# AUTOMATIC TEXT SUMMARIZATION BASED ON PRAGMATIC ANALYSIS

**MANISHA PRABHAKAR*, NIDHI CHANDRA****

*M. Tech. Student, Computer Science Department, Amity School of Science & Technology, Amity University, Noida, Uttar Pradesh, India
**Assistant Professor, Computer Science Department, Amity School of Science & Technology, Amity University, Noida, Uttar Pradesh, India

***Abstract-*** The rapid growth of online information has encumbered the user with colossal amount of information. It is difficult to access large amount of data. This problem has increased the research in the field of automatic text summarization. Automatic text summarization is a technique where the text is input to the computer and it returns the clipped and concise extract of the original text and also sustains the overall meaning and main information content. In this paper, text summarization technique is designed for the documents having the fixed format. The proposed system generates the summary of the fixed format documents by analyzing all the different parts of the documents. The system consists of five stages. In first stage each sentence is partitioned into the list of tokens and stop words are removed. In second stage, frequency usage is counted for each word. In third stage, assign POS tag for each weighted term and Word sense disambiguation is done. In the fourth stage, pragmatic analysis is performed. After Pragmatic Analysis, summarized sentences will be store in a database.

***Index Terms***- POS tagging, Pragmatic Analysis Text Summarization, Wordnet

## I. INTRODUCTION

Internet has developed the lives of many enthusiastic discoverers and researchers. Now-a-days, with the increasing demand of the internet huge volumes of information are arising continuously [1].Users finds it thorny to find the anticipated information swiftly and precisely [10].It is useful for the users to extract the main content from the original document instead of the original document. Automatic Text Summarization automatically extracts the main content from the original document correctly and comprehensively. The language of the summary extracted should be consistent and efficient. Automatic Text summarization can be categorized into two- extraction and abstraction [11].Extraction means to select the phrases or sentences having the highest score from the original text and combined to obtain the new shorter text without changing the source text. Abstraction means to probe and interpret the text by using linguistic methods. Mostly, extraction method is used to produce the summary in automated text summarization system. During the last twenty years, several researchers addressed that the automated part of speech tagging is a well known problem. Part of speech tagging is a technique for elucidation of the lexical categories. In POS tagging, a felicitous tag is assign to each word of the sentence.POS tagging is broadly used for lexical text analysis.POS tagging is a important task for the activities in natural language processing [7]. In POS tagging, it takes a sentence as input and assigns a unique tag to each word in the sentence. It is a firm conjecture that when it comes to index term extraction, the nouns carry most of the sentence meaning. Index term is a term that catches the essence of the sentence. In an ideal case, index terms should give excellent semantic representation. A subtask of POS Tagging is noun extraction in which every noun either proper or common is identifying in a document. Nouns are used as a most important feature to express the meaning of the text in natural language processing applications like text summarization, information retrieval, information extraction etc.

POS Tagging has a variety of techniques .It has two approaches- Supervised POS tagging and Unsupervised POS Tagging [4]. A pre tagged corpora is needed for Supervised Tagging Technique while it is not required in Unsupervised Tagging Technique.

Both techniques can be of two types: - Stochastic and Rule Based. Rule Based Technique requires a context rule for POS Tagging [5]. In rule based approaches, tags are assigned to the ambiguous and unknown words by using the contextual information. Stochastic Technique uses a Hidden Markov Model[9].The states mostly symbolize the POS Tags. The contingencies are calculated from the tagged corpus and the untagged corpus in order to calculate the most likely POS tags for the word of the sentence. Stochastic Training techniques can be categorized into Supervised and Unsupervised Stochastic Technique. Supervised stochastic technique requires only the pre tagged data in a huge volume to achieve high level of accuracy. On the other hand, Unsupervised stochastic technique does not need pre tagged data and it uses the computational methods to do the automatic word groupings or make tag sets and based on these groupings, estimate the probabilistic values required by stochastic taggers.

## II. PROPOSED METHODOLOGY

The following figure represents the diagram of the propose system. Proposed model has the following stages:-

### *A.* ***Database:***

Firstly make a repository that will act as a source of the system. This repository contains the files on which text summarization

can be done. From this repository, one paper at a time is taken on which apply our first step i.e. tokenization.

### *B.* ***Tokenization*:**

Each sentence is partitioned into the list of tokens. It is a part of the lexical analysis.

### *C.* ***Stop Word Removal*:**

Some words are insignificant that are highly used in the English sentences. They exist in superiority in the documents. Therefore, they do not provide the actual idea about the theme of the text. While scoring the sentences, these words can be discarded. For example- articles like ‘a’, ’an’, ’the’; ’by’ come into sight mostly in all the text but does not include much semantic information. As we already generated the list of tokens, now remove the stop words from that list and store these words into a separate file.

***List of some stop words considered are:-***

| | |
|---|---|
| a | ourselves |
| about | out |
| above | over |
| after | own |
| again | same |
| against | shan't |
| all | she |
| am | she'd |
| an | she'll |
| and | she's |
| any | should |
| are | shouldn't |
| aren't | so |
| as | some |
| at | such |
| be | than |
| because | that |
| been | that's |
| before | the |
| being | their |
| below | theirs |
| between | them |
| both | themselves |
| but | then |
| by | there |
| can't | there's |
| cannot | these |
| could | they |
| couldn't | they'd |
| did | they'll |
| didn't | they're |
| do | they've |
| does | this |
| doesn't | those |
| doing | through |
| don't | to |
| down | too |
| during | under |
| each | until |
| few | up |
| for | very |
| from | was |
| further | wasn't |
| had | we |
| hadn't | we'd |
| has | we'll |
| hasn't | we're |
| have | we've |
| haven't | were |

List 1: Stop Words

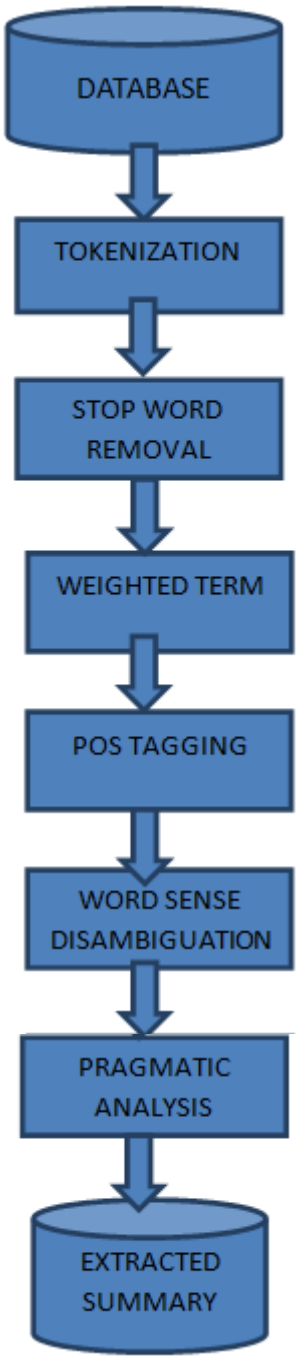

Figure 2: Proposed Methodology

### D. *Weighted Term*:

Frequency usage of each word is calculated after removal of stop words. The words having the higher frequency are the weighted terms. Frequency can be defined as the number of times term is repeated in the document. For example- a term occurring 10 times in a document is much more relevant than the term occurring 2 times.

### E. *POS Tagging*:

It is a process of assigning a felicitous tag to each word of the sentence[8].A large lexical database for English language named "Wordnet" can be used for POS tagging[6]. Wordnet consists of most English nouns ,verbs, adverbs and adjectives which are grouped into sets of cognitivesynsets, each showing a distinct concept.Wordnet is systematized by meaning that means the words are semantic similar which are in close propinquity[2].

### F. *Word Sense Disambiguation*:

It is a process of choosing pertinent senses of the word in a given context[3]. It removes the ambiguity of the word.
It is important for NLP applications like information retrieval, machine translation, part of speech tagging and Text processing.

Word sense disambiguation consists of 2 steps:-

1. Identify all the different senses for each word congruent to the text.
2. It involves assigning relevant sense for each word in context.

Word Sense Disambiguation can be reached by 2 approaches- Shallow approaches and Deep approaches. Access to the comprehensive body of world knowledge is assumed in Deep Approach. But these approaches are not favorable in practice because access to body of knowledge is possible in very limited domains. If knowledge exists, then this approach is better than the shallow approaches.

Shallow approaches consider the surrounding words and not even try to understand the text. These are not as stronger as Deep approaches but gives better results in practice because of limited knowledge domain[12].

### G. *Pragmatic Analysis*:

Pragmatics can be defined as the study of the meaning in context. It is an effort to get the intended meaning of the text[13].Pragmatic Analysis is only performed on scientific or fixed format articles[14].After POS tagging, pragmatic analysis is performed which analyzes the weighted terms .

This analysis gives a fixed format to the files and the sentences containing the weighted terms will be stored in a database. These are the summarized sentences. Any format file is converted into a fixed format by pragmatic analysis.Thus,the main content of the document can be identified to compose the summary by analyzing the pragmatic function.

### H. *Extracted Summary*

After pragmatic analysis, the summarized sentences get stored in a database. This summary is the result of text summarization. Hence,the file taken at the first stage is summarized into meaningful text through this process.

## III. RESULTS

The system break up the given text into tokens as shown in table I,then stop words are removed from the list of tokens,shown in tableII.Word Frequency of each word is calculated after stop word removal,shown in fig 3.Then,POS tag is assigned to the weighted terms by using Wordnet as shown in fig 4. Word Sense Disambiguation is performed while doing the Pragmatic Analysis. During Pragmatic Analysis,maincontent of the document is identified to compose the summary.

Table I: Tokenization

| |
|---|
| I. hi |
| II. How |
| III. Are |
| IV. You |
| V. Dear |
| VI. You |
| VII. Are |
| VIII. looking |
| IX. good |
| X. you |
| XI. are |
| XII. comfortable |
| XIII. this |
| XIV. kind |
| XV. of |
| XVI. environment |
| XVII. are |
| XVIII. you |
| XIX. going |
| XX. somewhere |
| XXI. what |
| XXII. you |
| XXIII. want |
| XXIV. dear |
| XXV. The |
| XXVI. person |
| XXVII. you |
| XXVIII. are |
| XXIX. looking |

Table II: After Stop word Removal

| |
|---|
| XXX. Dear |
| XXXI. Good |
| XXXII. Comfortable |
| XXXIII. Kind |

| |
|---|
| XXXIV. Environment |
| XXXV. Somewhere |
| XXXVI. Dear |
| XXXVII. person |
| XXXVIII. Don't |
| XXXIX. Here |
| XL. Want |
| XLI. Sir |
| XLII. Sun |
| XLIII. Rises |
| XLIV. east. |
| XLV. Kind |
| XLVI. attitude |
| XLVII. Fine |
| XLVIII. Dear |

```
C:\Program Files\Java\jdk1.5.0_16\bin>java GeneratedTokens
Please enter a text file name: License.txt
File has 18 words.

WORDS                FREQUENCY

attitude                     1
comfortable                  1
dear                         3
dont                         1
east                         1
environment                  1
fine                         1
good                         1
here                         1
kind                         2
person                       1
rises                        1
sirthe                       1
somewhere                    1
sun                          1
```

Figure 3: Showing words with their frequency.

```
zf
WORDS                PART OF SPEECH

dear                      NOUN
dear                 ADJECTIVE
dear                    ADVERB
east                      NOUN
east                 ADJECTIVE
east                    ADVERB
kind                      NOUN
kind                 ADJECTIVE
fine                      VERB
fine                      NOUN
fine                 ADJECTIVE
fine                    ADVERB
person                    NOUN
environment               NOUN
here                      NOUN
here                 ADJECTIVE
here                    ADVERB
attitude                  NOUN
comfortable          ADJECTIVE
somewhere                 NOUN
somewhere               ADVERB
sun                       VERB
sun                       NOUN
good                      NOUN
good                 ADJECTIVE
good                    ADVERB
rises                     VERB
rises                     NOUN
```

Figure 4: Part of speech tagging

```
Here are all the possible senses of 'customer' with pos: 'n'
========================================
Sense: 99984659
Description: someone who pays for goods or services
Synset: customer client
-------------------------

Here are all the possible senses of 'information' with pos: 'n'
========================================
Sense: 96634376
Description: a message received and understood
Synset: information info
-------------------------
Sense: 95816287
Description: knowledge acquired through study or experience or
instruction
Synset: information
-------------------------
Sense: 97237561
Description: formal accusation of a crime
Synset: information
-------------------------
```

Figure 5: Word Sense Disambiguation

Pragmatic Analysis is performed to extract the summary in a fixed format. When we perform the word sense disambiguation, all the synsets are generated. The sentences containing the synonyms of the words are removed .i.e repetitive sentences are removed and the main content is extracted and get stored in a database.

## IV. CONCLUSION

In this work, we proposed an automatic text summarization approach by pragmatic analysis. We used supervised POS tagging approach and implemented it in Java. By pragmatic analysis, the intended meaning of the text is obtained and the text summary is generated and stored in a database. The systems produced the high quality compressed summary and provide better results than the manual summarization.

## REFERENCES

[1] J. Steinberger, "Using Latent Semantic Analysis in Text.Summarization and Summary Evaluation," Proc.ISIM 04, 2004.

[2] Gonzalo, J., Verdejo, F., Chugur, I., and Cigarran, J.: Indexing with WordNetsynsets can improve text retrieval. Proceedings of the COLING/ACL '98 Workshop on Usage of WordNet for NLP. Montreal (Canada) (1998) 38–44.

[3] PrabhakarPande Lakshmi Kashyap Manish Sinha, Mahesh Kumar and Pushpak Bhattacharyya.Hindi word sense disambiguation. In International Symposium on Machine Translation, Natural Language Processing and Translation Support Systems, Delhi, India, November 2004.

[4] Marcus, M. P., Marcinkiewicz, M. A. and Santorini, B. Building a large annotated corpus of English: The Penn Treebank. Computational Linguistics **19**(2), 1993.

[5] L. Bahl and R. L. Mercer, Part-Of-Speech assignment by a statistical decision algorithm, IEEE International Symposium on InformationTheory, pages: 88 - 89, 1976.

[6] Chai, Joyce Y. and Biermann, Alan W.: A WordNet based rule generalization engine for meaning extraction, to appear at Tenth International Symposium On Methodologies For Intelligent Systems (1997).

[7] Ratnaparkhi, A. A maximum entropy part-of-speech tagger. Proceedings 1st Conference on Empirical Methods in NaturalLanguage Processing, EMNLP, 1996.

[8] Tapanainen, P. and Voutilainen, A. ; "Tagging Accurately – Don't Guess If You Don't Know", Proceedings of 4th ACL Conference on Applied Natural Language Processing, ACM, Stuttgart, 1994.

[9] Edmundson, H.P. New Methods in Automatic Extraction. Journal of the ACM 16(2), 264–285, 1968.

[10] Daelemans, W., Zavrel, J., Berck, P. and Gillis, S.MBT: A memory based part-of speech tagger generator. Proceedings 4thWorkshop on Very Large Corpora, pp. 14–27. Copenhagen, Denmark,1996.

[11] Rafeeq Al-Hashemi, Text Summarization Extraction System (TSES) Using Extracted Keywords, International Arab Journal of e-Technology, Vol. 1, No. 4, June 2010 pp 164-168

[12] Dang, H.T. & Palmer, M., 2002. Combining Contextual Features for Word Sense Disambiguation.Proceedings of the SIGLEX/SENSEVAL Workshop on WSD, Philadelphia, July 2002, P88-94.

[13] F. Yetim, "DISCOURSIUM for cooperative examination of information in the context of the pragmatic web," 2nd InternationalConference on the Pragmatic Web, Tilburg, The Netherlands, 2007,pp.29-40

[14] Atifi H., Matta N. 2000, Pragmatic analysis and modelling of argumentation messages in computer mediated communications, Proceedings du Workshop :Cooperative Models Based on Argumenation In Problem Solving, N. Matta, M. Lewkowicz and M. Zacklad (Eds).

## AUTHORS

**First Author** – ManishaPrabhakar,M.Tech Student,Computer Science Department, Amity School of Science & Technology, Amity University, Noida, UttarPradesh, India, **manisha.elegant@gmail.com**

**Second Author-** Nidhi Chandra, Assistant Professor, Computer Science Department, Amity School of Science & Technology, Amity University, Noida, UttarPradesh, India srivastavanidhi8@gmail.com.