text. As the compression rate increases, the summary will be larger, and more irrelevant matter is contained. While the compression rate decreases the summary to be short, more information is lost. In fact, the quality of summary is acceptable when the compression rate is 5-30%.

3.2.2 Extractive Summarization Techniques :

An extractive summarization process consists of choosing important sentences, paragraphs etc. from the source document and uniting them into briefer form. The significance of sentences is determined based on statistical and linguistic features of sentences. The techniques of extractive based approach is given below in brief:

(i) Term Frequency-Inverse Document Frequency Method

It is a numerical statistic which indicates how valuable a word is in a given document. The TF-IDF value increases proportionally to the number of times a word occurs in the document. This method is mainly based upon in the weighted term-frequency and inverse sentence frequency paradigm, where sentence-frequency is the number of sentences in the document that include that term. These sentence vectors are then ranked by similarity to the query and the highest ranking sentences are selected to be a part of the summary. Summarization is query-oriented. The hypothesis pretended by this method is that if there are ‘‘more specific words’’ in a given sentence, then the sentence is relatively more important. The main words are usually nouns. This method performs a comparison between the term frequency (tf) in a document -in this case each sentence is treated as a document and the document frequency (df), which means the number of times that the word occurs along all documents. The TF/IDF score is calculated as follows:

$$\mathrm{tfidf}(t,d,D) = \mathrm{tf}(t,d) \times \mathrm{idf}(t,D)$$

Example

Document1 Document 2

Term	Term Count	Term	Term Count
This	1	This	1
Is	1	is	1
A	2	another	2
Sample	1	example	3

we have term count tables of a corpus consisting of only two documents. The calculation of tf–idf for the term "this" is performed as follows:

In its raw frequency form, tf is just the frequency of the "this" for each document. In each document, the word "this" appears once; but as the document 2 has more words, its relative frequency is smaller.

$\mathrm{tf}(\text{“This”},d1) = 1/5 = 0.2$

$\mathrm{tf}(\text{“This”},d2) = 1/7 \approx 0.14$

An idf is constant per corpus, and **accounts** for the ratio of documents that include the word "this". In this case, we have a corpus of two documents and all of them include the word "this".

$\mathrm{Idf}(\text{“This”},D) = \log(2/2) = 0$

So tf–idf is zero for the word "this", which implies that the word is not very informative as it appears in all documents.

tfidf(“This”,d1) = 0.2 x 0 = 0

tfidf(“This”,d2) = 0.14 x 0 = 0

tf(“example”,d1) = 0/5 = 0
tf(“example”,d2) = 3/7 ≈ 0.429
Idf(“example”,D)= log(2/1) = 0.301

tfidf(“example”,d1) = 0 x 0.301 = 0

tfidf(“example”,d2) = 0.429 x 0.301 = 0.13

(ii) Cluster Based Method

In this method, the semantic nature of a source document is taken and shown in natural language by a set of triplets. This triplet includes subjects, verbs and objects related to each sentence. Cluster these triplets using common information. The triplets information are considered as the basic unit in the process of summarization. More identical the triplets are, the more the information is meaningless repeated; thus, a summary may be constructed using a sequence of sentences related the computed clusters.

(iii) Graph Theoretic Approach

In this technique, every sentence is represented as a node. If the two sentences share some common words, these two sentences are connected with an edge. This representation provides two outputs. First, The segments consists of in the graph (that is those sub-graphs which are disconnected to the other sub graphs), collect dissimilar topics covered in the documents. Second, identification of the important sentences in the document by using the graph-theoretic method. The important sentences in the segments are represented as the nodes with high cardinality (number of edges connected to that node), and hence carry higher preference to be included in the summary. The graph theoretic method may also be used easily for envision of inter and intra document uniformity.

(iv) Machine Learning Approach

In this approach, the summarization process is constructed as a classification problem and the training dataset is used for reference. Sentences are grouped as summary sentences and non-summary sentences based on the attributes that they contain. The classification probabilities are gained statistically from the training data, using Bayes’s rule: where, s is a sentence from the document collection, F1, F2...FN are attributes used in classification. S is the summary to be generated, and P (s∈< S | F1, F2, ..., FN) is the probability that sentence s will be chosen to form the summary given that it exhibits attributes F1,F2...FN.

(v) LSA Method

Latent Semantic Analysis is an algebraic-statistical method that produces invisible semantic structures of words and sentences. It is an unsupervised method that does not need any practice or external expertise. LSA method takes the context of the input document and retrieve the information such as which words are used together and which common words are seen in different sentences. More occurrences of common words among sentences shows that the sentences are semantically linked. The context of a sentence is determined using the words it consists of, and meanings of words are determined using the sentences that includes the words. An algebraic method, Singular Value Decomposition is used to determine the relationship between sentences and words. Along with having the power of creating interrelations among words and sentences, SVD has the power of noise cutback, which helps to enhance correctness. The following shows that how LSA can represent the meanings of words and sentences:

Example : Three sentences are given as an input to LSA.
d0: ‘The man walked the dog’.
d1: ‘The man took the dog to the park’.
d2: ‘The dog went to the park’.

After performing the calculations we can see that d1 is more related to d2 than d0; and the word ‘walked’ is related to the word ‘man’ but not so much related to the word ‘park’. These kinds of analysis can be made by using LSA and input data, without any external knowledge.

(vi) Text summarization With Neural Networks:

In this approach, each document is transformed into a list of sentences. Each sentence is expressed as a vector [f1,f2,...,f7], composed of 7 attributes. Seven attributes of a document are:

1) f1- Paragraph follows title
2) f2- Paragraph location in document
3) f3- Sentence location in paragraph
4) f4- First sentence in paragraph
5) f5- Sentence length
6) f6- Number of thematic words in the sentence
7) f7- Number of title words in the sentence

The first step of the process contains training the neural networks to get the kinds of sentences that should be included in the summary. Once the network has got the features that must remain in summary sentences, we need to identify the trends and interrelations among the features that are inherent in the most of the sentences. This is attained by the feature fusion phase, which includes two steps: 1) excluding uncommon traits; and 2) collapsing the effects of common attributes.

(vii) Text Summarization Based on Fuzzy Logic:

This approach includes each attribute of a text such as sentence length, similarity to title, similarity to key word etc. as an input of fuzzy system. Then, it inputs all the standards essential for summarization, in the knowledge base of system. Thereafter, a value from zero to one is attained for each sentence in the output based on sentence characteristics and the available standards in the knowledge base. The attained value in the output determines the degree of the importance of the sentence in the final summary. The input membership function for each attribute is divided into three membership functions which are composed of insignificant values (low L), very low (VL), medium (M), significant values (High h) and very high (VH). The important sentences are extracted using IF-THEN rules according to the attribute criteria. The fuzzy logic system contains four components: fuzzifier, inference engine, defuzzifier, and the fuzzy knowledge base. In the fuzzifier, rigid inputs are expressed into linguistic values using a membership function to be used to the input linguistic variables. After fuzzification, the inference engine refers to the rule base containing fuzzy IF-THEN rules to derive the linguistic values. In the last step, the output linguistic variables from the inference are converted to the final rigid values by the defuzzifier using membership function for representing the final sentence score.

(viii) Query Based Extractive Text Summarization

In this approach, the sentences in a given source document are scored based on the repetition counts of words or phrases. The sentences having the query phrases are given higher scores than the ones having single query words. Then, the sentences with maximum points are included into the resultant summary together with their structural context. Pieces of text may be obtained from various parts or subparts of the text. The output summary is the merger of such pieces. The number of obtained sentences and the size to which their context is shown depends on the summary frame size which is fixed to the size of the screen that can be seen without scrolling. In the sentence extraction algorithm, whenever a sentence is chosen for the inclusion in the summary, some of the headlines in that context are also chosen. The query based sentence extraction algorithm is given below:

Algorithm:

```
1: Rank all the sentences according to their score.
2: Add the main title of the document to the summary.
3: Add the first level-1 heading to the summary.
4: While (summary size limit not exceeded)
5: Add the next highest scored sentence.
6: Add the structural context of the sentence (if any and not already included in the summary)
7: Add the highest level heading above the extracted text (call this heading h).
8: Add the heading before h in the same level.
9: Add the heading after h in the same level.
10: Repeat steps 7, 8 and 9 for the next highest level headings.
```

A one more query-specific summarization method views a document as a set of interconnected text fragments and focuses on keyword queries .

4. EVALUATING THE SUMMARIZATION SYSTEMS

Summary evaluation has a great significance for text summarization. Using intrinsic or extrinsic measures, summaries can be assessed. Intrinsic methods are used to determine summary ideality using human assessment and extrinsic methods determine the same through a experimental performance measure such as the information retrieval-oriented task. Evaluation methods are useful in marking the quality and confiding of the summary. Evaluating the attributes like comprehensibility, coherence, and readability is really hard. System evaluation might be accomplished manually by professionals. To check the trait of summary, the manually expert system is used. When number of sentences selected by the system match with the human gold standard, the qualitative evaluation is completed. The ROUGE evaluator tool is used which consist of precision, recall and F-measure, to measure the quantitative assessment of the summary .

5. CONCLUSION

Automatic text summarization is an early challenge but the latest research trend deviates towards rising ways in minutes of meeting, previews of movies, reviews of book and movie, digest, biography, weather forecast, stock market reports, biomedicine, histories, information summary for businessman, product review, government officials, education domains, researchers online search through search engine to receive the summary of relevant pages found, emails and blogs and medical field for tracking patient's medical history for further treatment. This is due to the fact that there is information make full in these fields, especially on the World Wide Web. Automated summarization is a valuable field in Natural Language Processings research. It automatically creates a summary from one or more source files. The objective of extractive document summarization is to automatically take out a number of indicative sentences, passages, or paragraphs from the main document. Text summarization methods based on Neural Network, Graph Theoretic, Fuzzy and Cluster have, to an extent, succeeded in making an effective summary of a document. Both extractive and abstractive methods have been researched. Most summarization techniques are based on extractive methods. Abstractive method is identical to summaries made by humans. Abstractive summarization as of now requires heavy machinery for language generation and is difficult to replicate into the domain specific areas.

REFERENCES

Aarti P., Komal P., Dipali N., Roshani A., (2015). Automatic Text Summarization**.** International Journal of Computer Applications (0975 – 8887) Volume 109 – No. 17.

Nabil A., Mohammed M., Noureddine R., (2015). Automatic Texts Summarization: Current state of the art. Journal of Asian Scientific, Research, 5(1):1-15, ISSN(e):2223-1331/ISSN(p):2226-5724

Deepali K. G. and Namrata M., (2016). A Review Paper on Text Summarization. International Journal of Advanced Research in Computer and Communication Engineering Vol. 5, Issue 3, March 2016

Roshna C., and Udit C.,y (2017). Automatic Text Summarization. International Journal of Computer Applications (0975 – 8887) Volume 161 – No 1,

Mehdi A., Seyedamin P., Mehdi A., Saeid S., Elizabeth D. T., Juan B. G., Krys K., (2017). Text Summarization Techniques: A Brief Survey. arXiv:1707.02268v3[cs.CL]