Advanced Technique for Text Summarization

Ms.Ketaki S.Jawale
Department of Computer Science and Engineering
Priyadarshini bhagwati College of Engineering
Nagpur (MS) India

Ms. Nandini P. Kolhe
Department of Computer Science and Engineering
Priyadarshini bhagwati College of Engineering
Nagpur (MS) India

Asstt.Prof Archana A. Nikose
Department of Computer Science and Engineering
Priyadarshini bhagwati College of Engineering
Nagpur (MS) India

Abstract— Text Summarization is condensing the source text into a shorter version preserving its information content and overall meaning. It is very difficult for human beings to manually summarize large documents of text. Text Summarization methods can be classified into extractive and abstractive summarization. An abstractive summarization method consists of understanding the original text and re-telling it in fewer words. It uses various methods to examine and interpret the text and then to find the new concepts and expressions to best describe it by generating a new shorter text that conveys the most important information from the original text document.

We present a framework for fast generating concise and coherent summaries in domain independent, document summarization. The proposed generation approach, called cut-and-paste, generates summaries through reusing the input document. Rather than using the extracted document sentences directly for producing summaries, the cut-and-paste approach edits the sentences in some way so that they are more concise, coherent, and appropriate for summaries. We specially investigate two effective techniques, sentence reduction and sentence combination, for transforming extracted sentences into appropriate summary sentences. The system is designed to be a general generation tool portable to any independent, document summarizer in need of a generation component.

Keywords— Text summarization, cut and paste, automatic program.

I. INTRODUCTION

1.1 Natural Language Processing

Natural language processing (NLP) is a field of computer science, artificial intelligence, and linguistics concerned with the interactions between computers and human (natural) languages. As such, NLP is related to the area of human–computer interaction. Many challenges in NLP involve natural language understanding, that is, enabling computers to derive meaning from human or natural language input, and others involve natural language generation.

Actually, Natural language processing (NLP) is the ability of a computer to understand what a human is saying to it. NLP is the ability of a computer program to understand human speech as it is spoken. NLP is a component of artificial intelligence (AI).

The development of NLP applications is challenging because computers traditionally require humans to speak to them in a programming language that is precise, unambiguous and highly structured or, perhaps through a limited number of clear voice commands. Human speech, however, is not always precise. It is often ambiguous and the linguistic structure can depend on many complex variables, including slang and social context.

1.2 What is Text Summarization

The amount of information available today is tremendous and the problem of finding the relevant pieces and making sense of these is becoming more and more essential. Nowadays, a great deal of information comes from the Internet in a textual form. The challenge of finding relevant documents on the web is mainly handled by information retrieval techniques utilized in search engines such as Google, Bing, Yahoo, etc. Search engines usually return thousands of pages for a single query, and even the use of sophisticated ranking algorithms can't provide us the exact information we are looking for.

A typical user goes through the top-ranked pages and tries to find the relevant pieces of information he or she is interested in, manually. Obviously, a short summary of the retrieved pages would be very helpful in such situations. In general, construction of summaries is an ideal way to cope with the information overload. A summary is a shortened version of a text that contains the main points of the original content.

Automatic summarization is the creation of a summary by a computer program. Although automatic summarization is a topic of research nowadays. In general, creation of a good summary requires a lot of intelligence. Like many other natural language processing (NLP) tasks, a high quality automatic summarization will require understanding of a natural language, at least to a certain degree. NLP tasks that are quite challenging, such as machine translation, speech recognition, domain specific question answering, etc. Although none of these problems are near to be solved yet, the results are promising to be useful. Improving the quality of automatic summarization to this level of usefulness is the motivation behind the increasing amount of research in the field.

I. LITERATURE SURVEY

[1] Cut and paste based summarization handout, Dept. CS, Colombia University.To summarize is to reduce in complexity, and hence in length, while retaining some of the essential qualities of the original document. Titles, keyword, table of content and abstract that might be considered as forms of summary.

2.1 Types of summarization techniques:

There are two main types of summarization techniques:

1. Summarization based on abstraction method
2. Summarization based on extraction method
3. Summarization based on hybrid method

2.2 **Hybrid summarization technique:**

[2] Allen, J. (1995). Natural Language Processing. The Benjamin/Cummings Publishing Company, Inc.Hybrid based summarization technique is the combination of the both other techniques that is abstract based and extract based summarization technique. The originality of the technique lies on the use of term co-occurrence property of the text. It could detect the number of subjects. The proposed technique summarizes the document in proportion to the subject treated in a document. It considers the conceptual property of the text algorithm and based on word synonymy prevents similar sentences to be included in the summary. It also preserves the cohesion of the summarized text.

Difference between abstraction and extraction based summarization:

Abstraction	**Extraction**
This method takes into account the meaning of various phrases present in the document and based on the meaning, it creates the summary of the document.	This method does not concerns with the meaning of the phrase.
It takes into account the meaning of the sentences and then creates the summary of the document.	It takes into account the position and length othe phrases and based on that creates the summary of the document.
Abstraction can condense a text more strongly than extraction.	Extraction does not condense a text more strongly than abstraction.

We proposed a new sentence reduction based on decision tree model where semantic information is used to support reduction process. The decision tree model is also extended to cope with the changeable order between original sentences and reduced sentences.

Our objective is to design, implement and evaluate an extraction-based automatic summarization framework. Use the system to experiment with several different summarization methods.

System Architecture :-

Following are the major operations performed in Cut and Paste based approach:

1. Sentence Reduction:

This is a frequently used technique. Through this method, humans select a single sentence from the document, remove less important material from the sentence, and then use the reduced sentence in the summary. The deleted material can be at any granularity: a word, a phrase, or a clause.

Generalization

Humans may borrow a block of text from the document and then replace certain phrases or clauses with a more generalized high-level description. For example, they replaced a proposed new law that would require Web publishers to obtain parental consent before collecting personal information from children with legislation to protect children's privacy on-line.

Reordering: The borrowed sentences from the document do not necessarily retain their precedence orders in the summary. For example, a concluding sentence in the document may be placed at the beginning of as an opening sentence

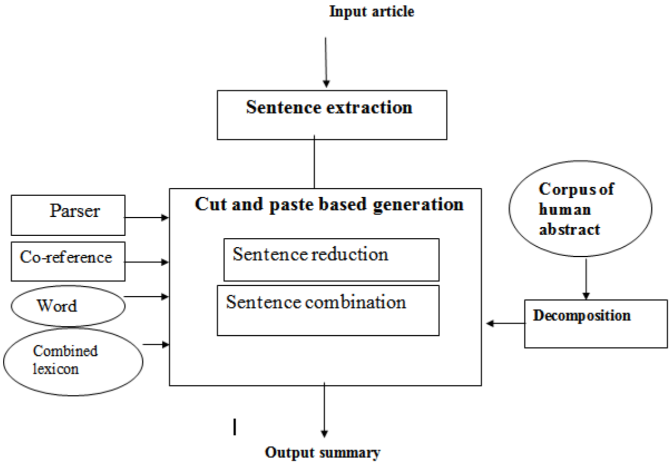


Fig. system Architecture of text summarization

A. PROPOSED SYSTEM

In summarizing document, people may perform a changeable order to ensure the summary document is smooth and coherence. This fact requires a new sentence reduction with the order of reduced sentence is different from the original. In addition to using sentence reduction for text summarization, the information of syntactic is not enough. The semantic information of original sentences should be incorporated with reduction process to enhance the accuracy of reduction process. This fact is also similar to the behavior of human in reduction sentence that they can understand the meaning of original sentences to ensure that important words is remained in reduced sentences.

To satisfy the new requirements mentioned above, we proposed a new sentence reduction based on decision tree model where semantic information is used to support reduction process. The decision tree model is also extended to cope with the changeable order between original sentences and reduced sentences.

4.1 Operations Performed in Cut and Paste Based approach:

2. Sentence Reduction:

This is a frequently used technique. Through this method, humans select a single sentence from the document, remove less important material from the sentence, and then use the reduced sentence in the summary. The deleted material can be at any granularity: a word, a phrase, or a clause.

a. **Document sentence**: James Rittinger, an attorney for the company pointed out that several west features such as syllabuses and headnotes still can't be legally copied.

b. **Summary sentence**: An attorney for the company noted that syllabuses and head- notes of West still cannot be copied.

c. **Sentence combination**

Humans also generate a summary sentence by combining material from several sentences in thedocument. This is another frequently used technique. It can be used together with sentence reduction, as shown in the following example, which also uses paraphrases.

Text sentence 1: But it also raises serious questions about the privacy of such highly personal information wafting about the digital world.

Text sentence 2: The issue thus fits squarely into the broader debate about privacy and security on the internet, whether it involves protecting credit card number or keeping children from offensive information.

Summary sentence: But it also raises the issue of privacy of such personal information and this issue hits the head on the nail in the broader debate about privacyand security on the internet.

3. Syntactic transformation

In both sentence reduction and sentence combination, syntactic transformations may be involved. The structure of this combined sentence is based on that of the first sentence in the document, but the position of the subject has been moved.

4. Lexical paraphrasing

Humans may borrow a block of texts from the original document and then replace certain phrases with their paraphrases.

5. Generalization

Humans may borrow a block of text from the document and then replace certain phrases or clauses with a more generalized high-level description. For example, they replaced a proposed new law that would require Web publishers to obtain parental consent before collecting personal information from children with legislation to protect children's privacy on-line.

6. **Reordering:** The borrowed sentences from the document do not necessarily retain their precedence orders in the summary. For example, a concluding sentence in the document may be placed at the beginning of as an opening sentence.

While creating the cloud environment will have to go to the cloud link where we get the particular cloud will have to select the cloudlets i.e. the amount of space on the Cloud. Will have

to create a WAR file and then will deploy the application on the cloud.
Then connectivity with the cloud takes place in which the cloud is getting connected and the deployed application will then executed. After the connection is being established data of the application is saved on the cloud

II. REWRITING LARGER SENTENCE INTO SMALLER SENTENCES

I. SHIFT-operator transfers a first word from the input list into CSTACK. It was written in mathematic by the label SHIFT

II. REDUCE-operators pops the k syntactic trees located
at the top of CSTACK and combine them into a new tree. These operators are formulated as REDUCE (k, x), in which k is an integer and X is a grammar symbol.

III. DROP-operators are used to remove from the input list subsequences of word that correspond to syntactic constituents to RSTACK. Both REDUCE-operators and DROP-operators are used to derive the structure of the syntactic tree of the short sentence. They were written as DROP x with X is a grammar symbol.

IV. ASSIGN TYPE-operators are used to change the label of trees at the top of the CSTACK. These POS tags may be different from the POS tags in the original sentence. These operators are written as ASSIGN TYPE (X), which x are POS tags.

V. RESTORE-operators take the kth element in RSTACK to remove that element into the Input list. These operators are designed with the assumption that a sub-tree was removed from the input list still affects the current decision. We also formulated it as RESTORE k where k is an integer.

A DROP x operators deletes from the input list all words that are spanned by constituent x in t and store them into CSTACK. The operator RESTORE is designed to restore some words in RSTACK to generate a small tree s. With these operators, the order of words within a small tree s can be changed in comparing with the word order of the large tree t.

5.1 Employing sentence reduction decision tree model:
This includes the following cases:

5.1.1 Generating learning cases

In this part, we associate with each configuration of our shift-reduce-drop-restore, rewriting model a learning case.

Input: a smaller tree n and inputList, CSTACK, RSTACK are empty.
Output: Learning cases to rewrite a large tree into a small tree.

```
 void GenerateLearningCase (Tree*n)
{
1. If n is leaf node
 1.1 Searching on the remainder part of the Input List. if found
remark the position i then goto the step 1.2.
 else continue searching on RSTACK;
if found
{
- Call RESTORE operator;
 - generate a learning case;
}
Else do nothing;
1.2 Find and do all DROP operators between the first element
and        element in the input list.
- Call SHIFT operator;
- Call ASSIGN TYPE with its parameter is label of node's
parent.
 2. else
 {
for each child c in n do
 GenerateLearningCase(c);
if n is a part of speech return;
```

- Call REDUCE operator and generate a learning case with its parameters are the label of n and the number of children in the node n.
}

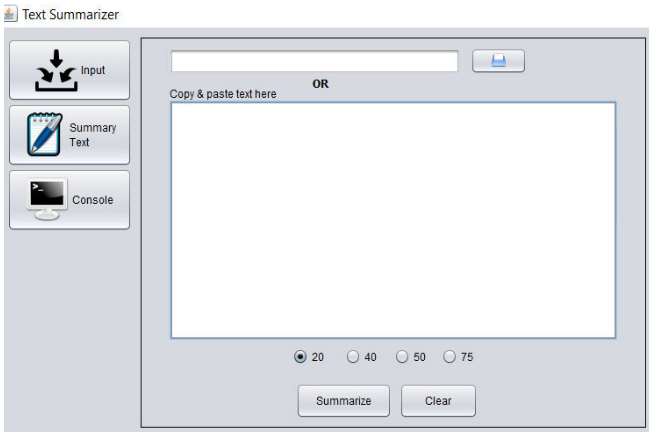
5.1.2 Process of sentence reduction:

Input: an input sentence
Output: a reduced sentence
Step 1. The input sentence is parsed into a syntax tree.
Step 2. The syntax tree is enriched semantic information.
Step 3. Create an input list and set CSTack and RStack to empty.
Step 4. Call a traversal procedure to obtain a reduced syntax tree
Step 5. Generate a reduced sentence from the reduced syntax tree Traversal **procedure**

```
 Input: Input list, CSTack, RStack
 Output: A reduced tree
While (not terminal condition) {
 Feature=get contextual feature ();
Action= get action (feature);
 Parameter=get parameter (action);
Switch (action)
{
Case SHIFT: SHIFT ();
 Case ASSIGN TYPE: ASSIGN TYPE (parameter);
 Break;
Case REDUCE: Reduce (parameter);
 Break;
 Case DROP: Drop (parameter);
 Break;
Case RESTORE: Restore (parameter);
 Break;
 }
 }
```

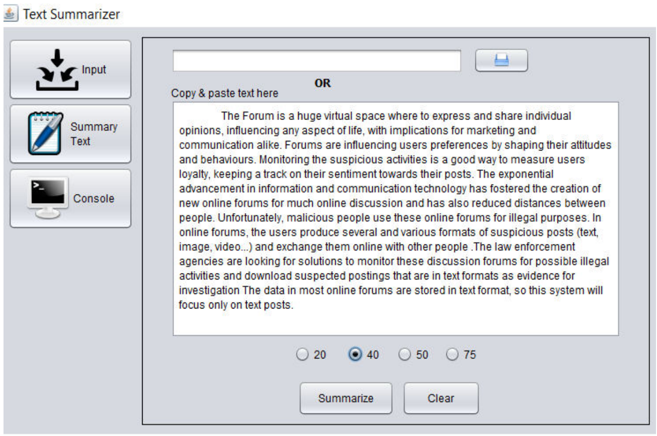
VI. Experimental Results And Discussion

Fig1 :Input Window



● 20 ○ 40 ○ 50 ○ 75

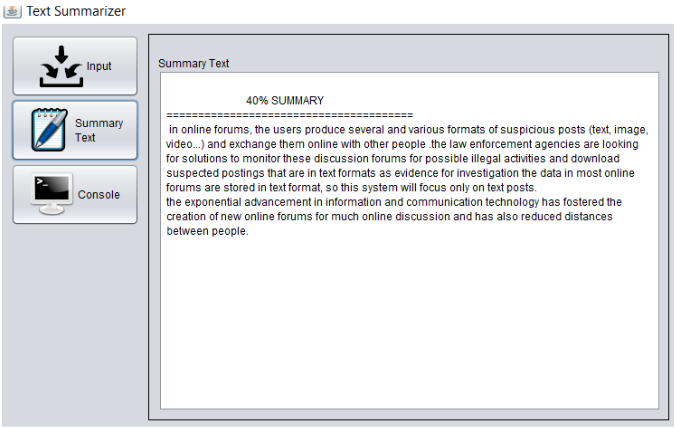
Fig2 : Input Text Window



The Forum is a huge virtual space where to express and share individual opinions, influencing any aspect of life, with implications for marketing and communication alike. Forums are influencing users preferences by shaping their attitudes and behaviours. Monitoring the suspicious activities is a good way to measure users loyalty, keeping a track on their sentiment towards their posts. The exponential advancement in information and communication technology has fostered the creation of new online forums for much online discussion and has also reduced distances between people. Unfortunately, malicious people use these online forums for illegal purposes. In online forums, the users produce several and various formats of suspicious posts (text, image, video...) and exchange them online with other people .The law enforcement agencies are looking for solutions to monitor these discussion forums for possible illegal activities and download suspected postings that are in text formats as evidence for investigation The data in most online forums are stored in text format, so this system will focus only on text posts.

○ 20 ● 40 ○ 50 ○ 75

Fig3 : Summarize window



40% SUMMARY
==
in online forums, the users produce several and various formats of suspicious posts (text, image, video...) and exchange them online with other people .the law enforcement agencies are looking for solutions to monitor these discussion forums for possible illegal activities and download suspected postings that are in text formats as evidence for investigation the data in most online forums are stored in text format, so this system will focus only on text posts.
the exponential advancement in information and communication technology has fostered the creation of new online forums for much online discussion and has also reduced distances between people.

We concentrate our presentation in two main points: (1) the set of employed features; and (2) the framework defined for the trainable summarizer, including the employed classifiers.

A large variety of features can be found in the text-summarization literature. In our proposal we employ the following set of features:

(a) Mean-TF-ISF. Since the text processing tasks frequently use features based on IR measures. In the context of IR, some very important measures are term frequency (TF) and term frequency ´ inverse document frequency (TF-IDF). In text summarization we can employ the same idea: in this case we have a single document d, and we have to select a set of relevant sentences to be included in the extractive summary out of all sentences in d. Hence the notion of a collection of documents in IR can be replaced by the notion of a single document in text summarization. Analogously the notion of document – an element of a collection of documents – in IR, corresponds to the notion of sentence – an element of a document – in summarization. This new measure will be called term frequency ´ inverse sentence frequency, and denoted TF-ISF(w,s).The final used feature is calculated as the mean value of the TF-ISF measure for all the words of each sentence.

(b) Sentence Length. This feature is employed to penalize sentences that are too short, since these sentences are not expected to belong to the summary. We use the normalized length of the sentence, which is the ratio of the number of words occurring in the sentence over the number of words occurring in the longest sentence of the document

(c)Sentence-to-Centroid Cohesion. This feature is obtained for a sentence *s* as follows: first, we compute the vector representing the centroid of the document, which is the arithmetic average over the corresponding coordinate values of all the sentences of the document; then we compute the similarity between the centroid and each sentence, obtaining the raw value of this feature for each sentence. The normalized value in the range [0, 1] for *s* is obtained by computing the ratio of the raw feature value over the largest raw feature value among all sentences in the document. Sentences with feature values closer to 1.0 have a larger degree of cohesion with respect to the centroid of the document, and so are supposed to better represent the basic ideas of the document

III. CONCLUSION

Text summarization is an important utility for many tasks in NLP. We have described a cut-and-paste technique to generate concise, coherent summaries fast and reliably in automatic text summarization.Specially, we have studied two effective techniques Sentence reduction and sentence combination. We have presented an algorithm that allows rewriting a long sentence into reduced sentence with the order of short sentence is able to be different from the original sentence. The semantic information of the original sentence was very useful for sentence reduction problem.

Now we have studied for the implementation of Sentence reduction algorithm. In next phase we will study the implementation of Sentence combination techniques.

REFERENCES

[1] Cut and paste based summarization handout, Dept. CS, Colombia University.

[2] Art of abstracting. ISI press, Philadelphia.

[3] ANSI1997.Guidelinesfor abstracts. Technical reportZ39.141997

[4] Summary Generation through Intelligent Cutting and Pasting of the Input Document Ph.D. Thesis Proposal

[5] Allen, J. (1995). Natural Language Processing. The Benjamin/Cummings Publishing Company, Inc.

[6] [Aone et al., 1997] Aone, C., Okurowski, M., Gorlinsky, J., and Larsen, B. (1997). A scalable summarization system using robust nlp. In Proceedings of ACL/EACL'97 workshop on summarization, Madrid, Spain.

[7] [Barzilay and Elhadad, 1997] Barzilay, R. and Elhadad, M. (1997). Using lexical chains for text summarization. In ACL/EACL-97 summarization workshop, Madrid, Spain.

[8] [CELEX, 1995] CELEX (1995). The CELEX lexical database—Dutch, English, German. CD-ROM. Centre for Lexical Information, Max Planck Institute for Psycholinguistics, Nijmegen.

[9] [Dalianis and Hovy, 1993] Dalianis, H. and Hovy, E. (1993). Aggregation in natural language generation. Proceedings of the 4th European Workshop on Natural Language Generation.

[10] I. Mani , "automatic summarization", john Benjamin's