I J C T A, 9(11) 2016, pp. 5159-5168

AUTOMATED TITLE GENERATION IN ENGLISH LANGUAGE USING NLP

Nandini Sethi [*], Prateek Agrawal [**], Vishu Madaan[***], Sanjay Kumar Singh[****] and Anuj Kumar[*****]

Abstract: This paper describes an approach for generating or finding the central theme of the story or document in English language without reading the whole document. It takes a story as input and produces the title after applying various approaches like frequency prioritizing, noun and adjective combination or idiom based title. Title-generation will be helpful to the editors of newspaper or technical writers to easily finding the central them without reading the whole article. In our paper, we have described complete algorithm for generating the title of English article or story. The performance of the system is totally dependent on the size of Database. This application can be helpful for students as informative tool, teachers for teaching different structure of sentences and helps in structure analysis.

Key Words: NLTK, HMM, Part-of-Speech.

1. INTRODUCTION

Artificial Intelligence is very vast field. This field is having many complex sub fields. One of the subfield is natural language processing which is 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 the 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 [5]. For the language processing, python I is the powerful language. Python provides a toolkit to process the language. This toolkit provides the good result [6][8]. The training corpus is the most important element for language processing. AI is the challenge to build computational models and approaches of cognitive processes. Various approaches for NLP are given below [7]:

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

[*] School of Computer Science Engineering, Lovely Professional University, Punjab nandinisethi2104@gmail.com
[**] School of Computer Science Engineering, Lovely Professional University, Punjab prateek061186@gmail.com
[***] School of Computer Science Engineering, Lovely Professional University, Punjab vishumadaan123@gmail.com
[****] School of Computer Science Engineering, Lovely Professional University, Punjab sanjayksingh.012 gmail.com
[*****] School of Computer Science Engineering, Lovely Professional University, Punjab kakrananuj1}@gmail.com

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.

Different ways to generate the title:

1. Frequency based suggestive title.
2. Noun adjective combination based suggestive title.
3. Idiom based suggestive title.

The major objective is to suggest the titles for document. Some kind of compositional semantic tool that perform the text analysis will 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:

1. Lexical analysis
2. Part of Speech tagging
3. Discourse Analysis
4. Frequencies of lexemes
5. Various analysis on frequencies and texts

Various activities involved in the process of title generation:

Part-Of-Speech Tagger

POS 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, and 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 tags the text. It also counts the words [9].

Discourse Analysis

Discourse analysis deals with converting all pronouns into equivalent nouns. It makes the connection between the two sentences. Each individual sentence is dependable to previous sentence and makes influence of that [10].

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.

1. Ram is reading a book.
2. 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 than 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 and 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 starts with pronoun then no need to perform discourse over text.

2. PREVIOUS WORKS

In the existing Title generation system, R.Jin (2002) describes a new probabilistic approach for title generation. He performs 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 [1]. Cedric Lopez (2012) describes 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 their system they provide the percentage of all categories. More is the noun candidate leads to the quality improvement of the title. This is implemented for French language [2]. Paul E. Kennedy (2001) 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 are considered 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 consist 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 [3]. Mario Barcala (2002) explains 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 this is built in previous model proper noun identifier extracts the proper noun [4]. Most of the existing system focuses on the single approach either focuses on nouns or some vocabulary document while in proposed system there are different ways to generating the title of the given story.

3. OUR APPROACH

Proposed system is done in various steps lexical analysis, part of speech tagging, discourse analysis, frequencies of tokens, prioritize the frequencies etc shown in Figure 1

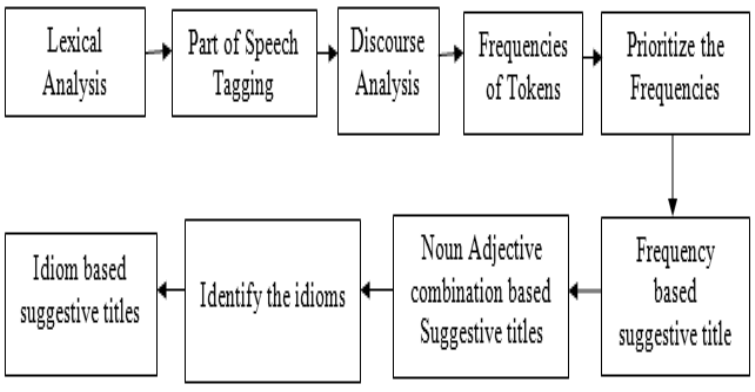


Figure 1: Processing steps of proposed approach

3.1 Lexical Analysis

Lexical analysis is concept of characters. Making a sequence of characters forms meaningful string. The sequence of characters is known as tokens. Tokens are nothing but a word [11].

1. The text is entered in the Rich Text Box. Sequence of characters is picked and gets the word to perform the further text analysis.
2. “ ” (space) between the two tokens is considered as the completion of single word.
3. “.”, ”?” And”!” characters are considered as the completion of the sentence.
4. Other Punctuation marks likewise (‘, ”” , ; ,) are also consider.

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 considered 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

Ws: = white space, sep: = separator, punch: = punctuation

For each (char c in text)

```
{
        If (c =ws or sep or punch)
        {
                Textbox: = word
                word: =””
        }
        Else
        {
                word: = word + c
        }
}
```

3.2 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 different parts are called tags. The tags collectively called tag set. A Single word can be classified as both noun and verb. Moreover, there are some derivations on words can change the 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 assigning 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.

Here, NN for noun, VB for verb, ADJ for adjective. All the tags are collectively known as tag set. Part-of-speech tagging can be writing like POST.

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

3.3 Discourse Analysis

Our system will perform discourse analysis by converting all pronouns into equivalent nouns. It makes the connection between the two sentences. Each individual sentence is dependable to previous sentence and makes influence of that sentence [10].

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.

The Figure 2 shows the discourse analysis performed by our system

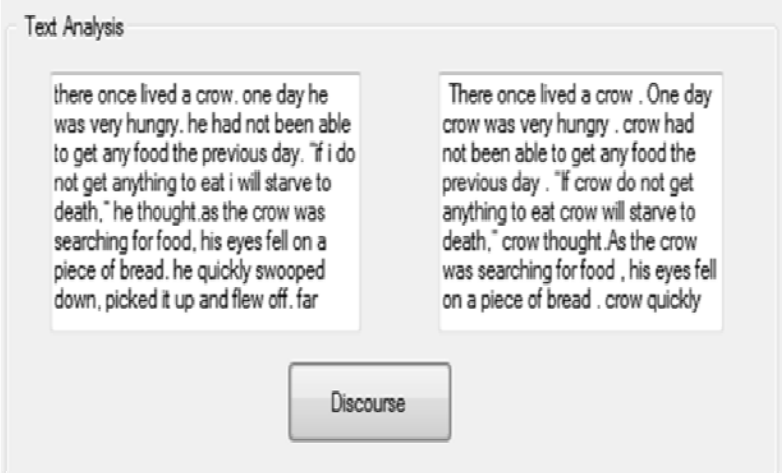


Figure 2: Discourse Analysis

3.4 Frequency of Tokens

The noun and adjective both are the most important in sentence structure. We need to calculate the occurrences of both elements. By calculating the occurrence means how frequently they are occurring in the story or paragraph. The process of discourse analysis will increase the accuracy of counting the frequency of the nouns as all the pronouns are replaced with their corresponding or associated nouns.

After calculating the frequency of nouns and adjectives then our system will prioritize or sort the nouns and adjectives in descending order. For this prioritizing our system will use the sorting process in descending order.

The Figure 3 shows the frequency analysis performed by our system:

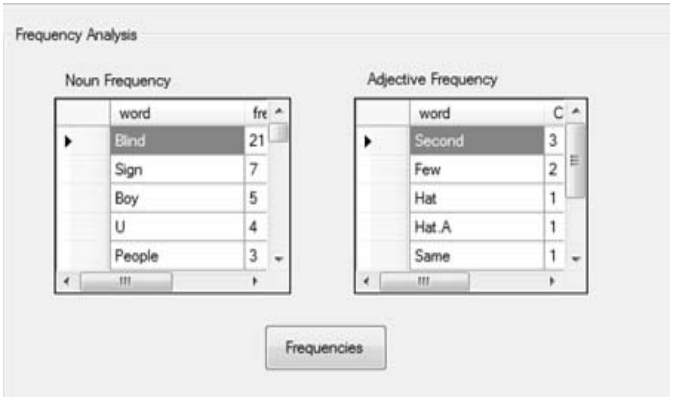


Figure 3: Frequency analysis

3.5 Suggestion of Title

The title will be suggested by three methods and all these methods performed by our system.

1. ***Frequency based suggestive title:*** The title word often found in the first sentence of the text. Nouns present at subject position are 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.
2. ***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. This combination will depend on the frequency of the nouns and adjective and collective presence of them in the story or article.
3. ***Idiom or proverb based suggestive titles:*** Some collection of words that provide some advice used in every culture are known as Idiom [12].

Example: When in Rome, do as the Romans.

4. Our system will search for the idiom or proverb in the story by matching with the database. If the idiom or proverb present in the story then it return it as the title of the story or article. As idiom or proverb present in the document is always being an appropriate title for the document.

Proposed algorithm

input_text = Story or paragraph

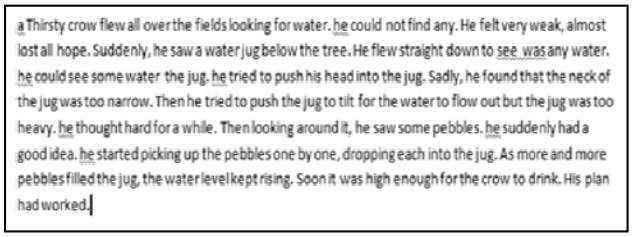
Title_generation (sentence or paragraph)

{

```
tagged_story = POS tagging (Input text);
c_text = Discourse analysis (tagged_story);
nouns [ ] = calculate_frequency_noun (c_text);
adjective [ ] = calculate_frequency_adj (c_text);
sort (nouns);
sort (adjective);
title [ ]  = highest frequency noun;
combo=adjective [i] +" "+ nouns[i];
if (input_text.contains (combo))
{
        title [ ] = combo;
}
idiom = list_of _idioms;      //fetched from database
if (input_text.contains (idiom))
{
title [ ] = idiom;
}
}
```

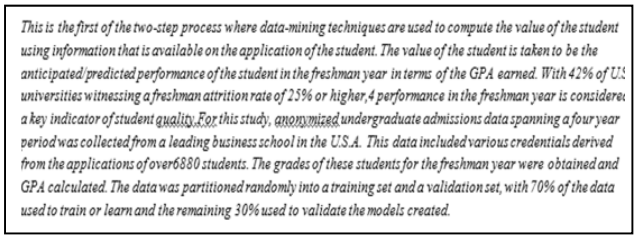
4. RESULTS

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. Below, we show the results of performing analysis for text using our compositional semantic tool.



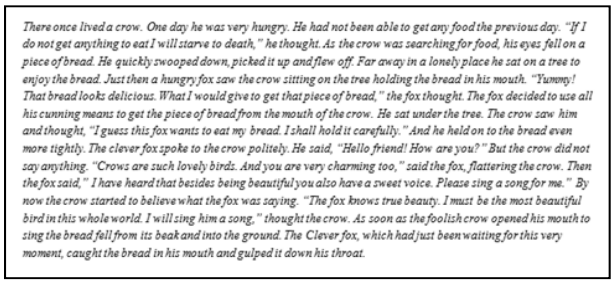
a Thirsty crow flew all over the fields looking for water. he could not find any. He felt very weak, almost lost all hope. Suddenly, he saw a water jug below the tree. He flew straight down to see was any water. he could see some water the jug. he tried to push his head into the jug. Sadly, he found 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 4: Story 1



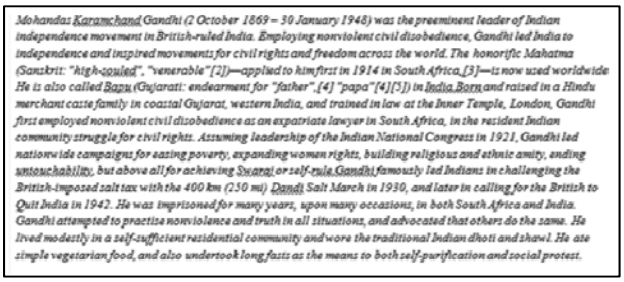
This is the first of the two-step process where data-mining techniques are used to compute the value of the student using information that is available on the application of the student. The value of the student is taken to be the anticipated/predicted performance of the student in the freshman year in terms of the GPA earned. With 42% of U.S universities witnessing a freshman attrition rate of 25% or higher,4 performance in the freshman year is considered 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 5: 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 6: 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 to independence and inspired movements for civil rights and freedom across the world. The honorific Mahatma (Sanskrit: "high-souled", "venerable"[2])—applied to him first in 1914 in South Africa,[3]—is now used worldwide. He is also called Bapu (Gujarati: endearment for "father",[4] "papa"[4][5]) in India.Born and raised in a Hindu merchant caste family in coastal Gujarat, western India, and trained in law at the Inner Temple, London, Gandhi 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 led nationwide campaigns for easing poverty, expanding women rights, building religious and ethnic amity, ending untouchability, but above all for achieving Swaraj or self-rule.Gandhi famously led Indians in challenging the British-imposed salt tax with the 400 km (250 mi) Dandi Salt March in 1930, and later in calling for the British to Quit India in 1942. He was imprisoned for many years, upon many occasions, in both South Africa and India. Gandhi attempted to practise nonviolence and truth in all situations, and advocated that others do the same. He lived modestly in a self-sufficient residential community and wore the traditional Indian dhoti and shawl. He ate simple vegetarian food, and also undertook long fasts as the means to both self-purification and social protest.

Figure 7: Story 4

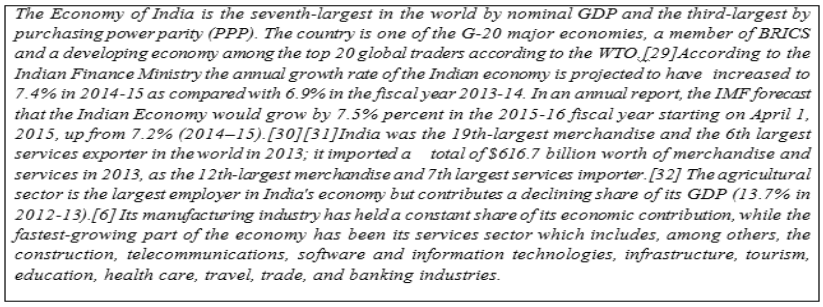

The Economy of India is the seventh-largest in the world by nominal GDP and the third-largest by purchasing power parity (PPP). The country is one of the G-20 major economies, a member of BRICS and a developing economy among the top 20 global traders according to the WTO.[29]According to the Indian Finance 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 8: Story 5

Table 2
Result Analysis

S.No	*Figure. No*	*Actual Title*	*Generated by Tool*
1	4	Thirsty Crow	Thirsty Crow
2	5	Data-Mining	Data-Mining Techniques anticipated/predicted performance Student
3	6	Foolish crow	Foolish crow Clever fox Hungry fox
4	7	Role of Gandhi in Indian independence	Resident India India Indian Independence Civil Rights
5	8	Indian Economy	Traditional India Indian Economy developing economy Economy agricultural sector annual growth

Judgment by Human beings

Now, above are 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 of view and can give different title.

5. CONCLUSION

In this proposed system, Title generation of English story or document using NLP is performed. This system can be used by scholars, technical writers, students and teachers. 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. 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. 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.

References

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

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

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

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

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

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

[7] http://www.mind.ilstu.edu/curriculum/protothinker/natural_language_processing.php //accessed on 12th January 2016

[8] http://nltk.org/book_1ed. // accessed on 10th January 2016

[9] http://www.cs.columbia.edu/~mcollins/courses/nlp2011/notes/hmms.pdf // accessed on 2nd January 2016

[10] http://219.244.160.5/yy/uploadfiles/201006/20100606095519379.pdf // accessed on 17th January 2016

[11] https://www.fbi.hda.de/fileadmin/personal/b.harriehausen/NLP/NLP_WS1112/NLP_Tokenisation__3__01.pdf //accessed on 2nd February 2016

[12] http://www.phrasemix.com/collections/the-50-most-important-english-proverbs. // accessed on 11th February 2016