LOVELY PROFESSIONAL UNIVERSITY

Compositional Semantics Tool for Automatic Title Generation in English Language using NLP

A Dissertation Proposal
submitted

By

SEEMA

To

Department of Computer Science and Engineering

In partial fulfilment of the Requirement for the

Award of the Degree of

Master of Technology in Computer Science and Engineering

Under the guidance of

Mr. Prateek Agrawal

(June 2015)

ABSTRACT

Natural language provides the computational challenges, such as tokenization, tagging, classification, machine translation, text summarization, question answering, information extraction, and building syntactic and semantic representations, topic modeling etc. In this dissertation we would deal with automatic title generation by a new structured based tool. Title is concise form that can help people understand a document's main idea without reading of the entire document. Our tool will summaries the topic and generates the appropriate title. The title can be a normal or can be phrase. Scan the whole paragraph and choose the best suited phrase. The tool will help the editors of the newspapers or magazines can get the title or summaries the whole news very easily. Moreover they can get the appropriate phrase as well. It will be informative tool for the children. They can understand the different parts of the sentence. The tool would develop in various phases such as lexical analysis, lexemes categorization, Discourse analysis, Frequencies of tokens, Prioritization based on frequency, Synonyms and antonyms, Frame the sentence or title based on the analysis, Match the suggestive sentences with phrase or idioms as for as possible.

ACKNOWLEDGMENT

I would like to thanks to **Mr. Prateek Agrawal** for assigning this work and helping me in this dissertation-I. I would also thanks to Mr. Krishan Bansal,Mr.Deepak Kumar,Ms. Rupinder kaur, classmates and friends without whose help it is impossible for me to complete my work. I would also thank my family for supporting me.

Seema
41200390

CERTIFICATE

This is to certify that **Seema** has completed M.Tech dissertation proposal titled **Compositional Semantics Tool for automatic Title generation in English Language using NLP** under my guidance and supervision. To the best of my knowledge, the present work is the result of her original investigation and study. No part of the dissertation proposal has ever been submitted for any other degree or diploma.The dissertation proposal is fit for the submission and the partial fulfilment of the conditions for the award of M.Tech Computer Science and Engg.

Date:-_____________

Signature of Advisor

Name: Mr. PRATEEK AGRAWAL
Asst. Professor
COD : Intelligent Systems

DECLARATION

I hereby declare that the dissertation proposal entitled, **Compositional Semantics Tool for Automatic Title Generation in English Language using NLP** submitted for the M.Tech Degree is entirely my original work and all ideas and references have been duly acknowledged. It does not contain any work for the award of any other degree or diploma.

Date: ____

INVESTIGATOR
Seema
Regn. No.: 41200390
M.Tech CSE

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER-1

INTRODUCTION

1.1 Background

Any ordinary language which is used by humans to communicate with each other is know as Natural language processing. It is having two forms: written form and spoken form. Natural Language Processing comes into picture in the early 1970’s when a 20 years old psychology student Richard Bandler and an associate professor of linguistics John Grinder work together at university of California. They work together on Neuro-linguistic Programming. They claim that there is a connection between neurological processes, language and behavioural pattern. Artificial field is very vast field. This field is having many complex sub fields. It is very difficult to understand those fields and implementing any new idea over that. similarly ,One of the subfield is natural language processing which is not difficult and complex to understand more over a challenge to implementing new ideas over it. We have some more terms related to natural language processing like

I. **Natural Language Understanding** In natural language processing, we try that the natural language is processed by computer which is possible only if computer understand natural language or machine can read natural language.

II. **Natural Language Generation (NLG)** After reading or understanding the natural language ,it is store in machine or computer in machine language. Now next work is to represent it in again natural language form just like a translator. conversion of machine form language into natural language.

The past ten years have seen fast and significant growth in the commercialization of NLP, NLP is currently a vital technology for business. Two trends are mainly responsible for this: the ascent of Natural Language Engineering (NLE) and statistical NLP. Both are based on formal theories and crudely handcrafted systems and also the weakness of previous approaches. Title means to understand a large article or any text document in two or three words. So that we can get the idea about article without reading it. and after making an idea we can decide to read the whole article or not. Many times , only on he basis of tile we can decide that we should read the thing in detail or not.

1. 2 Natural Language Processing

When the ordinary language is handled with computer is known as Natural language Processing. various theories ,concepts, technologies, models and tools are used to making the processing over it. In these days NLP is the area of interest of many researchers. NLP having subfields for research likewise Information Retrieval (IR) , Text Summarization, Information Extraction, Topic Modelling, Tokenization , Classification, Machine Translation build the semantic and Syntactic Representation.NLP provides the huge database to create robust models, tools and approaches.

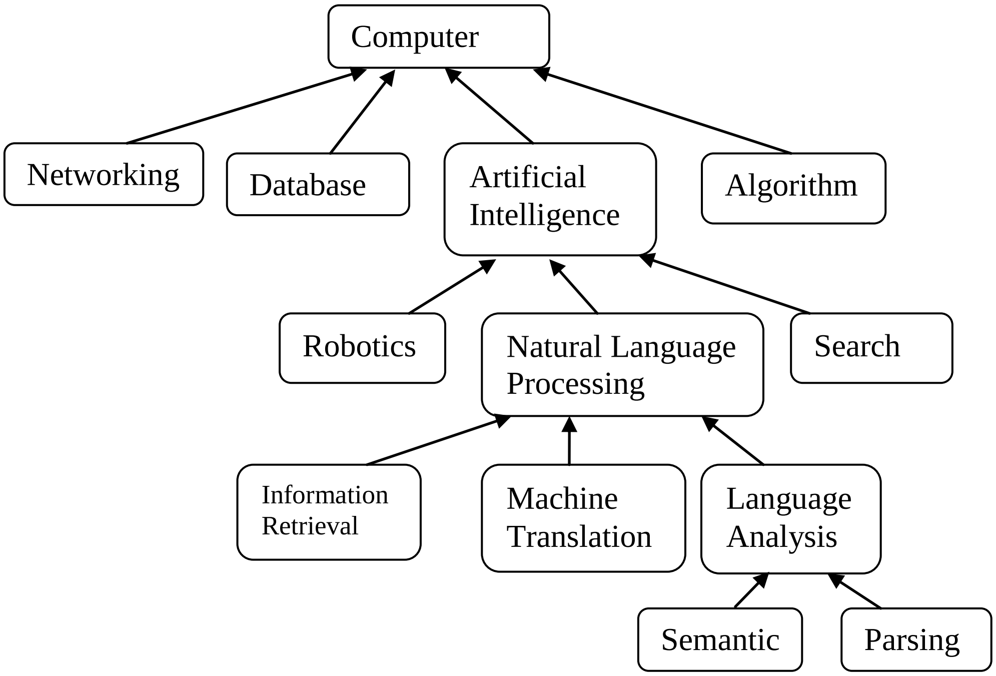


Figure 1: Hierarchal Structure of Language Analysis

1.3 Levels of Language processing

Technically language processing is known as linguistic analysis. Linguistics analysis include all

kind of analysis. There are the various steps of linguistic analysis. There are the various levels likewise sound(Phonology), word formation(Morphology), sentence structure(Syntax) , Meaning(Semantics) and understanding(Pragmatics). All these levels are processed only then the machine can understand the language.

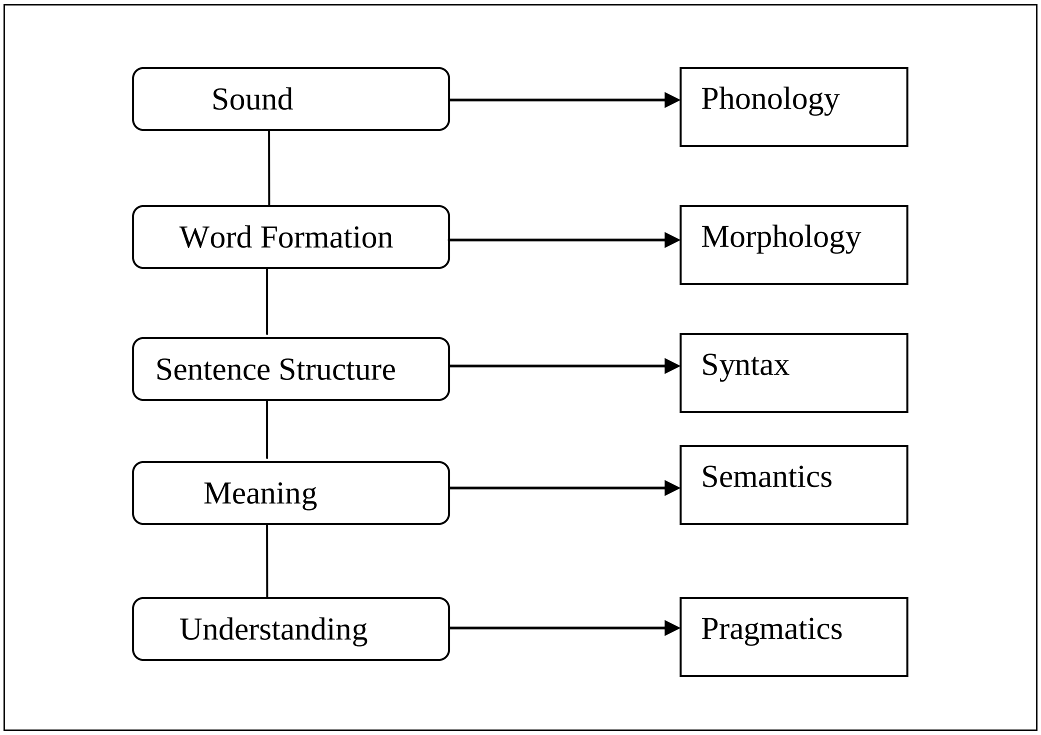


Figure 2: Levels of Language Processing

We will discuss the each level of processing

I. **Phonology:** is the concept of sound. Processing on the sound to make is understandable by machine and store in computer.

II. **Morphology:** is the concept of word of any language. But here the concept is about structure only.

 For example: word “generalization” has two parts. “General” and “tion”.

III. **Lexical Analysis:** is concept of characters. Making a sequence of characters for meaningful string. The sequence of characters is known as tokens. Tokens are nothing but a word.

e.g. Let us consider a sentence, “Prime minister visited nepal”

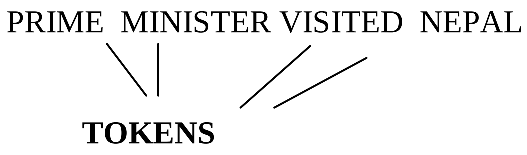


Figure 3: Structure of Sentence

IV. **Syntactic Analysis:** After analysing the word. Next concept is of syntax. Here, syntax means correctness of sentence or rules of sentence or grammar of sentence.

E.g.: The sentence “song rani sing” is wrong grammatically.

V. **Semantic Analysis:** concept of meaning. Grammatical correctness does not make the sentence meaningful.

E.g.: The sentence “ Colourful White Flowers . . .” colourful white does not make sense.

VI. **Discourse Analysis:** concept of dependency. If there are five sentences in a paragraph, all must be dependent or related to each other.

E.g.: Ram has two sons. He is a king.

Here word “He” depends on the previous context.

Pragmatic Analysis: It extracts the knowledge or understanding from the document. It interprets the main theme of document.

1.4 Major Areas in NLP

One of the subfield of AI is natural language processing which is not difficult and complex to understand more over a challenge to implementing new ideas over it.NLP also provides sub fields such as:

i. **Automatic summarization:** generate the summary only with important lines of a document. It is important for very large documents.

ii. **Coreference resolution:** It makes the connection between the two sentences. Each individual sentence is dependable to previous sentence and makes influence of that.corefernce means use the reference of the previous sentence in coming sentences. That reference might be noun , adverb, adjective etc. It builds the relationship between sentences. it implements the dependencies between the sentences.

iii. E.g. "sheela go to delhi by bus. It follows long route", "long route" is a referring by bus.

iv. **Discourse analysis:** concept of dependency. If there are five sentences in a paragraph, all must be dependent or related to each other.

v. E.g.: Ram has two sons. He is a king. Here word “He” depends on the previous context.

vi. **Machine translation:** process of translation of one language to another just like a translator. Both the sub fields, natural language generation and natural language understanding. NLU includes machine reading of natural language and NLG includes language representation. This process includes the lexemes, grammar, semantics and facts of different languages.

vii. **Morphological segmentation:** deals with the formation of the word. It divide the words into different parts. As the structure of the language is difficult morphological segmentation become complex. English language has a simple morphology. more over this language is having limited forms of a single word likewise "run", "runs" ,"running", "runner" etc. but some language like Manipuri having thousands of possible forms of single word.

viii. **Named entity recognition (NER):** identify the named entities from the document. Here name refer to name of person, place organization. Named entities identification is required because capitalization is performed in some languages. Likewise in English language capitalization is performed on named entities. Moreover, in English language the first character of the sentence also capitalizes. Other languages like Hindi, Urdu, and Spanish do not have capital characters and do not perform capitalization. Named entities reorganization is very dominating area of research. a lot of research work has done in this field. Named entities reorganization includes the lexemes, grammar, semantics and facts of languages.

ix. **Natural language generation:** Conversion of machine readable form data to human readable form of data.

x. **Natural language understanding:** In natural language processing, we try that the natural language is processed by computer which is possible only if computer understand natural language or machine can read natural language. It accept the special notations of natural language then understand it and store it. Conversion of natural language to machine language. Various model, approaches and ontology are available to understand natural language.

xi. **Optical character recognition (OCR):** Determine the corresponding text from the image.

xii. **Part-of-speech tagging:** Determine the various parts of the sentence and mark them with respective tag. A single word can have different tags. As that particular word used in different context. For example, "bank" is a noun ("He set on the bank of the river") or verb ("today online banking is possible"); other example is "mouse" is a noun. Mouse is living being. Mouse , is input device of computer. Other word like watch. As a noun watch is shows the time of a day. But as a verb watch means to look at something. So, the word is same but by using in the different contexts its meaning gets changed.

xiii. **Parsing:** Perform the grammatical analysis of a given sentence. In natural languages, grammar is uncertain and multiple analysis on a single sentence. with different parsers grammatical analysis perform differently.

xiv. **Question answering:** Determine the answer of given a human-language question. some of them has direct single word answer likewise what is your father's name. But some of them are descriptive answers likewise how was your holidays.

xv. **Relationship extraction:** extract the relationship between the named entities present in the sentence. E.g. Ram is husband of sita.

xvi. **Sentence breaking:** From the given a text, find the sentence boundaries. Periods or other punctuation marks are use mark sentence boundaries.

xvii. **Sentiment analysis:** identify the sentiments from the given text like happy, sad, agree, satisfy ,good ,bad, confuse, tired. when the online reviews are held then sentiment analysis is performed over there.

xviii. **Speech recognition:** identifying the speech from the text. In other simplest way we can say that extract the textual information from speech or sound. Sound provides the information that can be store in textual form.

xix. **Speech segmentation:** Separate the words from the sound clip. It is a sub part if speech recognition process.

xx. **Topic segmentation and recognition:** separate the segments of the text of document then find the topic of the segment.

xxi. **Word segmentation:** Divide passage into different tokens called words. This is possible because space is available as a separator of words in a language. But not all the language have separation symbol like Jananese. For this kind of languages word segmentation become a challenging task.

xxii. **Word sense disambiguation:** Words having different meaning with respect to context
in which words are using. word like watch. As a noun watch is shows the time of a day. But as verb watch means to look at something. So, the word is same but by using in the different contexts its meaning gets changed.

xxiii. **Information retrieval (IR):** concept of finding, retrieving and storing textual information. Individual retrieval is dominating area in online world of information.

xxiv. **Information extraction (IE):** concept of meaningful information finding. For this firstly perform the information retrieval. It includes the lexemes, grammar, semantics and facts of languages, named entities reorganization.

xxv. **Speech processing:** process the speech and converts text-to-speech. So text can be speak by machine. Research work on speech is a vast area in itself. It is processed in various levels. . It accept the special notations of natural language then understand it and store it.

1.5 Three Major Aspects of Natural Language Understanding Theory:

I. **Syntax:** Generally syntax means the structure. But in natural language processing syntax means the rules and grammar of any language. That rules are required to follow by all the sentences and dictionary of that language.

II. **Semantics:** semantic means the meaning. the meaningful sentences are semantic and the sentences that do not have meaning called non-semantic. Grammatical correctness does not make the sentence meaningful. it verifies the structure only. it does not tell something informative or meaningful. E.g.: The sentence “ Colorful White Flowers . . .” colorful white does not make sense.

III. **Pragmatics:** After getting the meaning from semantic analysis. Now need to check the feasibility of meaning. Existence in real world or not.

1.6 Part of speech

In English language, words are classified or categorized into different parts such as noun ,pronoun ,adjective etc is known as part of speech. the differ parts are called tags. the tags collectively called tagset. A Single word can be classified as both noun and verb. More over, There are some derivations on words can change the its part of speech.

E.g. The word “function” can be used in different part of speech.

We can see the **functioning** of car. (verb)

The **function** of prime minister will hold there. (noun)

This heater can **functional** even at very low temperature. (adjective)

Now, computers are **functionally** built to save electricity. (adverb)

There are the various parts of speech for a single word

Philosophy (noun)

Philosopher (noun)

Philosophize (verb)

Philosophical (adjective)

Philosophically (adverb)

1.7 Text mining

Text mining is the technique for information extraction from textual data. Data mining & knowledge discovery in database(KDD) technique processed the structured database. The text mining include all the specialized technique to operate textual data & extract the information from collection of texts. A document is a piece of text & we can perform a lot of analysis on the text like word stemming(removing suffix) ,finding lexemes , phrase matching synonyms normalization, tagging ,getting senses ,anaphora resolution & role determination.

1.8 Lexical resources

To perform the processing on the natural language a huge database is required which is a linguistic database or corpus.

1.9 Existing approaches for title generation

Two approaches toward title generation are: Text summarization based approach, Statistical based approach

1.9.1 Text summarization based approach

I. **Single document Vs Multiple document**: System will summarize a single document then it can be used to summarize several documents.

II. **Exact Vs abstract** : summary can be form by extracting particular text units from document.

III. **Generic Vs user Focus**: Generic summary is based on whole of the text. While user focus is based on the users focused part of document.

IV. **Indicative Vs informative:** indicative summary shows the addressing topics in document and informative is based on the main concept of the document.

1.9.2 Statistical based approach

In recent, for title generation statistical approach is used. In statistical approach, statistical learning algorithms are applied to automatically generate titles for documents. Moreover, the training corpus is used to generate the appropriate title. This train corpus is used for tagging. The most general concept on this approach is there the large database of document-title pairs. Map the relation between title and document. and titles and then will apply the statistical model to generate titles for unseen documents title generation process is divide into two firstly getting the main words and then organize in proper sequence.

1.10 Part of Speech tagging:

1.10.1 First order Hidden Markov Model

Hidden Markov Model (HMM) is a based on statistical approach. This model is visualized as set of states which are interlocked. These set of states are interchanges with each other. With the set of states, one can move to another state. This is called transition between the states. In an HMM, the actual sequence of change of state is unknown or hidden. So, the name of the model is hidden markov model. Let us suppose, w0,w1,w2 are three states of the system. There are probability to change the states of the system from w0 to w0,w0 to w1,w0 to w2 & so on. Markov process is like WFSA(Weighted Finite State Automata)

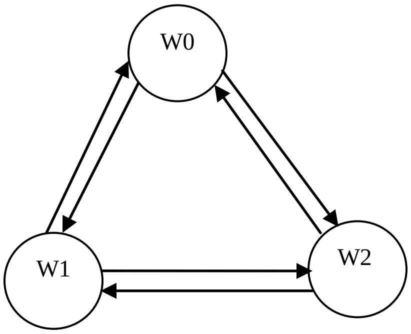


Figure 4: Transition of state of System

I. N is used for tagging the system. Tag for all the state of system.

II. A= {a_{kl}},The state transition probability distribution.

III. The probability a_{kl} is the probability that the process will move from the state k to state l.

$$A = \begin{bmatrix} a_{00} & a_{01} & a_{02} \\ a_{10} & a_{11} & a_{12} \\ a_{20} & a_{21} & a_{22} \end{bmatrix}$$

Figure 5: Probability of Transition

1.10.2 Second order Hidden Markov Model In second order markov model, one more element comes into picture i.e. emitted values or output symbols. when there is change in state then their must be emission of some values called output symbol.

I. M is no. of distinct output symbols.

II. w_0,w_1,w_2 are various state of the process. v_0,v_1 and v_2 are three emitted values.

III. B, The observational symbol probability distribution $b_j(k)$.

$$B = \begin{bmatrix} b_{00} & b_{01} & b_{02} \\ b_{10} & b_{11} & b_{12} \\ b_{20} & b_{21} & b_{22} \end{bmatrix}$$

Figure 6: Probability of Transition

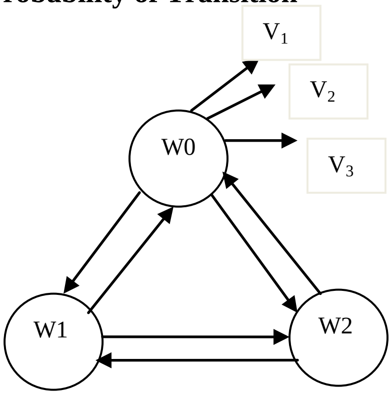


Figure 7: State of the System

Hidden Markov Model for part of Speech Tagging

I. N is used for tagging the system. Tag for all the state of system means all the words.

II. M is the output symbol or emitted value. For the part of speech tagging ,M is the no. of words in the lexicon.

III. For part of speech, Transition probability distribution means model will move from tag t_k to $t_{l.}$ This probability can generate the trained corpus.

IV. For part of speech, B is observational symbol probability distribution is the probability that the word wk will be emitted when system is at tag $t_{l.}$

V. Pi is the initial state distribution, For part of speech, The probability that the sentence will begin with tag t_k

1.10.3 Rule-based Tagging

In 1992, Brill introduced a POS Tagger which is based on rules. There is a trained corpus is available to the system. From the trained corpus system can drive lexical & contextual information.

1.10.3.1 Transformation based error-driven learning (TEL)

Brill's corpus based on Transformation-based error driven learning approach. In the figure, The input text is an unannoted corpus which corpus which passes to the first state annotator. It apply some tagging. The output produced by initial state tagger is temporary corpus. Now it is compared with goal corpus which has been manually tagged. Temporary corpus passed through the next stage that is of Learner. Learner replaces the temporary corpus with analysis. The tagger uses TEL two times: Once for lexical analysis for tagging unknown words and once for conceptual rules. Learning lexical rules:

P(T/W):= Freq(W,T)/Freq(W)

P(T/W) Most likely tag for word.

Freq(W,T) Freq. of word with specific tag.

Freq(W) Freq of word in annoted corpus

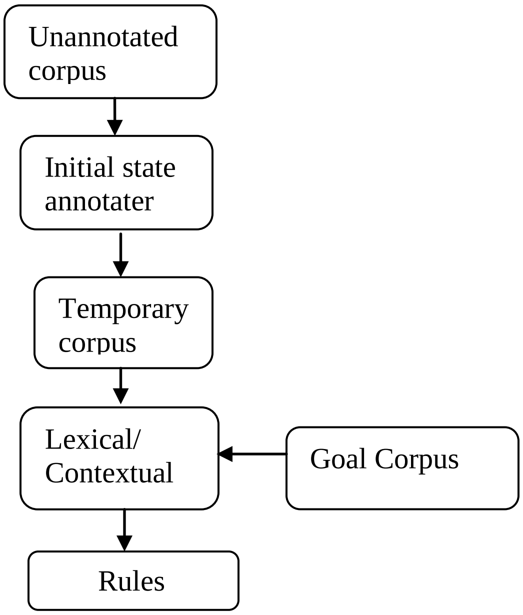


Figure 8: Rule Based Tagging Steps

CHAPTER-2

REVIEW OF LITERATURE

Marie-Catherine de Marne_e et.al [5] they describe the structure of the sentence and stanford university provides us the Stanford CoreNLP tool. It can perform the tokenization ,generation of parse tree. Parse tree is very useful in natural language understanding process. with the help of the parse tree machine readable form can easily generated which is helpful in natural language understanding. Java language is used for develop this tool. firstly it load the text file. then it load the parser. The parser given by standford i.e. edu.stanford.nlp.parser.ui.parser. Stanford university has given a great contribution in the field of natural language processing. Tool is available at link[22]

Apache OpenNLP Development Community[6] this community give a great contribution in the are Natural language processing. they provide a toolkit that can perform many tasks over text. such as separation of words from sentences. POS tagging, Find name etc. Tool can perform its work with the help of models. some of the model files are not found or my be not available. A different models are available for different tasks over the sentence. This tool is developed in dot net technologies. Tool is available at link [23]

Hamish Cunningham et.al[7] they describe the working of GATE tool. It is general architecture of text analysis. Various analysis are performed over the text.

GATE can perform the various task:

I. It is based on the Named entities reorganization. It identifies the name of person, address, states ,countries. There is corpus is available for the tool.
II. the tools and components can easily extendible.
III. The tools and its documentation is freely available.Its documentation, tutorials and online demonstrations are available on its Web site.

Willian R. Hersh et.al [8] they describe the decision support system. The system is Electronic medical record (EMR)System is a decision support system based on the medical corpus of 842 MB . Medical corpus includes all the reports related to patients, medicines, doctor prescriptions, discharging report. they describe the impotence of corpus for any decision making systems. Corpus must be consist and must be updated time to time. Only then it can lead to decision support system.

R.Jin et.al [9] describe a new probabilistic approach for title generation. Rong jin perform the probabilistic theories on text to extract the title. In very general terms they divide the process in two parts: finding the appropriate words and prepare a sequence. The words are extracted from documents and store in 'information source". "Information source" is like a temporary storage of analysed document. Now on the basis of these extracted words. Title can be generated. Now the order or sequence of these words is also very important step. It is like a pragmatic analysis.

Cedric Lopez,Violaine Prince et.al[10] they describe the survey on the titles. in title percentage of noun, percentage of adjective, percentage of adverb etc. In all kind of document, in 90% cases noun is the title. This is the highest percentage. In the paper they provide the percentage of all categories.

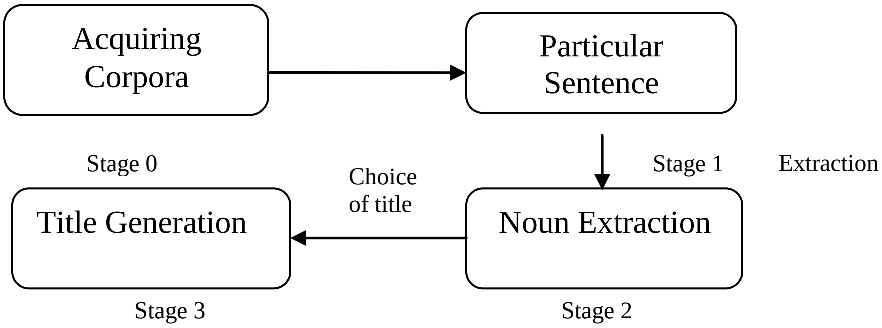


Figure 9: Steps of Text Mining Technique

Step 0 Acquiring Corpus: Find the large corpus

Step1 Particular Sentence Determination: probably, first and second sentence is candidate for title.

Step 2 Noun Extraction: find the nouns from document

Step 3 Title Generation: select the title from extracted candidate noun phrases.

More is the noun candidate leads to the quality improvement of the title.This is implemented for french language.

Paul E. Kennedy et.al [11] describes the non-extractive approach. The title can be generated without extracting the word from the document. In the model , the title and the document is consider as “the bag of words” Prepare a title vocabulary and a document vocabulary. Now, perform the estimation of probability of document word appear in the given document & title word appear in the corresponding title. This model consists the list of document word and tile word with assign the probabilities. They explain the EM (Estimate & Maximize) algorithm. They trained a word-pair model P(dw|tw) for 3 iterations with the corpus of 40000 transcripts of broadcast-news stories with human-assigned titles. They also build the language model. Extractive summarization is the most popular approach to generate titles.

M. Rajman et.al[12] describe the Text mining techniques which is useful to summarize the document. Data mining and text mining both are used for information extraction the only difference about the data on which these techniques work. Data mining works over structure database and text mining works over unstructured database. They explain two different TM works: information retrieval from indexed document and providing information question answering scenarios.

Yi Guo et.al [13] explain the sentence understanding stages such as sentence parsing and semantic processing. parsing is consider as a basic level but semantic understanding involves lexical and higher discourse integration. After that different models of semantic and syntactic behaviour of sentence are explained. Different models are brain based models, garden path models, syntax first & integrative module, working memory and semantic memory. CIParsing and SMCI (Sentence meaning construction index) explain the four dimensions of the sentence lexical, syntactic , grammatical and semantic dimension.

Fco. Mario Barcala et.al.[14] explain complex linguistic phenomena of proper noun reorganization. They explain the effectiveness of several methods. They show the result of several experiment perform to analyze the strategy of recognize proper noun. They propose a technique based on indexing . There are two sub parts: Proper noun trainer and Proper noun identifier. Proper noun trainer sub module use the trained dictionary to set the candidate proper nouns. It identifies the words begin with capital letter and its non-ambiguous position. These words include in the dictionary which is further used by next sub module. It also identifies sequences of capitalized words check that connectives are valid like the preposition of and definite articles. All possible segmentations of these sequences are measured. Based on the trained dictionary which is built in previous model proper noun identifier extracts the proper noun. Then it detects the proper noun whether simple and compound and position of that.

Batuer Aisha et.al.[15] describe Morpheme analysis for Uyghur language processing. Uyghur language belongs to Altaic language family. It is based on Arabic script which consists of 32 letters, 8 vowels and 24 consonants. Each letter in the language have different shape at the beginning, middle, and end of a word. Uyghur is written from right to left in text like urdu language and words splits by a blank space in sentences like hindi and English languages. The existing Uyghur word morpheme analysis methods are rule based. These methods work on statistical based analysis but give some problems during experiments.

Sara Stymne [16] explains the tree bank corpora in the which each sentence has been annotated with syntactic analysis. Generally penntree bank(PTB) is used. PTB developed at university of Pennsylvania in 1999.There are different PTB labels are used such as NNP for Proper noun, CD for Cardinal number, NNS for Noun, plural, JJ for Adjective, MD for Modal ,VB for verb, base form , DT for Determiner , NN for Noun, singular , IN for Preposition ,S for Declarative clause ,NP for Noun phrase , ADJP for Adjective phrase ,VP for Verb phrase ,PP for Prepositional, ADVP for Adverb phrase, RRC for Reduced relative ,WHNP for *Wh*-noun phrase ,NAC for Not a constituent. Treebank grammar has a nested structure. Its grammar can be transformed in various ways. It consists 29,846 rules.

Xinyan Xiao et.al.[17] they join the tokenization model and decoding phase. Tokenization and decoding both are the steps of machine translation. In simple machine translation (SMT) tokenization is first step. on the basis of input string SMT divided in two types: string based system that take string as input and tree based system take tree as a input. Basically they consider the Chinese language for tokenization which is the challenging task for tokenization because Chinese language has no space between the words in the sentence. So, it is difficult to extract the tokens. For these kind of language ,errors are produced in SMT systems. So, they propose tokenization and translation.

Maria Soledad Pera et.al[18] they introduce the approach of CorSum for summarizing the document. It is based on word similarity. Moreover, word correction factor also computed it can summarize multilingual document. It improves the searching process on the web. The naive bayes classifier is trained with it. It captures the main part of the document. Firstly, Two algorithms are used to set the ranking. The sentences with higher rank are including in the summery. Working of CorSum is also based on the ranking. It selects the most representative sentences from the document D & that is the summery of the document D. The performance and the output quality is verified by Document Understanding Conference (DUC) 2002 dataset. DUC-2002 includes 533 news articles out of which approximately 10 from from popular news collections such as the Wall Street Journal, AP Newswire, Financial Times, and LA Times.

Lucas Antiqueira et.al.[19] They work on the concept of complex network or graph of text that represent the nodes corresponding to the sentences & those nodes are connected by common nouns. This network detects the text feature. There are the various strategies used in CN-Summ are used named as Degree Strategies, Shortest Path Strategies , Locality Index Strategy , d-Ring Strategies, k-Core Strategies ,w-Cut Strategies , Community Strategy , A Voting Strategy.

H. P. Luhn [20] design the auto-abreact approach. Text is scanned by IBM-704 data process machine in machine readable form. Then individual word frequency is analyzed. Measure the significance for individual word & sentence also. The sentences having highest score are extracted & consider as "auto-abstract".

Hiroshi ISHII et.al.[21] They present the word importance approach. They consider the noun as a most important element of the sentence. Moreover, its position is also very important. The noun present at subject position is more important then noun present at object position. This approach is based on four steps. First, assign the score to the nouns in each sentence. Second, based on the score find the importance values of each sentence. Third, Higher importance valued sentences are selected. Finally perform the coherency test which is final decision step for which sentences are considered in the summery.

CHAPTER-3

PRESENT WORK

3.1 Problem Formulation

Artificial field is very vast field. This field is having many complex sub fields. It is very difficult to understand those fields and implementing any new idea over that. Similarly, One of the subfield is natural language processing which is not difficult and complex to understand more over a challenge to implementing new ideas over it.Title means to understand a large article or any text document in two or three words. So that we can get the idea about article without reading it and after making an idea we can decide to read the whole article or not. Many times, only on he basis of tile we can decide that we should read the thing in detail or not. Almost all the expert systems based on knowledge base and inference engine [1]. For the language processing, python I is the powerful language. Python provides a toolkit to process the language. This toolkit provides the good result [2][27]. The training corpus is the most important element for language processing [3]. AI is the challenge to build computational models and approaches of cognitive processes [4]. To study the books and other research papers I fond that title generation is the dominant area of research in language processing. It follows the Lexical analysis, Part of Speech tagging ,Discourse Analysis, Frequencies of Tokens, Prioritize the Frequencies Various approaches are given below approaches for NLP[26]

I. **Symbolic Approach:** concept is related to natural language generation i.e. representation of knowledge
II. **Statistical Approach:** concept of mathematical techniques. Probabilistic approaches are implemented.
III. **Connectionist Approach:** combination of above two approaches i.e. using the statistical approach represent the knowledge.

Wordnet is a online resource for database. It performs the tokenization and tagging as well. It is freely available resource[25].

3.2 Objective

objective is the task or group of tasks that are involved in our research work or any work. The objective of my work is to get idea about the document without reading it. The major objective is to suggest the titles for document. some kind of compositional semantic tool that perform the text analysis to generate the titles. compositional semantic means analysis performed over number of sentences. compositional analysis can perform analysis over all the sentences. The major objective includes the following:

I. Lexical analysis
II. Part of Speech tagging
III. Discourse Analysis
IV. Frequencies of lexemes
V. Various analysis on frequencies and texts

3.3 Methodology

Our methodology is based on the parts of sentences. Firstly we need to separate the each word from sentence and find its type or tagging. For this I utilize the HMM Tagger develop in VC# in dot net technology.

3.3.1 Research design

In the text analysis the primary step is to perform lexical analysis of the document. It is following Part-of-speech analysis, discourse analysis. There are the following phases are need to follow.

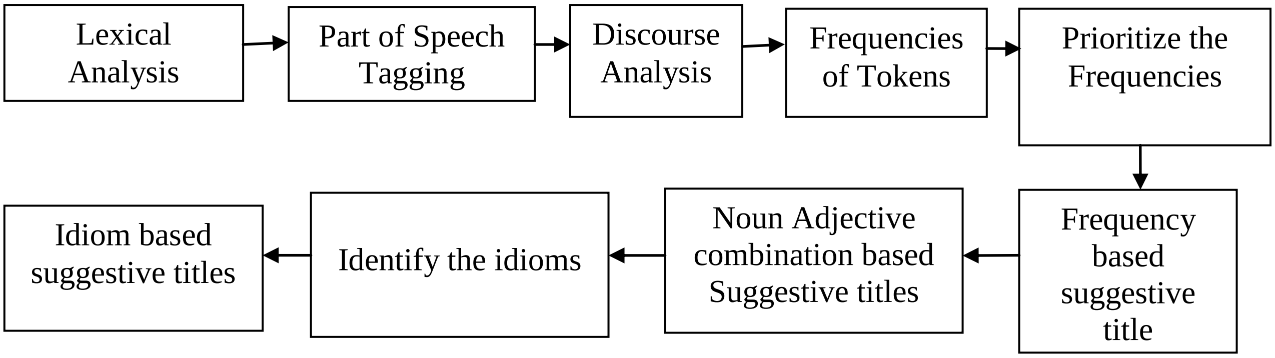

Lexical Analysis
Part of Speech Tagging
Discourse Analysis
Frequencies of Tokens
Prioritize the Frequencies
Frequency based suggestive title
Noun Adjective combination based Suggestive titles
Identify the idioms
Idiom based suggestive titles

Fig 10: Diagrammatic Representation of Various Phases.

I used the .NET technology to implement the idea. This technology provides the robust programming tools and framework. Most of the tools in these days are develop in this technology. It supports the multiple programming languages and support the other languages. It supports the intellisense. Moreover the HMM tagger that is I use for tagging also develop in .NET technology with graphical user interface. It loads the corpus first. It has a large corpus. Then it implements the add-one smoothing over the corpus for more accurate results. After that it tag the text. It also counts the words.

3.3.2 Proposed solution

I. **Lexical Analysis:** is concept of characters. Making a sequence of characters for meaningful string. The sequence of characters is known as tokens. Tokens are nothing but a word.

   I. The text is entered in the Rich Text Box. Sequence of characters are picked and get the word to perform the further text analysis.

   II. “ ”(space) between the two tokens is considered as the completion of single word.

   III. “.”, ”?” and ”!” characters are considered as the completion of the sentence.

   IV. Other Punctuation marks likewise(‘, ””” , ; ,) are also consider.No. of sentences/words counted.

Algorithm: This is the algorithm that is used for lexical analysis. The whole text is scanned by character by character and there is the formation of the word. There are various algorithms are designed for lexical analysis. But this is the most general algorithm. The white space or punctuation mark is consider as the separator in English language.

```
Lex(c, text, word, TextBox)
c is to store the character
text is string of characters
word is used to store the word
TextBox is used show the result of processing

   I.   Ws := white space, sep := separator, punch:= punctuation mark
  II.   Foreach(char c in text)
```

```
III.  If (c =ws or sep or punch)
IV.   then
V.    TextBox := word
VI.   word := ”” 
VII.  Else
VIII. word := word + c
IX.   End if
X.    End foreach
```

II. **Part of Speech Tagging:** In English language, words are classified or categorized into different parts such as noun ,pronoun ,adjective etc is known as part of speech. the differ parts are called tags. the tags collectively called tagset. A Single word can be classified as both noun and verb. More over, There are some derivations on words can change the its part of speech.

E.g. The word “function” can be used in different part of speech.
We can see the functioning of car. (verb)
The function of prime minister will hold there. (noun)

Tags are the labels that are assign to the words such as Ram is good boy, Now the words of the sentence are tagged as Ram/nn is /vb good/adj boy/nn. So here, nn for noun, vb for verb, adj for adjective. All the tags are collectively known as tagset. Part-of-speech tagging can be writing like POST.

Part-Of-Speech Tagger is a software tool that perform the lexical analysis and assign tags to each lexeme likewise noun, pronoun ,verb etc. In these days no. of POS Taggers are available named as Stanford Tagger, OpenNLP Tagger, HMM Tagger, NLTK(natural language toolkit). To implement idea we uses HMM Tagger designed in c# language. It is based on HMM approach. Hidden Markov models are especially known for their application in worldly pattern recognition such as part-of-speech tagging, speech, handwriting, gesture detection, musical keep score following and bioinformatics .It provides the robust tool for part-of-speech tagging. It uses the Brown corpus. It is the first major corpus of English the Brown Corpus. It is a linguistic corpus which is tied with part-of-speech tagging. It is developed in 1960, by Henry Kucera and W.

Nelson Francis, at Brown University. It consists of about 1,000,000 words of English language. the HMM tagger that is I use for tagging also develop in .NET technology with graphical user interface. It loads the corpus first. It has a large corpus. Then it implements the add-one smoothing over the corpus for more accurate results. Smoothing filters the results .it makes the corpus more robust and increases the accuracy. After that it tag the text. It also counts the words.

Table 1:Brown Corpus Tag Set

Tag	**Definition**
NN	singular or mass noun
NN$	possessive singular noun
NNS	plural noun
JJ	Adjective
RB	Adverb
VB	verb, base form
VBD	verb, past tense
VBG	verb, present participle/gerund
VBN	verb, past participle
VBP	verb, non 3rd person, singular, present
PRP	Personal pronoun

RBR	comparative adverb

III. **Discourse Analysis:** Converting all pronoun into equivalent noun. It makes the connection between the two sentences. Each individual sentence is dependable to previous sentence and make influence of that.

E.g. Ram is boy. He is king.

After implementing the discourse analysis , the pronoun he is changes to Ram. Ram is boy. Ram is king.

If we see the structure of the sentence we will find that noun is the most important part of sentence. It is like a root of the sentence. Next important thing about a noun is the position of noun. It matters a lot.

I. Ram is reading a book.

II. Book is read by Ram.

There are the two nouns Ram and book in both the sentences. In I, Ram is more impotent then Book. But in II, book is more important then Ram. So, noun present at subject potion is more impotent then noun present at object position. In our approach for the sentence, noun present in subject position is called primary noun & at object position is called secondary noun.

E.g. Ram is reading a book.

Here ram is primary noun and book is secondary noun.

There are various cases are consider. There are the various cases

Case I Single Noun in the first sentence.

Ram is eating. He is good boy.

There is a single noun (Ram) in the first sentence. The pronoun (He) in the second sentence must be replaced by noun(Ram).After implementing the discourse over the sentence. Ram is eating. Ram is good boy.

Case II More than one noun in the first sentence.

Ram is a king. He lives in ayodhaya.

There are two nouns (Ram and King) in first sentence. The pronoun (He) in the second sentence must be replaced by First noun (Ram) not by second noun (King).

For the more than two nouns in the sentence, First noun is **primary noun** and other nouns are the **secondary nouns**. In above example, Ram is the primary noun and king is the secondary noun.

Ram is a king. Ram lives in ayodhaya.

Case III Conjunction between the nouns

Ram and Seeta are good friends. They play games.

If there is a conjunction (and ,or) between the nouns then pronoun is replaced by both the nouns with conjunction.

Ram and Seeta are good friends. Ram and Seeta play games.

Case IV Paragraph start with pronoun

He is good boy. He plays games.

If the paragraph start with pronoun then no need to perform discourse over text.

IV. **Frequency of Tokens:** The noun and adjective both are the most important in sentence structure. We need to calculate the occurrences of both elements.

V. **Suggestion of Title :** The title will be suggested by three methods.

I. **Frequency based suggestive title:** [10] the title word often found in the first sentence of the text. [21] nouns present at subject position is very important in a sentence. It is consider as the root of the sentence. So, as the noun is very important the highest frequency noun can be suggested as a title.

II. **Noun Adjective combination based Suggestive titles:** Noun is considered as the first important element in the structure of sentence. After the noun, Adjective is also very important element. The combinations of noun and adjective are suggest as a title.

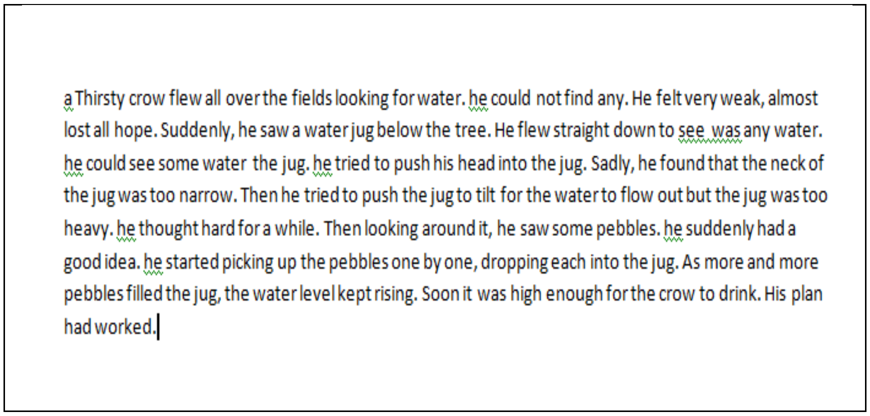
III. **Idiom based suggestive titles:** Idiom or phrase present in the document is always be a appropriate title for the document.

CHAPTER-4
RESULTS AND DISCUSSIONS

We obtain very good results from the implementation of our program based on the above phases of tile generation. Compositional Semantics Tool for Automatic Title Generation was reliably performed various analysis on text to provide the results. The tool was tested for different documents of English language. Results obtained were satisfactory for text of English language.

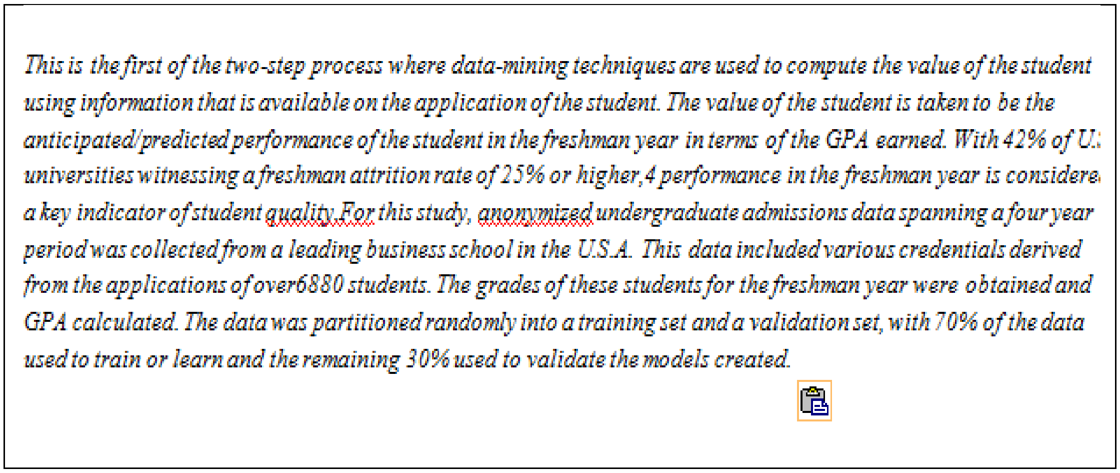
The lexical analysis and part of tagging are performed on the test with the help of HMM tagger.

After that the discourse analysis is performed on the text. Then the frequency analysis and noun/adjective combination analysis performed. Moreover finding the phrase from the text or finding the relevant token of phrase/idiom. Sometimes no idiom or phrase occurs in he text document so phrase based title does not show result. This is not a serious limitation. Below, we show the results of performing analysis for text using our compositional semantic tool.



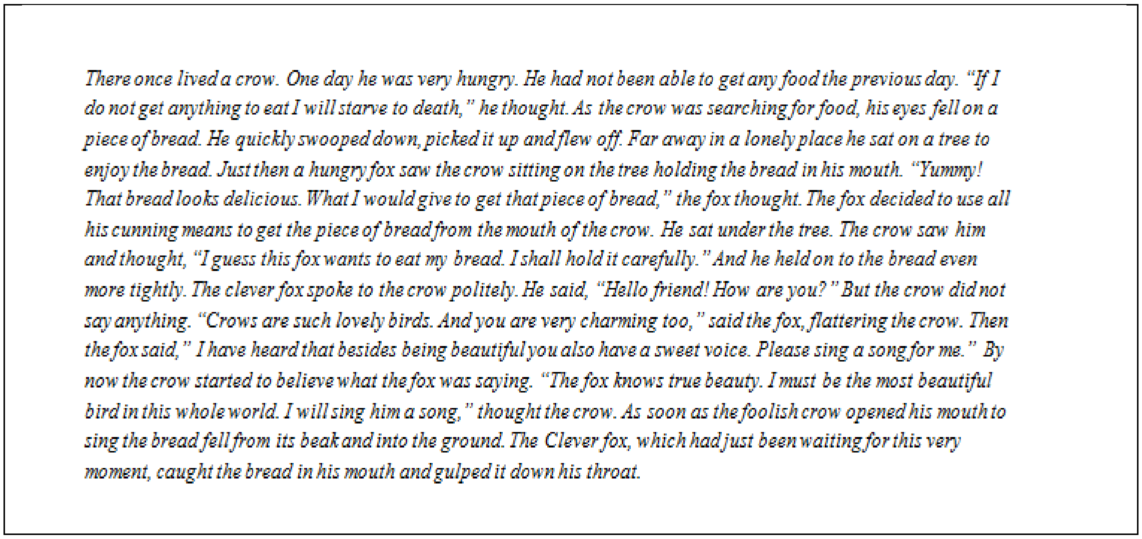
a Thirsty crow flew all over the fields looking for water. he could not find any. He felt very weak, almost lost all hope. Suddenly, he saw a water jug below the tree. He flew straight down to see was any water. he could see some water the jug. he tried to push his head into the jug. Sadly, he found that the neck of the jug was too narrow. Then he tried to push the jug to tilt for the water to flow out but the jug was too heavy. he thought hard for a while. Then looking around it, he saw some pebbles. he suddenly had a good idea. he started picking up the pebbles one by one, dropping each into the jug. As more and more pebbles filled the jug, the water level kept rising. Soon it was high enough for the crow to drink. His plan had worked.

Figure 11: Story1



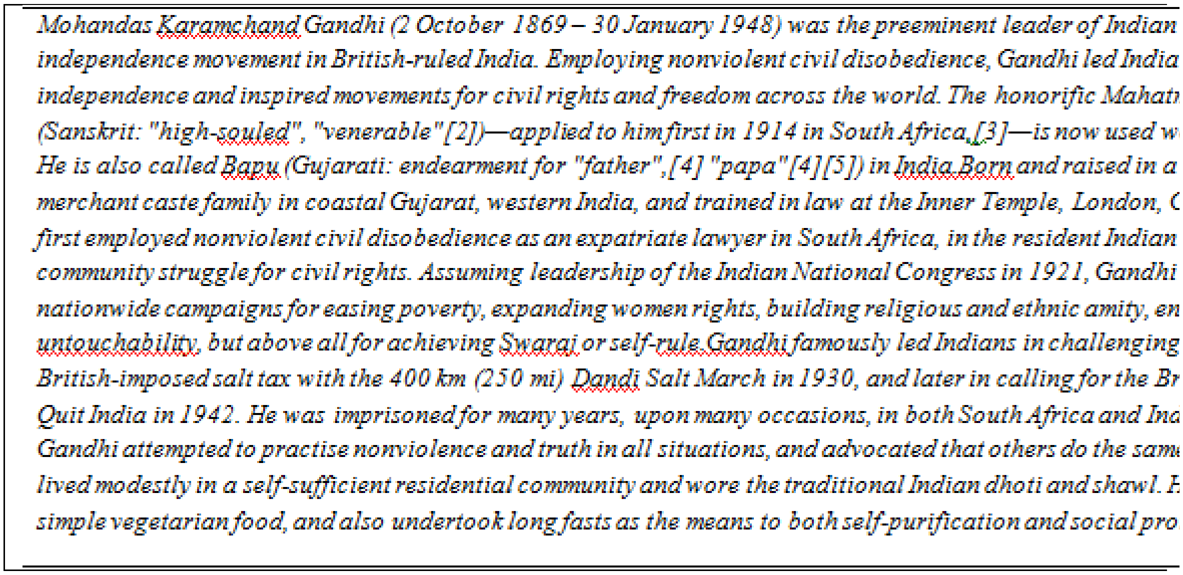
This is the first of the two-step process where data-mining techniques are used to compute the value of the student using information that is available on the application of the student. The value of the student is taken to be the anticipated/predicted performance of the student in the freshman year in terms of the GPA earned. With 42% of U. universities witnessing a freshman attrition rate of 25% or higher,4 performance in the freshman year is considere a key indicator of student quality.For this study, anonymized undergraduate admissions data spanning a four year period was collected from a leading business school in the U.S.A. This data included various credentials derived from the applications of over6880 students. The grades of these students for the freshman year were obtained and GPA calculated. The data was partitioned randomly into a training set and a validation set, with 70% of the data used to train or learn and the remaining 30% used to validate the models created.

Figure 12: Story 2



There once lived a crow. One day he was very hungry. He had not been able to get any food the previous day. "If I do not get anything to eat I will starve to death," he thought. As the crow was searching for food, his eyes fell on a piece of bread. He quickly swooped down, picked it up and flew off. Far away in a lonely place he sat on a tree to enjoy the bread. Just then a hungry fox saw the crow sitting on the tree holding the bread in his mouth. "Yummy! That bread looks delicious. What I would give to get that piece of bread," the fox thought. The fox decided to use all his cunning means to get the piece of bread from the mouth of the crow. He sat under the tree. The crow saw him and thought, "I guess this fox wants to eat my bread. I shall hold it carefully." And he held on to the bread even more tightly. The clever fox spoke to the crow politely. He said, "Hello friend! How are you?" But the crow did not say anything. "Crows are such lovely birds. And you are very charming too," said the fox, flattering the crow. Then the fox said," I have heard that besides being beautiful you also have a sweet voice. Please sing a song for me." By now the crow started to believe what the fox was saying. "The fox knows true beauty. I must be the most beautiful bird in this whole world. I will sing him a song," thought the crow. As soon as the foolish crow opened his mouth to sing the bread fell from its beak and into the ground. The Clever fox, which had just been waiting for this very moment, caught the bread in his mouth and gulped it down his throat.

Figure 13: Story 3

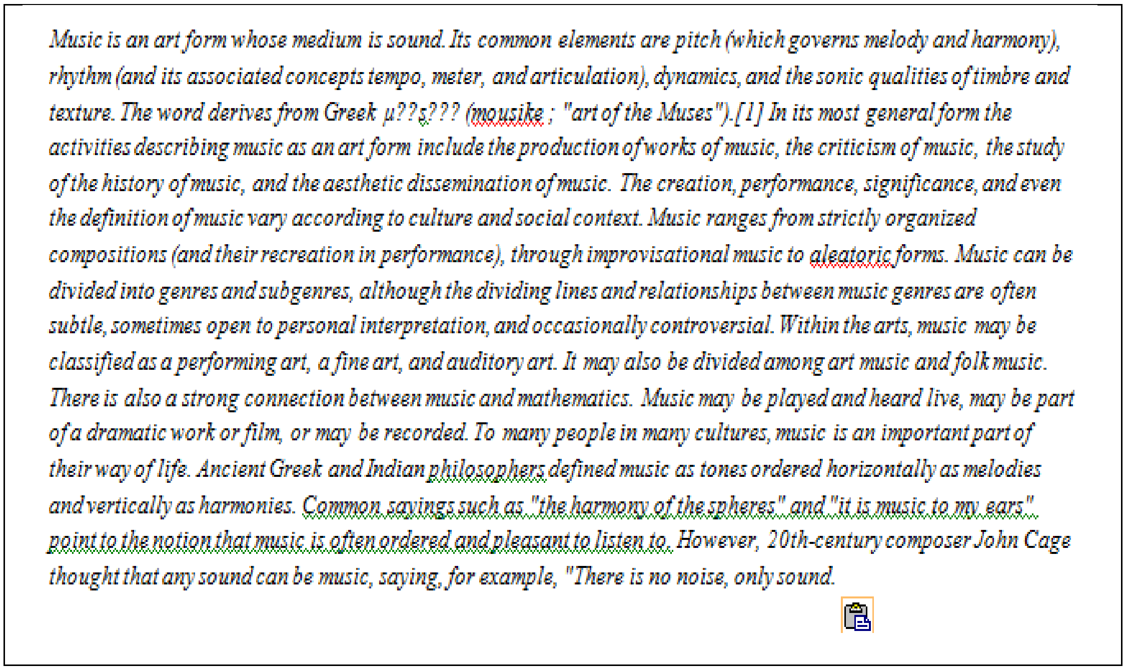


Mohandas Karamchand Gandhi (2 October 1869 – 30 January 1948) was the preeminent leader of Indian independence movement in British-ruled India. Employing nonviolent civil disobedience, Gandhi led India independence and inspired movements for civil rights and freedom across the world. The honorific Mahat (Sanskrit: "high-souled", "venerable"[2])—applied to him first in 1914 in South Africa,[3]—is now used w He is also called Bapu (Gujarati: endearment for "father",[4] "papa"[4][5]) in India.Born and raised in a merchant caste family in coastal Gujarat, western India, and trained in law at the Inner Temple, London, (first employed nonviolent civil disobedience as an expatriate lawyer in South Africa, in the resident Indian community struggle for civil rights. Assuming leadership of the Indian National Congress in 1921, Gandhi nationwide campaigns for easing poverty, expanding women rights, building religious and ethnic amity, en untouchability, but above all for achieving Swaraj or self-rule.Gandhi famously led Indians in challenging British-imposed salt tax with the 400 km (250 mi) Dandi Salt March in 1930, and later in calling for the Br Quit India in 1942. He was imprisoned for many years, upon many occasions, in both South Africa and Ina Gandhi attempted to practise nonviolence and truth in all situations, and advocated that others do the sam lived modestly in a self-sufficient residential community and wore the traditional Indian dhoti and shawl. F simple vegetarian food, and also undertook long fasts as the means to both self-purification and social pro

Figure 14: Story 4

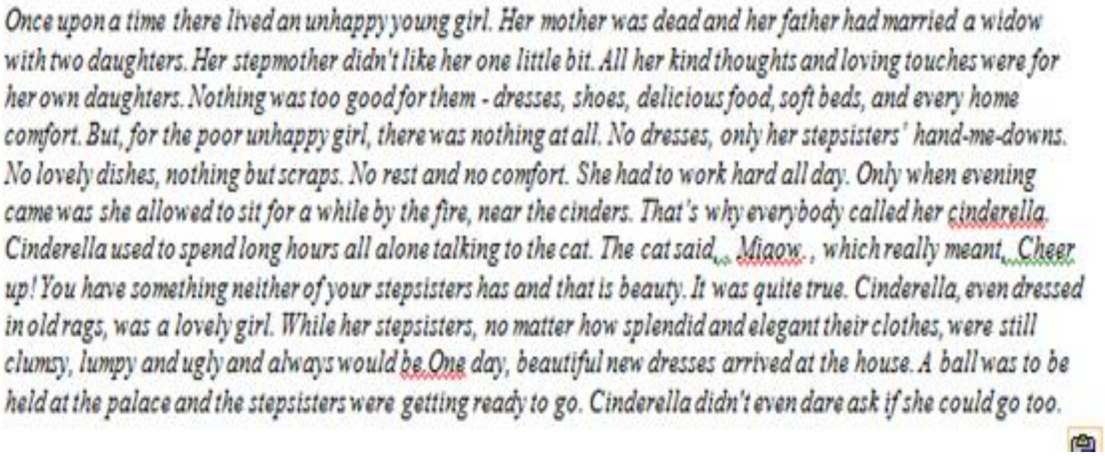
The Prime Minister visited his ailing guru Swami Atmasthananda Maharaj, the president of Ramakrishna Math and Mission Order, in Kolkata on Saturday. Within hours of his first visit to Kolkata after assuming the office of Prime Minister, Narendra Modi on Saturday visited his ailing guru Swami Atmasthananda Maharaj, the president of Ramakrishna Math and Mission Order, at a city hospital.“The Prime Minister and Swamiji spoke in Gujarati. Swami Atmasthananda Maharaj also gave a small piece of chocolate to the Prime Minister,” Swami Satyadevananda, secretary of Ramakrishna Mission Seva Pratisthan said. He said that the 96-year-old blessed Mr. Modi putting his hand on his forehead. “It was a very informal meeting, the two met like guru-shishya. It did not seem as if he was the Prime Minister,” Swami Subhakarananda Maharaj told The Hindu.Mr. Modi urged Swami Subhakarananda Maharaj to sing a song for Swami Atmasthananda Maharaj. The monks at the Ramakrishna Mission also say that it was Swami Atmasthananda Maharaj who had advised him against becoming a monk and said that he was destined for other works.

Figure 15: Story 5



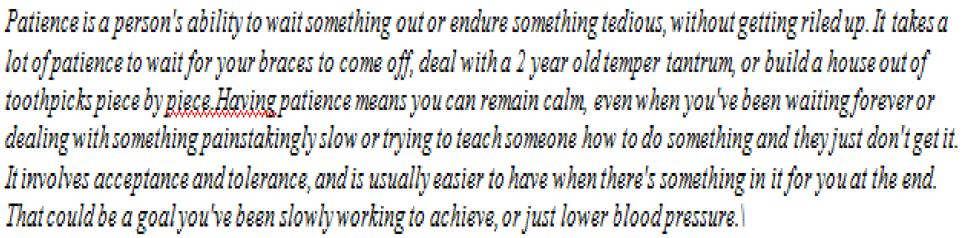
Music is an art form whose medium is sound. Its common elements are pitch (which governs melody and harmony), rhythm (and its associated concepts tempo, meter, and articulation), dynamics, and the sonic qualities of timbre and texture. The word derives from Greek µ??s??? (mousike ; "art of the Muses").[1] In its most general form the activities describing music as an art form include the production of works of music, the criticism of music, the study of the history of music, and the aesthetic dissemination of music. The creation, performance, significance, and even the definition of music vary according to culture and social context. Music ranges from strictly organized compositions (and their recreation in performance), through improvisational music to aleatoric forms. Music can be divided into genres and subgenres, although the dividing lines and relationships between music genres are often subtle, sometimes open to personal interpretation, and occasionally controversial. Within the arts, music may be classified as a performing art, a fine art, and auditory art. It may also be divided among art music and folk music. There is also a strong connection between music and mathematics. Music may be played and heard live, may be part of a dramatic work or film, or may be recorded. To many people in many cultures, music is an important part of their way of life. Ancient Greek and Indian philosophers defined music as tones ordered horizontally as melodies and vertically as harmonies. Common sayings such as "the harmony of the spheres" and "it is music to my ears" point to the notion that music is often ordered and pleasant to listen to. However, 20th-century composer John Cage thought that any sound can be music, saying, for example, "There is no noise, only sound.

Figure 16: Story 6



Once upon a time there lived an unhappy young girl. Her mother was dead and her father had married a widow with two daughters. Her stepmother didn't like her one little bit. All her kind thoughts and loving touches were for her own daughters. Nothing was too good for them - dresses, shoes, delicious food, soft beds, and every home comfort. But, for the poor unhappy girl, there was nothing at all. No dresses, only her stepsisters' hand-me-downs. No lovely dishes, nothing but scraps. No rest and no comfort. She had to work hard all day. Only when evening came was she allowed to sit for a while by the fire, near the cinders. That's why everybody called her cinderella. Cinderella used to spend long hours all alone talking to the cat. The cat said, Miaow. , which really meant, Cheer up! You have something neither of your stepsisters has and that is beauty. It was quite true. Cinderella, even dressed in old rags, was a lovely girl. While her stepsisters, no matter how splendid and elegant their clothes, were still clumsy, lumpy and ugly and always would be.One day, beautiful new dresses arrived at the house. A ball was to be held at the palace and the stepsisters were getting ready to go. Cinderella didn't even dare ask if she could go too.

Figure 17: Story 7



Patience is a person's ability to wait something out or endure something tedious, without getting riled up. It takes a lot of patience to wait for your braces to come off, deal with a 2 year old temper tantrum, or build a house out of toothpicks piece by piece.Having patience means you can remain calm, even when you've been waiting forever or dealing with something painstakingly slow or trying to teach someone how to do something and they just don't get it. It involves acceptance and tolerance, and is usually easier to have when there's something in it for you at the end. That could be a goal you've been slowly working to achieve, or just lower blood pressure.

Figure 18: Story 8

A Blind Boy Sat On The Steps Of A Building With A Hat By His Feet. He Held Up A Sign Which Said: "I Am Blind, Please Help." There Were Only A Few Coins In The Hat.A Man Was Walking By. He Took A Few Coins From His Pocket And Dropped Them Into The Hat. He Then Took The Sign, Turned It Around, And Wrote Some Words. He Put The Sign Back So That Everyone Who Walked By Would See The New Words. Soon The Hat Began To Fill Up. A Lot More People Were Giving Money To The Blind Boy. That Afternoon The Man Who Had Changed The Sign Came To See How Things Were.The Boy Recognized His Footsteps And Asked, "Were U The One Who Changed My Sign This Morning? What Did U Write?" The Man Said, "I Only Wrote The Truth. I Said What U Said But In A Different Way."What He Had Written Was: "Today Is A Beautiful Day & I Cannot See It."Do U Think The First Sign & The Second Sign Were Saying The Same Thing? Of Course Both Signs Told People The Boy Was Blind. But The First Sign Simply Said The Boy Was Blind. The Second Sign Told People They Were So Lucky That They Were Not Blind. Should We Be Surprised That The Second Sign Was More Effective?

Figure 19: Story 9

The Economy of India is the seventh-largest in the world by nominal GDP and the third-largest by purchasing power parity (PPP). The country is one of the G-20 major economies, a member of BRICS and a developing economy among the top 20 global traders according to the WTO.[29]According to the Indian Finance Ministry the annual growth rate of the Indian economy is projected to have increased to 7.4% in 2014-15 as compared with 6.9% in the fiscal year 2013-14. In an annual report, the IMF forecast that the Indian Economy would grow by 7.5% percent in the 2015-16 fiscal year starting on April 1, 2015, up from 7.2% (2014–15).[30][31]India was the 19th-largest merchandise and the 6th largest services exporter in the world in 2013; it imported a total of $616.7 billion worth of merchandise and services in 2013, as the 12th-largest merchandise and 7th largest services importer.[32] The agricultural sector is the largest employer in India's economy but contributes a declining share of its GDP (13.7% in 2012-13).[6] Its manufacturing industry has held a constant share of its economic contribution, while the fastest-growing part of the economy has been its services sector which includes, among others, the construction, telecommunications, software and information technologies, infrastructure, tourism, education, health care, travel, trade, and banking industries.

Figure 20: Story 10

Table 2: Describe the results of Title Generation of Compositional Semantics Tool

S.No.	Fig.No.	Actual Title	Generated by Tool
1	11	Thirsty crow	Thirsty crow
2	12	Data Mining	Data-mining Techniques anticipted/predicted performance student validation set
3	13	Foolish crow	Foolish crow Clever fox Hungry fox
4	14	Role of Gandhi in Indian independence	Resident india India Indian Independence Civil Rights

			Indian Community Traditional India
5	15	Prime minister visited his ailing guru	ailing guru first visit other works prime
6	16	Music is art	Music Improvisational music Fine art only sound impotent part

7	17	Unhappy girl Cinderella	Unhappy girl
			lovely girl
			cindrella
			new dress
8	18	Patient	patient
			old temper
9	19	Second sign	Second sign
			hat u
			few coin
			blind
10	20	Indian Economy	Indian Economy
			developing economy

			economy
			agricultural sector
			annual growth

Judgment by Human beings

Now We have seen the results given by the tool. we make the comparison between human assigned results and results given by composite semantic tool. Also identify the quality of machine generated titles. The titles given by the tool is quite significant. Other thing is that in this case every person can think with different point f view and can give different title.

CHAPTER-5

CONCLUSION & FUTURE SCOPE

5.1 CONCLUSION

We have described about a tool for automatically generating a title of a single document. An improved way is suggested to perform semantic analysis and get the main idea (theme) of the document. The Quality of automatic generated title depends on corpus and the linguistic analysis. Corpora will include all the parts of sentence such as the adjective, adverb, verb, conjunctions etc. Nouns are countless so it will not be the part of corpus. Moreover all the phases of linguistic analysis must be performing properly. Phonology is a speech recognize phase so will not be consider in this analysis. So the analysis of corpus of language with the science of language will lead this theses work.

5.2 FUTURE SCOPE

This research solves the problem of Text Summarization in NLP. This research is to build such tool that performs the semantic analysis on text and generate the title. Here, the text will be the number of sentences like a paragraph or story or news or article. Study is restricted to the area of English language. The structure of the English sentences is complex. English language provides a huge database of words. Corpus will be designed. Corpora will include the adjective, adverb, verb, Conjunctions etc. Nouns are countless so it will not be the part of corpus. But will consider the noun in semantic analysis. The editors of the newspapers or magazines can get the title or summaries the whole news very easily. Moreover they can get the appropriate phrase as well. It will be informative tool for the children. They can understand the basic structure of the sentence likewise noun, pronoun, adjective, adverb verb. They can learn the different phrases. The tool will work on the basic structure of the sentence. This tool can be used in mobile applications. It can act as a web service and can be a small component of whole application. With the help of this tool teachers in easily teach the students about the parts of the sentence. Moreover , Anybody can easily perform the sentence analysis. It promotes the more researches in the field of Natural language processing.

CHAPTE-6

REFERENCES

I. BOOKS

[1]. D.W.Patterson(1990) Introduction to AI & Expert Systems, Prentice Hall.

[2]. Mark Lutz (2009, September), Learning Python, O'Reilly Media , 4th edition.

[3]. Mark Lutz (2009, September), Python Poket Reference, O'Reilly Media , 5th edition.

[4]. Perkins Jacob (2010, November), Python Text Processing with NLTK 2.0 Cookbook, Packt Publishing, 1st edition.

II. RESEARCH PAPERS

[5]. Marne_e Marie-Catherine de and Christopher D. Manning (2008) "Stanford typed dependencies manual" The Stanford Natural Language Processing Group, Stanford Parser v. 3.3.

[6]. Apache OpenNLP Developer Documentation(2006), Apache OpenNLP Development Community , Version 1.5.2.

[7]. Cunningham, Hamish, et al.(2002) "GATE: an architecture for development of robust HLT applications." Proceedings of the 40th annual meeting on association for computational linguistics. pp. 168-175

[8]. Hersh R. Willian , Campbell H. Emily , Evans A. David and Brownlow D. Nicholas (1996) "Empirical Automated Vocabulary Discovery Using Large Text Corpora and Advanced Natural Processing Tools" Proc AMIA Annu Fall Symp :159-63.

[9]. Jin Rong and Hauptmann A.G. (2002) "A new probabilistic model fore title generation" In proceeding of the 19th international conference on computational linguistics-volume 1 ,pp.1-7.

[10]. Lopez, C., Prince, V., & Roche, M. (2012) "How to title electronic documents using text mining techniques" International Journal of Computer Information Systems and Industrial Management applications , vol 4, 562-569.

[11]. Jin, R., & Hauptmann, A. G. (2001, March). Automatic title generation for spoken broadcast news. In Proceedings of the first international conference on Human language technology research .Association for Computational Linguistics. pp. 1-3

[12]. Rajman, Martin, and Romaric Besançon.(1998) "Text mining: natural language techniques and text mining applications." Data Mining and Reverse Engineering. Springer US . 50-64.

[13]. Guo Yi, Shao Zhiqing(2010) "Cognitive Learning for Sentence Understanding" INTECH Open Access Publisher.

[14]. Barcala, Francisco-Mario, Jesús Vilares, Miguel A. Alonso, Jorge Grana, and Manuel Vilares(2002) "Tokenization and proper noun recognition for information retrieval." In Database and Expert Systems Applications, 2002. Proceedings. 13th International Workshop on, pp. 246-250. IEEE.

[15]. Batuer, Aisha, Sun Maosong (2009)," A Uyghur Morpheme Analysis Method based on Conditional Random Fields", International Journal on Asian Language Processing, 19 no.2 , pp 69- 77.

[16]. Stymne Sara (2013) "Treebank Grammars and Parser Evaluation", Uppsala universitet , Syntactic Analysis(5LN455).

[17]. Xiao Xinyan , Liu Yang, Hwang Young-Sook,Liu Qun , Lin Shouxun (2010)"Joint Tokenization and Translation" ,Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010), pages 1200–1208, Beijing.

[18]. Pera, M. S., & Ng, Y. K. (2009, November). Classifying Sentence-Based Summaries of Web Documents. In *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on* (pp. 433-440). IEEE.

[19]. Antiqueira, L., & Nunes, M. D. G. V. (2010). Complex networks and extractive summarization. In the Extended Activities Proceedings of the 9th International Conference on Computational Processing of the Portuguese Language–PROPOR. Porto Alegre/RS, Brazil.

[20]. Luhn H. P. (1958)"The automatic creation of literature abstracts.IBM Journal of Research and Development", 2, pp. 159-165.

[21]. Ishii, H., Lin, R., and Furugori, T. (2001). An automatic text summarization system based on the centrality of word roles in sentences (in japanese). Information Processing Society of Japan SIG Notes, 2001(20):83-90.

III. WEBLINKS

[22]. http://nlp.stanford.edu/software/tagger.shtml

[23]. https://sharpnlp.codeplex.com/

[24]. http://gate.ac.uk

[25]. https://wordnet.princeton.edu/wordnet/

[26]. http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php

[27] http://nltk.org/book_1ed.

CHAPTER-7

APPENDIX

List of Abbreviations:

NLTK: Natural Language Toolkit

HMM: Hidden Markov Model

AI: Artificial Intelligence

SMT: Simple Machine Translation

HV: Helping Verb

NLP: Natural Language Processing

O: Object

POS: Part of Speech

S: Subject

V: Verb

DUC: Document Understanding Conference

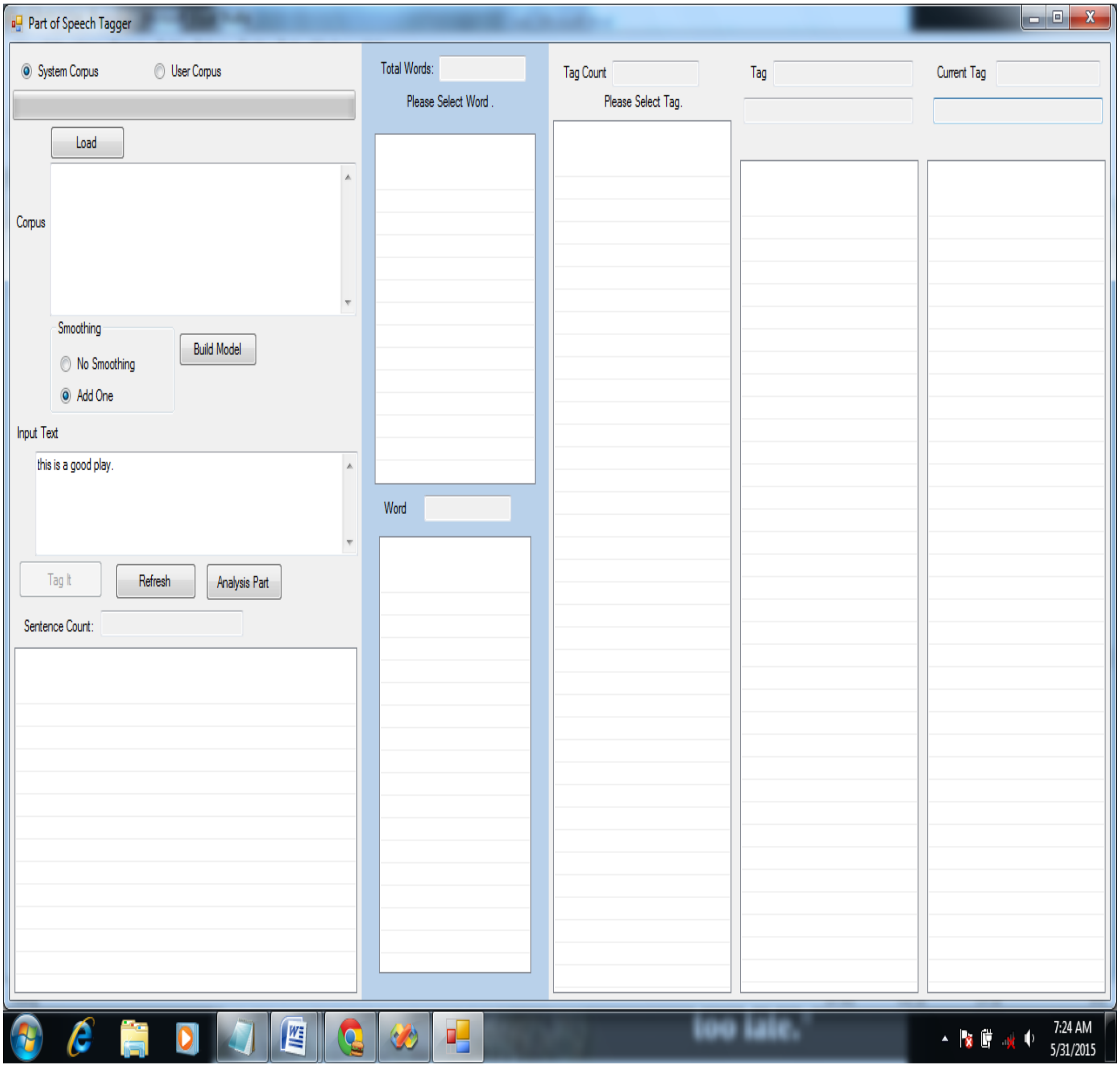
API: Application Program Interface

CLI: Command Line Interface

GATE: General Architecture of Text Engineering

KDD: Knowledge Discovery in Database

Execution outputs:

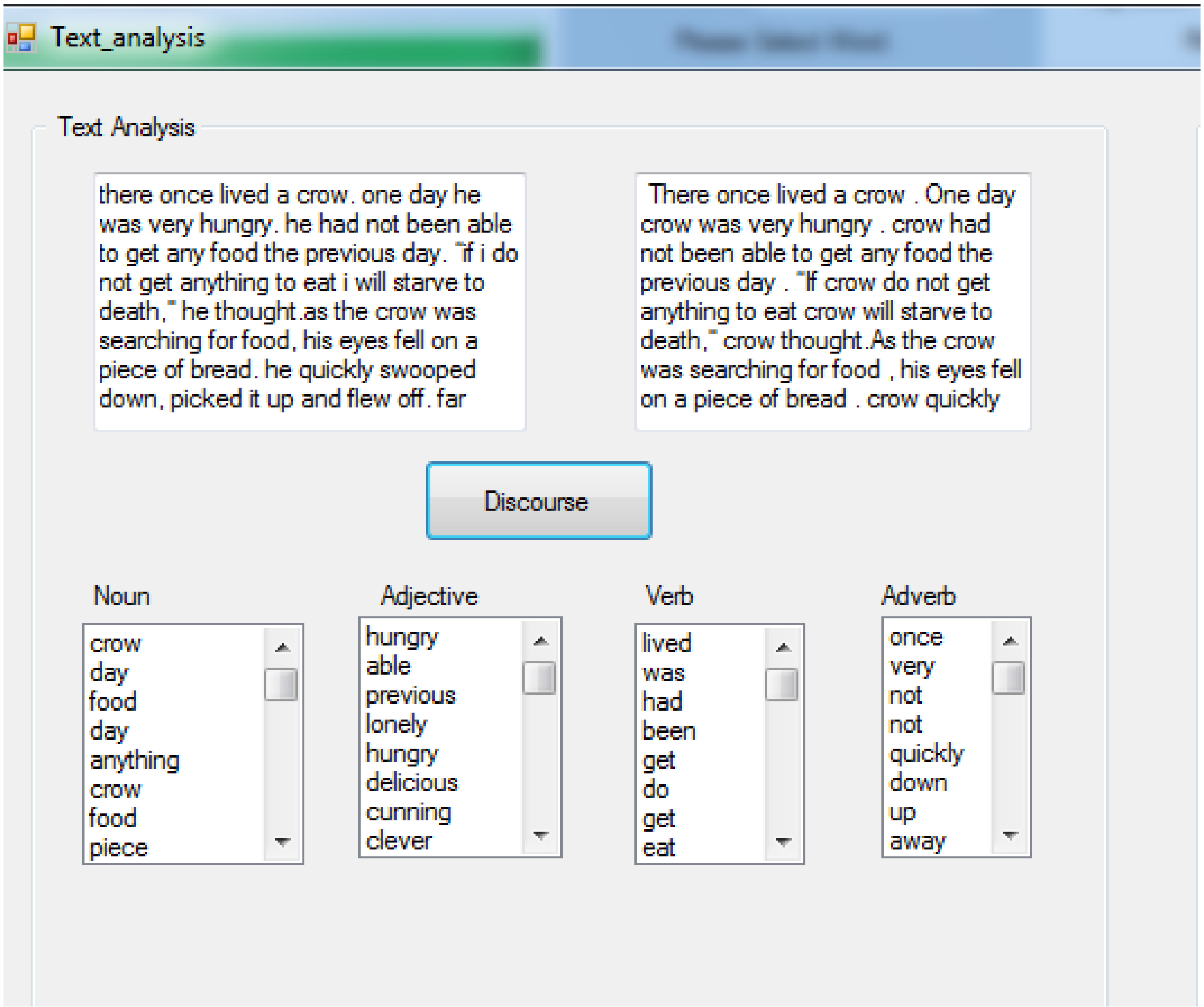


● System Corpus ○ User Corpus

Smoothing: ○ No Smoothing ● Add One

Input Text: this is a good play.

Figure 21: HMM Tagger Interface

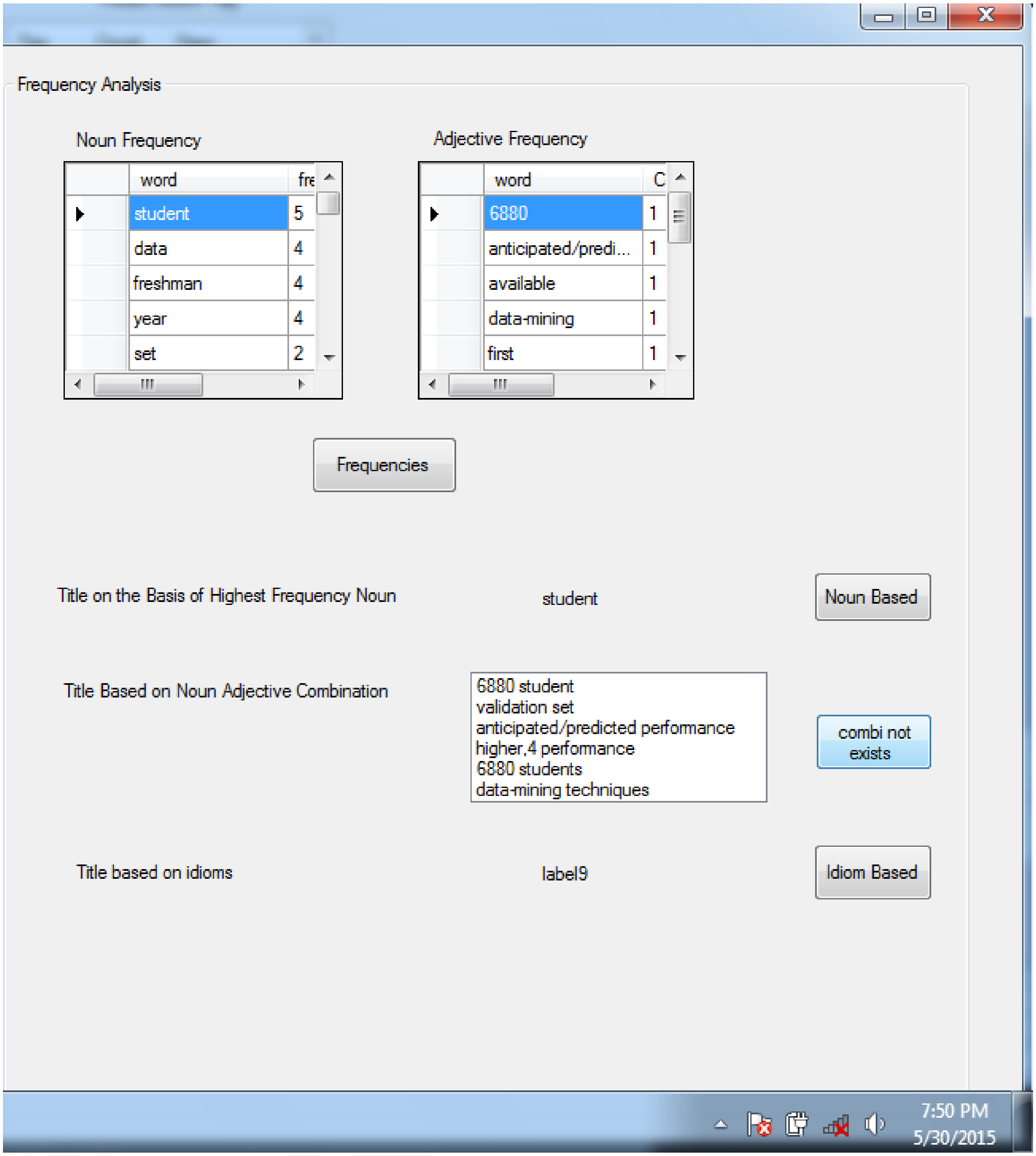


there once lived a crow. one day he was very hungry. he had not been able to get any food the previous day. "if i do not get anything to eat i will starve to death," he thought.as the crow was searching for food, his eyes fell on a piece of bread. he quickly swooped down, picked it up and flew off. far

There once lived a crow . One day crow was very hungry . crow had not been able to get any food the previous day . "If crow do not get anything to eat crow will starve to death," crow thought.As the crow was searching for food , his eyes fell on a piece of bread . crow quickly

```
Noun        Adjective    Verb      Adverb
crow        hungry       lived     once
day         able         was       very
food        previous     had       not
day         lonely       been      not
anything    hungry       get       quickly
crow        delicious    do        down
food        cunning      get       up
piece       clever       eat       away
```

Figure 22:Output of Figure13



Frequency Analysis

Noun Frequency

word	fre
student	5
data	4
freshman	4
year	4
set	2

Adjective Frequency

word	C
6880	1
anticipated/predi...	1
available	1
data-mining	1
first	1

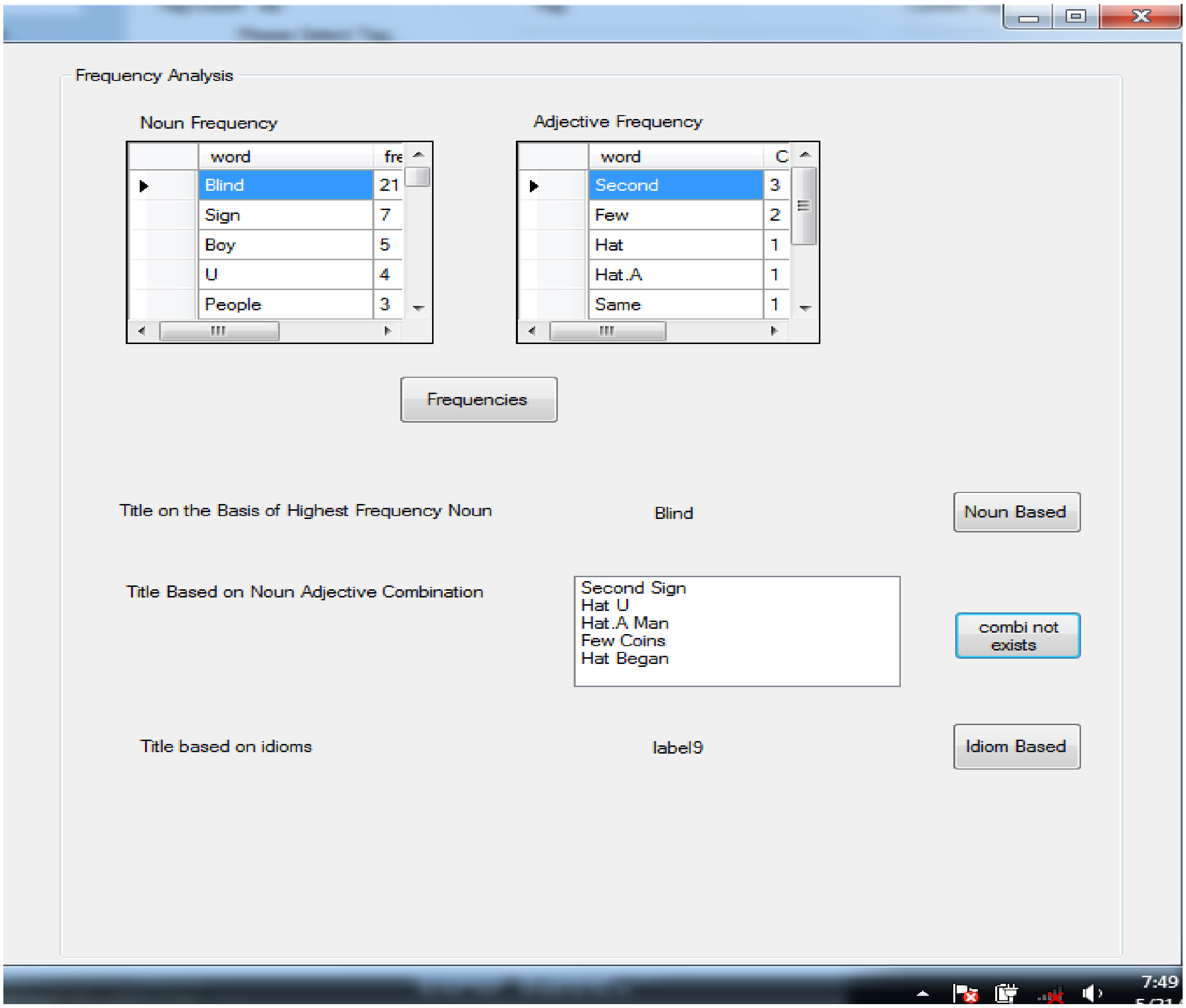
Title on the Basis of Highest Frequency Noun student

Title Based on Noun Adjective Combination

6880 student
validation set
anticipated/predicted performance
higher,4 performance
6880 students
data-mining techniques

Title based on idioms label9

Figure 23: Output of Figure 12



Frequency Analysis

Noun Frequency

word	fre
Blind	21
Sign	7
Boy	5
U	4
People	3

Adjective Frequency

word	C
Second	3
Few	2
Hat	1
Hat.A	1
Same	1

Title on the Basis of Highest Frequency Noun Blind

Title Based on Noun Adjective Combination

Second Sign
Hat U
Hat.A Man
Few Coins
Hat Began

Title based on idioms label9

Figure 24: Output of Figure 19