Chapter 5

Parikh matrix and Natural languages

Natural languages are those languages which are used by human beings either vocally or in written form in their day to day life for communication. Natural language and formal language are different to each other with respect to their configuration and utility. Much work has been done to establish the interrelation between natural language and formal language. In this present work interrelation between the two is established.

Computational Linguistics is a subject which has its origin blended in two subjects - Axiomatic method and Method of grammatical description. The Axiomatic analysis is a subject which starts with a set of axioms. The axioms are nothing but statements which are assumed to be true and there is no scope to prove them . But using these axioms one can prove various theorems and transfer truth from the axioms to other statements by means of a fixed set of logical rules. Grammar is the set of rules for natural language.

'Natural Language Processing' (NLP) is the computerized approach to process natural languages. There are some other types of languages like programming languages and more. These types of languages fall under 'Ar-

tificial Language Processing'. Compilers and interpreters are topics of 'Artificial Language Processing'. There are some relations between NLP and 'Artificial Language Processing'.Roots of NLP lie mainly in the subjects Linguistics, Computer Science and Cognitive Psychology. One of the main goals of NLP is translating one natural language to another natural language using computer. Instruction given to a machine in natural languages should be understandable to the machine. This is another goal of NLP. Information retrieval system which uses NLP can understand what the reader is really searching for.

If it is possible to digitize natural language then it will be definitely helpful in machine translation. So it is tried to make some bridge between natural languages and arithmetic.

Limitations or challenges of this process of digitizing words are that this process suffers from word-sense disambigution. Word sense disambiguation (WSD) is the problem of determining in which sense a word is used in a given sentence. Words can have different senses. Some words have multiple meanings. This is called Polysemy. For example: The word 'admit' has two meanings. One meaning is to get access. Clarifying sentence is-"They cannot admit non-members into their club." Second meaning is 'to confess'. The clarifying sentence is "I am willing to admit that I do mistakes." Sometimes two completely different word are spelled the same. This is called Homonymy. Distinction between polysemy and homonymy is not always clear. Sometimes two completely different words are spelled the same. Sometimes the same words can be used as a verb and a noun but with completely different meaning. For example: Tear(verb)- shred to pieces with force. Clarifying sentence is - "The little girl tears the newspaper everyday." Tear (noun)- a drop of clear salty liquid secreted from glands in a person's eye when he or she cries or the eye is irritated. Clarifying sentence

is - "Criticism that left me in tears."

Formal language is a superset of context-free language. Within the field of computer science, specifically in the area of formal languages, the Chomsky hierarchy is a hierarchy of classes of formal grammars. This hierarchy of grammars was described by Noam Chomsky in 1956. Natural Language Processing is a theoretically inspired variety of computational techniques. It is used for analysing and representing naturally occurring texts at one or more levels of linguistic analysis. Bengali is a natural language. Bengali language is an emerging area of investigation of NLP. Many research work are going on in the field of Bengali language. One of the major areas where the researchers are toiling more is the generation of a context-free grammar for Bengali language and the properties associated with this. In this present chapter, while Parikh matrix has been applied in Bengali letters, words and sentence, the concept of context-free grammar was the working force behind.

The methodologies to solve problems regarding Natural language processing are divided broadly into four types. These are

a) Symbolic

b) Statistical,

c) Connectionist,

d) Hybrid.

Among these four categories statistical approach of NLP is the inevitable part for this present work.

Statistical Approach- A statistical model which is frequently used in NLP is the Hidden Markov model (HMM) . Hidden Markov model is a finite state automaton that has a set of states with probabilities attached to transitions between states . Although outputs are visible, associated states are not directly observable, thus hidden from external observations. Each state pro-

duces one of the observable outputs with a certain probability. Statistical approaches have typically been used in tasks such as speech recognition, lexical acquisition, parsing, part-of-speech tagging, collocations, statistical machine translation, statistical grammar learning and so on.

Chomsky hierarchy was a significant turning point in the study of NLP [25, 27].

While NLP is a relatively recent area of research and application, as compared to other information technology approaches, there have been sufficient successes that suggest that NLP based information access technologies will continue to be a major area of research and development in information systems in near future. Scopes of NLP are manifold and immense in the field of robotics and in other NLP applications. Computer science, Statistics, Mathematics, Linguistic and so many other branches of information are working hand in hand to achieve the goal of NLP.

5.1 Representation of Bengali letters by Parikh matrices:

The study of natural language grammar has its root in the ancient studies about the subject . As for example, over two thousand years ago the great Indian grammarian Panini wrote his grammar of Sanskrit. It is still a valuable reference in learning Sanskrit. Bengali is a natural language. And it is possible to make a context free grammar for Bengali. The matter of developing context-free grammar for Bengali is an emerging area of study. In the research papers [58, 67, 44, 65, 5] this problem is investigated. Depending on the above mentioned research work some results regarding context-free grammar are applied to Bengali language. Parikh matrix is one of the tools concerning context-free language. In this chapter Bengali alphabets are rep-

resented by Parikh matrices. Bengali Alphabet is given below:

অ আ ই ঈ

উ ঊ ঋ

এ ঐ ও ঔ

ক খ গ ঘ ঙ

চ ছ জ ঝ ঞ

ট ঠ ড ঢ ণ

ত থ দ ধ ন

প ফ ব ভ ম

য র ল শ ষ স হ

ড় ঢ় য় ৎ ং ঃ ঁ

The alphabet is divided into two parts-Swarbarna and Byanjanbarna. The first 11 letters are known as Swarbarna and the rest are Byanjanbarna. It is an ordered alphabet. There are 50 letters altogether in Bengali alphabet. In this section Parikh matrix is used in Bengali letters. For this we enumerate each letter one by one maintaining the above sequence of the letters in the alphabet. That is we are taking অ as the 1st letter, আ as the 2nd letter, and so on. In this way we shall get ক as the 12th letter, খ as the 13th letter, and last one ঁ as the 50th letter.

By Parikh Matrices, অ is defined as follows:

$$\begin{pmatrix} & 2^{nd} & & & & & & & \\ 1^{st} & 1 & 1 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ & 0 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ & 0 & 0 & 0 & 1 & 0 & 0 & \dots & \dots & 0 \\ & 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)},$$

Similarly আ and all other swarbarnas are defined. Again ক is defined as follows:

$$\begin{pmatrix} & & & & & & 13^{th} & & & \\ & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ & 0 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & 1 & \dots & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 12^{th} & 0 & 0 & 0 & \dots & 1 & 1 & 0 & \dots & 0 \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)},$$

Similarly খ and all other byanjanbarnas are defined.

5.2 Representation of Bengali words by Parikh matrices:

Bengali language has 50 letters. Making "Sandhi bicched" that is resolution of Bengali words into respective letters we can use Parikh matrices to Bengali

words. As for example we can take the following word. If we want to write আম , by using Parikh Matrices, we have to find the matrix product of the two matrices corresponding to

আ and ম = আম.

$$\begin{array}{c} \\ 2^{nd} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{3^{rd}}{} \\ \begin{pmatrix} 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & \dots & \dots & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)} \end{array}$$

$$\cdot \begin{array}{c} \\ 36^{th} \\ \\ \\ \\ \\ \\ \\ \\ \end{array} \begin{array}{c} \overset{37^{th}}{} \\ \begin{pmatrix} 1 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 0 & 0 & \dots & \dots & 1 & 1 & \dots & \dots & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)} \end{array}$$

$$= \begin{pmatrix} & & & 3^{rd} & & & 37^{th} & & & \\ & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 2^{nd} & 0 & 1 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & 1 & 0 & 0 & \dots & \dots & \dots & 0 \\ & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 36^{th} & 0 & 0 & \dots & \dots & 1 & 1 & \dots & \dots & 0 \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)} .$$

For clarity another example is discussed here. Let us take the word ছবি . This word can be resolved as ছবি = ছ + ব + ই .By matrix product of the Parikh matrices of the letters ছ , ব and ই we can get the Parikh matrix of the word ছবি as follows:

$$= \begin{pmatrix} & & & & 4^{th} & & 19^{th} & & 35^{th} & & \\ & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ & 0 & 1 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & 0 \\ 3^{rd} & 0 & 0 & 1 & 1 & 0 & \dots & \dots & \dots & \dots & 0 \\ & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 18^{th} & 0 & 0 & \dots & \dots & \dots & 1 & \dots & \dots & \dots & 0 \\ & 0 & 0 & 0 & \dots & \dots & \dots & \dots & \dots & \dots & 0 \\ 34^{th} & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & 1 & \dots & \vdots \\ & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \dots & \vdots & \vdots \\ & 0 & 0 & 0 & 0 & 0 & 0 & \dots & \dots & \dots & 1 \end{pmatrix}_{(51)\times(51)} , a_{i,j} = 0$$

for the rest of the entries of the 51×51 matrix.

5.3 Representation of Bengali sentence by Parikh matrix:

For further investigation on the use of Parikh matrix on Bengali language the following simple sentence is taken. যদু ভাল ছেলে . This sentence is written by permutation of the separate words as follows:

যদু ভাল ছেলে
ভাল ছেলে যদু
ছেলে ভাল যদু
ভাল যদু ছেলে
যদু ছেলে ভাল
ছেলে যদু ভাল

All the above sentences are having the same meaning. Difference is that some of them are frequently used in the Bengali language and some are not. We can take a line as a whole. We use Parikh matrix in the above Bengali lines as follows:

For the word যদু = য +দ + উ the Parikh matrix is $a_{i,i} = 1, a_{5,6} = 1, a_{29,30} = 1, a_{37,38} = 1$ and 0 for the rest of the entries of the 51×51 matrix.

For the word ভাল =ভ + আ + ল the Parikh matrix is $a_{i,i} = 1, a_{2,3} = 1, a_{35,36} = 1, a_{39,40} = 1,$ and 0 for the rest of the entries of the 51×51 matrix.

For the word ছেলে =ছ +এ +ল +এ the Parikh matrix is $a_{i,i} = 1, a_{2,3} = 3, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{34,36} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 1,$ and 0 for the rest of the entries of the 51×51 matrix..

Then the matrix product of all the Parikh matrices of the words is taken. This gives the Parikh matrix of the line যদু ভাল ছেলে as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ and 0 for the rest of the entries of the 51×51 matrix.

The Parikh matrix of the line ভাল ছেলে যদু as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ for the rest of the entries of the 51×51 matrix.

The Parikh matrix of the line ছেলে যদু ভাল as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{34,36} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ and 0 for the rest of the entries of the 51×51 matrix.

The Parikh matrix of the line ভাল যদু ছেলে as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ and 0 for the rest of the entries of the 51×51 matrix.

The Parikh matrix of the line যদু ছেলে ভাল as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{34,36} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ and 0 for the rest of the entries of the 51×51 matrix.

The Parikh matrix of the line ছেলে ভাল যদু as $a_{i,i} = 1, a_{2,3} = 2, a_{5,6} = 1, a_{8,9} = 1, a_{29,30} = 1, a_{34,35} = 1, a_{34,36} = 1, a_{35,36} = 1, a_{37,38} = 1, a_{39,40} = 2$ and 0 for the rest of the entries of the 51×51 matrix.

5.4 Conclusion of the Chapter:

Depending on the previous works done on the context-free grammar for Bengali language by various researchers Parikh matrix is applied on Bengali language. The Parikh matrix of every Bengali letter is a 51×51 matrix. All the entries of the main diagonal of this matrix is 1 and every entry below the main diagonal has the value 0 but the entries above the main diagonal provide information on the concerning Bengali letter . Every word is a matrix product of these matrices. The entries above the main diagonal provide information about the concerned Bengali word. The tool Parikh matrix is used for Bengali language. With the advancement of context-free grammar for Bengali language this effort will produce more fruitful results. Like Parikh matrix many other tools of formal language can be used in Bengali language. Various results of Parikh matrix can also be applied to Bengali language. Similar work can be done for other natural languages for which context-

free grammar has been developed. Although development of a context- free grammar for a natural language is a challenging job still whenever it is possible one can arithmatise the natural language using Parikh matrix.

*****