[1] Rita Chakraborty

[2] Prof. Shikhar Kr. Sarma

N-Gram based Assamese Question Pattern Extraction and Probabilistic Modelling

Abstract: - N-gram probabilities provide valuable information in understanding, processing, and modelling various natural language processing tasks. They assign probabilities to the sequences of words and subsequently to the whole sentence. Such information is very essential to make more accurate predictions in machine learning based systems. Here in this paper we worked on finding Parts-of-Speech (PoS) sequence based Assamese question patterns. We derived the unique bi-grams and tri-grams of PoSs occurring in the patterns and also extracted the probabilities of them. We then tried to find the unique PoS patterns of Assamese questions. We also have tried to incorporate the probabilities of unique bi-grams and tri-grams and the combined bi-grams and tri-grams probabilities of all patterns. Our work is a novel approach of finding the probabilities of bi-grams and tri-grams of the patterns occurring in Assamese questions.

Keywords: Assamese language, Parts-of-Speech (PoS), N-gram Probability, Natural Language Processing (NLP), Question-Answering

I. INTRODUCTION

Analysis of linguistic level information requires lots of processing. They are highly unstructured; unsuitable for processing. Processing of linguistic data requires them to be converted into structured form so that analysis, synthesis and processing can proceed [1]. Textual data are the rich sources of information. It can be in spoken or written form. Both require processing to be carried out through some range of tasks. Finally, we get a format which would enable human-like language processing. Analysis of such kind will not only make knowledge extraction possible but may provide very useful information about the language itself [2]. Today, we have many advanced tools and technologies which have enabled the naturally occurring languages become digitally rich as well as highly resourceful.

We are basically working on Assamese language which is a low-resource language. It is a language for which less research works have been done in the field of Natural Language Processing (NLP). Researches are going on to make the language fit in the digital world. Tools and technologies have been developed for this language also. The language is highly morphological and linguistically rich. Sentences are formed as sequences of well-organized, PoS tagged words.

A. Assamese questions

An Assamese question always ends with a “?” mark. In our work, Questions may comprise of one or multiple PoS tagged word(s). For instance,

i. হ’লনে ? <V_VM>is a one word type question sentence

ii. কোনে<PR_PRQ>শ্যামক<N_NNP>গুৱাহাটীলৈ<N_NNP>নিলে<V_VM>?is a question sentence comprising of multiple words.

The questions in Assamese begin with question words like কি, কোনে, কাক, ক’লৈ, কেনেকুৱাetc. Questions are of different types. Some are very simple comprising of just one word; basically a Verb followed by the question word (as in i.). Some contains multiple words along with the question word [3].

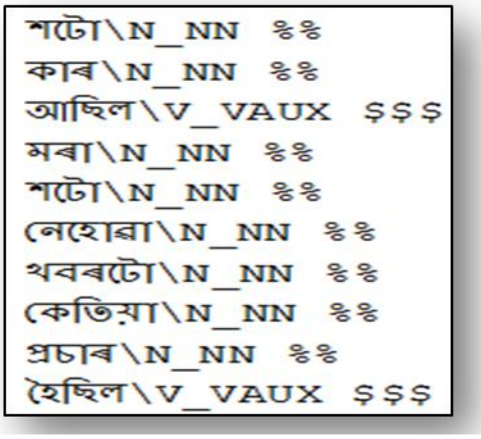
Computational processing of Assamese sentences require the questions to be spliced out once the “?” mark is encountered. We have a corpus with mixed sentence types (Both Assertive and Interrogative). It was a PoS

[1]Department of Computer Science and Information Technology, CDOE, Gauhati University, India 781014; tanushreegu@gmail.com

[2] Department of Information Technology, Gauhati University, India 781014; sks001@gmail.com

*Corresponding author: tanushreegu@gmail.com; sks001@gmail.com; Tel.: +91-8011003093

tagged corpus annotated with Bureau of Indian Standards (BIS) Tagset [4]. The corpus was developed by Department of Information Technology, Gauhati University. The corpus is duly tagged with BIS tagsets. It contains 1,53,457 duly annotated words and 3338 numbers of both assertive and question type Assamese sentences[5]. We initially tried to extract the Questions from the corpus. We derived 589 question sentences. A screenshot for the same has been provided in fig 1:



```
শটো\N_NN  %%
কাৰ\N_NN  %%
আছিল\V_VAUX  $$$
মৰা\N_NN  %%
শটো\N_NN  %%
নেহোৱা\N_NN  %%
খবৰটো\N_NN  %%
কেতিয়া\N_NN  %%
প্ৰচাৰ\N_NN  %%
হৈছিল\V_VAUX  $$$
```

Figure 1: Assamese Question Sentences

Each word in a question is duly tagged with PoS annotations. Next, we planned to generate only the PoS tagged sequence of each question.

B. N-gram Modelling

In computational linguistics, texts are formed as sequences of elements like characters, words, PoS annotations or tokens that appear continuously one after another. N-grams are the sequences of elements as they appear in the text. Traditional N-grams may be Unigrams (1-gram), Bi-grams (2-gram) or Tri-grams (3-gram). It is the simplest and extensively used concept in statistical language modelling. It applies the probabilistic approach to approximate the next probable sequence of words depending on sequences of words preceding it. We can even compute the probability of the entire sequence of words [6].

FOR INSTANCE, THE SENTENCE “WE WENT FOR A LONG WALK YESTERDAY” HAS:

- UNIGRAMS: “WE”, “WENT”, “FOR”, “A”, “LONG”, “WALK”, “YESTERDAY”.
- BI-GRAMS: “WE WENT”, “WENT FOR”, “FOR A”, “A LONG”, “LONG WALK”, “WALK YESTERDAY”.
- TRI-GRAMS: “WE WENT FOR”, “WENT FOR A”, “FOR A LONG”, “A LONG WALK”, “LONG WALK YESTERDAY”.

We shall attempt to apply this technique in Assamese language and also compute probabilities for each N-grams.

II. RELATED WORKS

Though considered to be simplest, N-gram approach is found to be an effective tool of study in the domain of Natural Language Processing (NLP).

Kostas Fragos and Christos Skourlas [7] proposed an N-gram based method for Authorship identification Problem. For that, they did character level segmentation of the author’s text writings. N-grams are formed for the piece of disputed text whose author is unknown. By counting the frequency of appearance of each N-gram, empirical distribution of N-grams for each author collection has been formed. To capture the abnormalities, it is compared with normal distribution of texts using the Kolmogorov-Smirnov test (KS-test). This test mainly deals with finding the difference between two underlying finite probability distributions. The author whose behaviour is found to be more abnormal is selected to be the correct author.

In their paper, William B. Cavnar and John M. Trenkle [8] proposed an N-Gram based technique to do categorization of texts. The N-gram frequency profiles are generated simply by reading and then counting the occurrences of all N-grams. The model generates tokens. Those tokens are scanned down to generate the possible N- order according to their number of occurrences or frequencies; resulting in an N-gram frequency profile of the document. The N-gram frequency profile of the training set as well as the N-gram profile of each article was also computed. Finally the distance between the sample profile and category profile was computed and the category with the smallest distance is picked.

E Brill, S Dumais and M Banko describe the architecture and system components of AskMSR question answering system [9]. Queries are reformulated simply by rewriting the strings. Each rewrite is fed as a search engine query and the corresponding page summaries are generated. N-grams are collected as possible answers from the page summaries. The summary text is processed according to the patterns specified by the rewrites and subsequently scored. Next, queries are analyzed based on question-type (*like what-type, who-type, where-type, when-type etc*). Filters are applied to the set of possible answers found during N-gram mining. These N-grans are rescored based on their features relevant to the filters. Finally, similar answers are merged and assembled into longer answers by overlapping smaller answer fragments. All subsequent candidates are checked to see if they can be tiled with the current answer. The longer tiled candidate gives the proper answer.

Ranking of passages is important as they have more probabilities to contain the answers. A Passage Retrieval System for Multilingual Question Answering System was proposed by JM Gómez Soriano, M Montes y Gómez, E Sanchis Arnal and P Rosso [10].The user question is fed into the search engine and N-gram extraction modules. The relevant passages are found out by the search engine. N-grams are produced both from the input question and the passages. Weightage is assigned to the passages in terms of the greater N-gram question structure found in the passage. With the increase in the number of considered passages; the percentage answers of generated also increase. They claimed the system is language independent which allows to be applied in multilingual question-answering systems.

When passed through noisy channel, it is not that easy to recover the words as they get garbled. In such situation, it is very important to estimate the probabilities in which words are fed into the noisy channel. PF Brown, VJ Della Pietra, PV Desouza, JC Lai and RL Mercer [11] discussed the method of how these estimates are made. Additionally, words are assigned to different classes based on their statistical behaviour of the surroundings from a large set of text. Classes are allocated to the unigrams, bigrams and trigrams. According to their frequencies in the text, two different words of the same class are distinguished. On the basis of their independent frequencies, if w_2 follows w_1 less often, then it is assumed that their mutual information is negative. On the other hand, if w_2 follows w_1 more often, then their mutual information is positive. The sticky pair is found out if their mutual information is greater than 0.Probabilities of words at random occurring near to each other are also derived. Words w_1 and w_2 are semantically sticky if their combined probability is more than their individual probabilities. It is observed from their experiment that the trigram model with their classes requires less storage (about one third as much as storage required to store each word).

A short keyword search query is sufficient to generate the ranked list of heterogeneous collection of documents available in World Wide Web (WWW). There is mostly term-based matching between the queries and documents since the keywords determine the relevance between the query and the documents. Christina Lioma and C. J. Keith van Rijsbergen [12] propose a POS based information in order to compute term weight or POS Information Score (PIS). Term weights or PIS form the integral part of a statistical IR modelling system and also important from the perspective to determine how informative the words are. Term weights are computed from grammatical information in terms of POS and the context in which they occur. A contiguous sequence of n-grams has been considered from a sample. Links are established between the terms and POS tags by relating a term to all possible n-grams that contain it. Probabilities are computed for each POS prior to computing that of POS n-grams. Similarly, computing PIS of a term requires first to map all term n-grams in which the term occurs to their corresponding POS n-grams. Finally, the probabilities of each of these POS n-grams are combined. PIS of a term is derived by finding the ratio between the probabilities of all possible n-grams containing the term to the total number of POS n-grams occurring in the collection of documents. BM25 model is used to match the document with query. Finally, PIS is integrated with the relevance weight between the query term and the document in order to find the relevance score between the document and the query. Experimental results show that PIS either improves the IR process or does not alter the performance of retrieval. Long queries contain more words resulting in more contribution of PIS. However, short queries mainly have keywords, where PIS tend to have less contribution. Therefore, long queries are more beneficial in IR process.

D Buscaldi, P Rosso, JM Gómez-Soriano and E Sanchis [13] have proposed a model on Passage Retrieval (PR) based Question-Answering (QA) system. The Question Classification and Analysis module uses a pattern based classifier. It analyzes the CLEF QA test set 2003-2006 and can handle questions for the classes. Those questions that do not match with any defined patterns fall under the category OTHER. Apart from question classification, the system does question analysis task in order to identify the constraints required during Answer Extraction. The JAVA Information Retrieval System (JIRS) is a passage retrieval system based on n-gram model. N-grams are the sequence of adjacent terms extracted from the sentence or the question. JIRS is based on the concept that the n-grams should appear near the answer at least once. Passages are retrieved and searches are made for all possible n-grams of the questions in the retrieved passages. Passages are given rates depending on the number and weights of the n-grams. The input question is also received by the Sentence Retrieval engine. It returns ranked list of sentences containing question keywords. These sentences form passages ranked by JIRS and later used in the Answer Extraction module. The answer coverage obtained by JIRS is found to be better than any other Information Retrieval (IR) system.

G McDonald, C Macdonald and I Ounis [14] have proposed an approach to identify the sensitive text in documents. The approach considers POS n-grams as their distribution indicate the amount of information they contain. The sensitivity load of each distribution measures identify the sequences of sensitive text occurring in a document. The documents are represented in terms of POS n-grams. Then the probabilistic approach is applied in order to measure the sensitivity load of these n-grams. A two-way contingency table was constructed indicating the sensitivity of text. Sensitivity or non-sensitivity depends on the documents that are either with POS or without POS. The dependency between the n-gram and sensitivity of text is measured by Chi-square test. It determines the amount by which the observed frequency of POS n-gram differs from the expected frequency. If the Chi-square score is greater than the distribution's critical value, then the distribution of the POS n-gram is regarded as sensitive text. The POS n-gram is sensitivity loaded. The sensitivity loaded CRF 10-gram achieves 0.9992 precision with an accuracy of 0.7282. In 67% document size the sensitivity loaded CRF method is able to identify 99% sensitive text.

Sidorov et al. [15] proposes a rule-based approach to automatic English Language Learning grammar error correction. It is a rule-based approach having few additional resources- a morphological analyzer in English is a dictionary containing 71000 words with corresponding POS annotation and a list of 250 common uncountable nouns used for finding the possibility of using nouns in plural form. The system also uses a training data having syntactic contents available in it. These contents are represented in terms of syntactic n-grams. In a syntax tree, the head word and its related dependent word are graphically showcased. Syntactic n-grams are constructed by starting at the root word and continue by considering each dependent word. Five types of errors are considered-noun number, incorrect preposition, choice of determiner or article, subject-verb agreement and verb form. The corpus is processed with the Stanford parser and used for the purpose of extraction of patterns for finding preposition error and also formulation of rules. Upon providing data along with corresponding POS annotation, the system performs with a precision of 25% which is considerably low because of the simplicity of the system. Since the system is rule-based, the system may be considered as the foundation for doing research in finding grammar error correction. Moreover, though this system cannot compete with ML based systems, it may be used as complementary to Machine Learning technique.

Incorporation of POS tagger based lexical and syntactic knowledge and a syntactic chunker into a traditional Information retrieval (IR) system have been proposed by Antonio Ferrandez [16]. IR systems basically consider the key terms of the query and documents are segmented based on these terms. Segmentation is relevant to the key terms. The proposal is termed as LexSIR (Lexical and Syntactic knowledge for Information Retrieval) and it considers two languages: Spanish and English. Instead of parsing the whole query, it is chunked into simple phrases like noun phrases or verb phrases. Depending on the lexical category of each query term, TP measure is calculated. Weight of each term is computed depending on remaining terms of each phrase. Weights are not added if terms are in the same phrase. Distance between terms is computed by measuring the intermediate phrases or sentences between terms. TP measures are applied to the phrases. Documents are segmented into sentences which in turn are segmented into phrases or entities. TP distances are measured for the entities also. The LexSIR algorithm finds out the similarity between a document and a query. Both the document and the query are PoS tagged and chunked into phrases. Stop words are removed. Estimate of weight of each query term is calculated. In this way, a set of weights will be generated for the whole query. Finally, the sum of all weights is also calculated. Apart from this, lowest penalization value for the distance between each query term is computed. Similarly, weight of the whole document is also calculated. For each pair of query phrases, penalization value for the distance between each query term is computed. Query terms are checked to see if they exist in a document. The proposed method provides consistent result in an IR system had a mean average precision over 0.5, with linguistically inflected like Spanish and also with different corpora as well as queries with varied lengths.

Text categorization is important for automated handling of documents. William B. Canvar and John M. Trenkle [17] propose an n-gram based text categorization system which works reasonably well for classifying articles also. The system begins with text samples belonging to some pre-existing subject categories. Next, each category is represented by generating n-gram frequency profile. Once a new document comes in for classification, the system first tries to generate its n-gram frequency profile. This new profile is compared with all existing profiles in order to compute the profile distance. The document is classified as belonging to the category having the smallest distance. N-gram frequency is an effective way of classifying documents. The system achieves 99.8% language classification rate.

Computational study and analysis of any naturally occurring language requires development of structured text corpus. Assamese also has its own. SK Sarma, H Bharali, A Gogoi, R Deka and A Barman [18] have presented their experience during building of Assamese raw text corpus. The corpus is unannotated and approximately 1.5 million (total 1,577,750 words). The corpus mainly focuses on its internal structure; i.e. the genres included and the length and number of individual text sample. There are three main categories- media, learned material and literature. These are again sub-divided into sub-categories. Media has the sub-categories newspaper and magazine. Learned material is categorized into science and arts. Similarly, literature is categorized into short fiction, criticism, theatre, novel, trivia, art and craft letter and didactic material. While selecting the genres, poetries are not considered. All the necessary domains are considered during selection of learned material. Finally, computerized entry of the corpus requires not only typing of text but entering its metadata also. The metadata includes the genre of text, its type, names of author, editor and publisher, date and place of publication and the page numbers. This corpus serves as an important language tool for NLP researchers.

Assamese has its own PoS tagger. In their paper, AK Barman, J Sarmah and SK Sarma [19] proposed two methods of tagging- Conditional Random Field (CRF) based and Transformation Based Learning (TBL). CRF is a probabilistic approach for labeling data. Provided the observation sequence O= {O1, O2, O3, ………….. , Or}, the conditional probability of the state sequence S= {S1, S2, S3, ………….. , Sr} is calculated. Feature templates of unigram and bigram contextual features used by the CRF tagger are found out. The CRF module is trained up to generate all features to make probabilities during tagging. TBL is a machine learning based approach which begins with a simple solution of the problem. Transformations are applied to select the best one and it continues until the selected transformation does not change the data or there is no more transformation to be applied. Linguistic features are considered for PoS tagging task. A heuristic function predicts the value of the feature. A learning module determines the correct value of the features. To find the performance, Precision and Recall of the taggers are calculated and finally combined to calculate the F-measure of overall performance. It has been observed that TBL gives more accurate result with a performance of 87.17% than CRF with 67.73%.

Assamese WordNet is a lexical repository and a tremendous source of information for the language. There are some major components which best describe each word of a WordNet- its Unique Identification Number, part-of-speech category, Synonyms arranged in accordance with frequency in which they are being used, concepts behind the synset defined by the term gloss. The WordNet is divided into four categories- Core, Common, PAN-Indian and Universal. An approach to classify documents using Assamese WordNet has been proposed by J Sarmah, N Saharia and SK Sarma [20]. The proposed structure has three major phases- pre-processing phase, tuned categories and classification phase. During pre-processing, tokenization, stop word removal, sorting, stemming and threshold value generation of texts are conducted. Document classification basically takes place in supervised way and the system has some pre-defined categories. These categories must be tuned so that the document to be classified to one of those categories. The tuned categories - News, Sports, Science and Arts include only those terms that are in the synset. Each category is passed through the pre-processing phase. Each term of the category is checked against the synset of each WordNet category except the Core. Finally, the document classification phase classifies an input document into one of those tuned categories. It also goes through the pre-processing steps. The most frequent words of the document which also appear in the synset are derived. For each such term, an extended form (including the term and its synset) is also created. These extended terms are tokenized so that they can be checked against each of the tuned categories. The document is categorized to the highest number of matching terms in a tuned category. The system attains an accuracy of 90.27%.

Stemmers are important for developing Information Retrieval (IR) systems. A Gogoi, N Baruah, SK Sarma and RD Phukan [21] have developed an Assamese stemmer capable of stemming suffixes from words. Suffixes are collected manually and divided into eight different categories: plural words, case words, definitive words, pleo words, extra words, indefinitive words, verbs and kinships. The dataset contains 20000 Assamese words. A word may contain more than one suffix of different types. The proposed approach starts by dictionary search of the input word. If it is there, simply the word is outputted. Else, the word is searched to check if any sub-word contains suffixes present in the suffix list. Then those suffixes are removed and again a dictionary search is made with the remaining stemmed word. If they match, simply it is outputted; otherwise again the checking in the suffixes list is performed. Also, the PoS annotation of the word is checked. If it is a proper noun, it is declared as the correct stemmed word. Otherwise, replace it with the original word and display it. The proposed stemmer attains 86.16% accuracy.

Named entity Recognition (NER) plays a very crucial role in developing information extraction system, machine translation system or a question-answering system. G Talukdar, PP Borah and A Baruah [22] have proposed an NER system in Assamese language emphasizing on individual features contributing towards recognition of entities like person, location or organization. The system follows supervised approach of learning using Naive Bayes classification model and considers four features to train the classifier. The first word of a compound proper noun (sequence of proper nouns) may represent a named entity. Similarly, the last word of a compound proper noun (sequence of proper nouns) may also represent a named entity. Previous word is another important feature which may indicate the presence of a named entity in the next position. Finally, the next word also gives information about the presence of a particular named entity from a sequence of words. It indicates that the named entity is present in the previous position within the sequence. There is a PoS tagged corpus with 5000 words. A named entity tag list is maintained to keep track of the named entities present in each sentence. The training phase of the system starts with tokenization of each word of the corpus. Then it is checked whether a sequence of words in tagged with proper noun (NNP). The four features are implemented and the output produced is classified as person, location and organization.

III. Probability of N-grams

Probability of an event is the likelihood that the event will occur. N-gram modelling applies the probabilistic approach to predict which word will follow a certain sequence of words. Probabilistic theories are basically applicable in ambiguous, noisy inputs. They are essential in applications like speech recognition, spelling correction or grammatical error correction.

Our work mainly derives the Bi-gram and Tri-gram probabilities of Assamese questions. For that we use the following approximation rule for a pattern with length N.

$$P\ (W_N \mid W_{N-1}) = C\ (W_{N-1}\ W_N)\ /\ C\ (W_{N-1}) \qquad \text{---------} \qquad 1$$

Here, P and C stand for Probability and Count of a pattern. Considering the pattern N_NN Nloc N_NN N_NN CC_CCD N_NN N_NN N_NN N_NN of length 9, we can generate the probabilities of all Bi-grams using rule 1 in the following ways:

P (Nloc | N_NN) = C (N_NN Nloc) / C (N_NN)

P (N_NN | Nloc) = C (Nloc N_NN) / C (Nloc)

P (N_NN | N_NN) = C (N_NN N_NN) / C (N_NN)

P (CC_CCD | N_NN) = C (N_NN CC_CCD) / C (N_NN)

P (N_NN | CC_CCD) = C (CC_CCD N_NN) / C (CC_CCD)

P (N_NN | N_NN) = C (N_NN N_NN) / C (N_NN)

P (N_NN | N_NN) = C (N_NN N_NN) / C (N_NN)

P (N_NN | N_NN) = C (N_NN N_NN) / C (N_NN)

However, it is worth mentioning here that the probability function P (N_NN | N_NN) = C (N_NN N_NN) / C (N_NN) has duplicate entries. Therefore, we shall derive the probabilities of only those Bi-grams which are unique in nature. Finally, combined probability of a particular pattern is achieved by multiplying the individual probability of each Bi-gram occurring in the pattern in the following way **[**23**].**

P (Nloc | N_NN)* P (N_NN | Nloc) * P (N_NN | N_NN) * P (CC_CCD | N_NN) * P (N_NN | CC_CCD) * P (N_NN | N_NN) * P (N_NN | N_NN) * P (N_NN | N_NN)

Similar is the case with generating probabilities of Tri-grams. For that we use the following approximation rule for a pattern with length N.

$$P\ (W_N \mid W_{N-2}\ W_{N-1}) = C\ (W_{N-2}\ W_{N-1}\ W_N)\ /\ C\ (W_{N-2}\ W_{N-1}) \qquad \text{---------}\ \ 2$$

In simple notation, it can be understood that if a sequence of words W_1 W_2 exists, then the probability that W_3would follow the sequence is determined by the approximation rule 2.

Considering the same question pattern as above, rule 2 generates the probabilities of Tri-grams existing in this pattern.

P (N_NN | N_NN Nloc) = C (N_NN Nloc N_NN) / C (N_NN Nloc)

P (N_NN | Nloc N_NN) = C (Nloc N_NN N_NN) / C (Nloc N_NN)

P (CC_CCD | N_NN N_NN) = C (N_NN N_NN CC_CCD) / C (N_NN N_NN)

P (N_NN | N_NN CC_CCD) = C (N_NN CC_CCD N_NN) / C (N_NN CC_CCD)

P (N_NN | CC_CCD N_NN) = C (CC_CCD N_NN N_NN) / C (CC_CCD N_NN)

P (N_NN | N_NN N_NN) = C (N_NN N_NN N_NN) / C (N_NN N_NN)

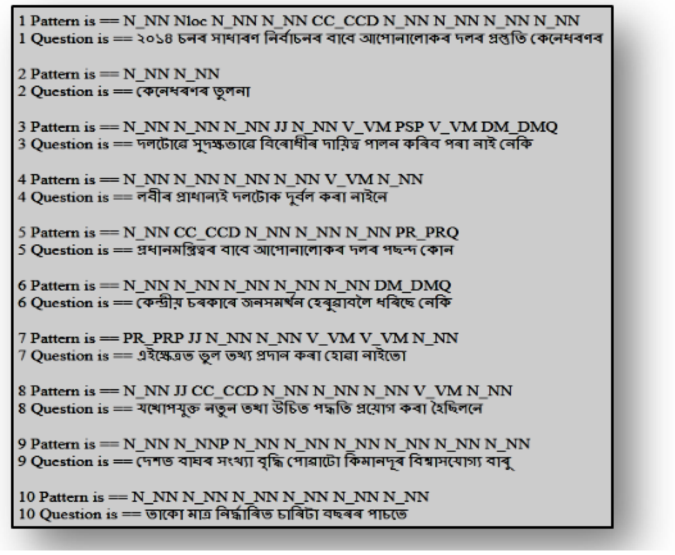
P (N_NN | N_NN N_NN) = C (N_NN N_NN N_NN) / C (N_NN N_NN)

We can observe here that the probability P (N_NN | N_NN N_NN) = C (N_NN N_NN N_NN) / C (N_NN N_NN) has duplicate entries. Our task would be to generate the unique trigrams and their frequencies. The combined probability will be achieved by multiplying the individual Tri-gram probabilities.

P (N_NN | N_NN Nloc) * P (N_NN | Nloc N_NN) * P (CC_CCD | N_NN N_NN) * P (N_NN | N_NN CC_CCD) *P (N_NN | CC_CCD N_NN) * P (N_NN | N_NN N_NN) * P (N_NN | N_NN N_NN)

IV. Our Work

We intend to derive Bi-gram and Tri-gram probabilities of the Assamese question patterns. Out of 3338 numbers of annotated sentences, we have derived 589 numbers of question patterns. We removed the duplicates and got 511 unique patterns in terms of PoS tags. The question sentences have also been generated. The PoS tagged unique question patterns are shown in fig 2:



```
1 Pattern is == N_NN Nloc N_NN N_NN CC_CCD N_NN N_NN N_NN N_NN
1 Question is == ২০১৪ চনৰ সাধাৰণ নিৰ্বাচনৰ বাবে আপোনালোকৰ দলৰ প্ৰস্তুতি কেনেধৰণৰ

2 Pattern is == N_NN N_NN
2 Question is == কেনেধৰণৰ তুলনা

3 Pattern is == N_NN N_NN N_NN JJ N_NN V_VM PSP V_VM DM_DMQ
3 Question is == দলটোৱে সুদক্ষতাৰে বিৰোধীৰ দায়িত্ব পালন কৰিব পৰা নাই নেকি

4 Pattern is == N_NN N_NN N_NN N_NN V_VM N_NN
4 Question is == নবীৰ প্ৰাধান্যই দলটোক দুৰ্বল কৰা নাইনে

5 Pattern is == N_NN CC_CCD N_NN N_NN N_NN PR_PRQ
5 Question is == প্ৰধানমন্ত্ৰিত্বৰ বাবে আপোনালোকৰ দলৰ পছন্দ কোন

6 Pattern is == N_NN N_NN N_NN N_NN N_NN DM_DMQ
6 Question is == কেন্দ্ৰীয় চৰকাৰে জনসমৰ্থন হেৰুৱাবলৈ ধৰিছে নেকি

7 Pattern is == PR_PRP JJ N_NN N_NN V_VM V_VM N_NN
7 Question is == এইক্ষেত্ৰত ভুল তথ্য প্ৰদান কৰা হোৱা নাইতো

8 Pattern is == N_NN JJ CC_CCD N_NN N_NN N_NN V_VM N_NN
8 Question is == যথোপযুক্ত নতুন তথ্য উচিত পদ্ধতি প্ৰয়োগ কৰা হৈছিলনে

9 Pattern is == N_NN N_NNP N_NN N_NN N_NN N_NN N_NN N_NN
9 Question is == দেশত বাঘৰ সংখ্যা বৃদ্ধি পোৱাটো কিমানদূৰ বিশ্বাসযোগ্য বাৰু

10 Pattern is == N_NN N_NN N_NN N_NN N_NN N_NN
10 Question is == তাকো মাত্ৰ নিৰ্দ্ধাৰিত চাৰিটা বছৰৰ পাচতে
```

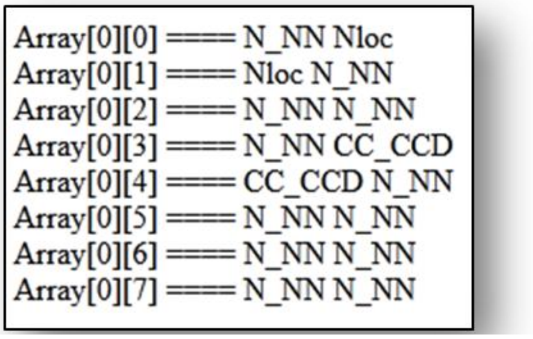
Figure 2: Unique patterns of Assamese questions

We next moved on by generating the Bi-gram and Tri-gram sequences of the patterns. We got 3414 Bi-grams and 2891 Tri-grams.

A. Bi-gram Probabilty

Our next task was to find the numbers of occurrences of Bi-grams and Tri-grams throughout the whole patterns. We also have derived the unique PoS tags and their frequencies of occurrences.

For instance, considering the Bi-gram pattern in Fig 3, we want to calculate the probability of occurrence of Nloc after N_NN.



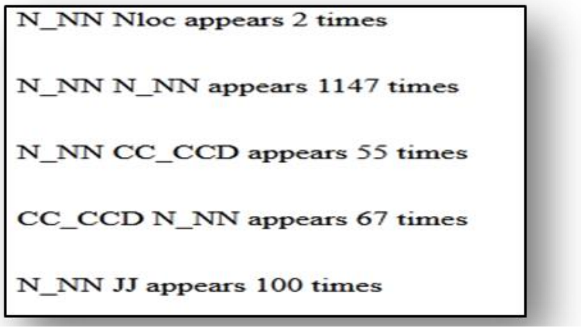
```
Array[0][0] ==== N_NN Nloc
Array[0][1] ==== Nloc N_NN
Array[0][2] ==== N_NN N_NN
Array[0][3] ==== N_NN CC_CCD
Array[0][4] ==== CC_CCD N_NN
Array[0][5] ==== N_NN N_NN
Array[0][6] ==== N_NN N_NN
Array[0][7] ==== N_NN N_NN
```

Figure 3: A Bi-gram pattern

Here, we apply rule 1 in order to find the probability. The probability function would be:

P (Nloc | N_NN) = C (N_NN Nloc) / C (N_NN)

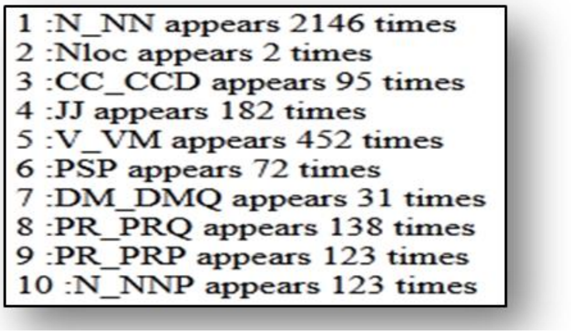
This way we have derived the probabilities of all Bi-grams that are unique throughout the patterns. Bi-gram N_NN N_NN appears highest 1147 number of times. We have got 199 such unique Bi-grams. Fig 4 showcases some unique Bi-grams and their corresponding frequencies.



```
N_NN Nloc appears 2 times

N_NN N_NN appears 1147 times

N_NN CC_CCD appears 55 times

CC_CCD N_NN appears 67 times

N_NN JJ appears 100 times
```

Figure 4: Unique Bi-grams with their frequecies

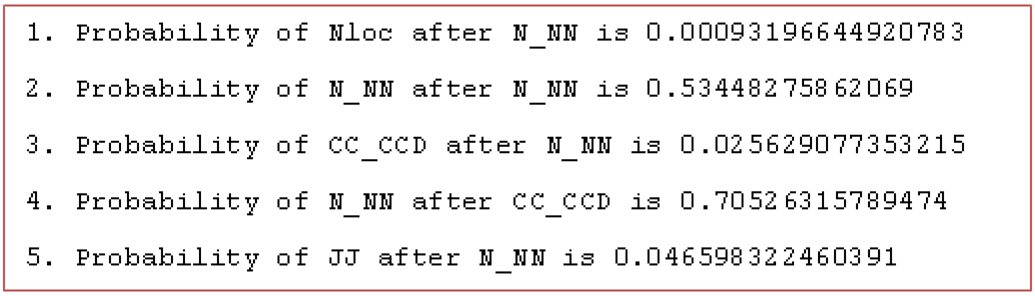
Another important task that we need to perform is to determine the number of times the unique tags appear. We found 30 such tags with N_NN appearing 2146 number of times. The next figure showcases some unique PoS tags with their frequencies of appearances in the corpus.



```
1 :N_NN appears 2146 times
2 :Nloc appears 2 times
3 :CC_CCD appears 95 times
4 :JJ appears 182 times
5 :V_VM appears 452 times
6 :PSP appears 72 times
7 :DM_DMQ appears 31 times
8 :PR_PRQ appears 138 times
9 :PR_PRP appears 123 times
10 :N_NNP appears 123 times
```

Figure 5: Unique PoS tags with their frequencies

As both our unique Bi-grams and unique PoS tags are ready with their respective frequencies, we can start generating the probabilities of the Bi-grams appearing in the patterns. We applied the probability function and tried to determine the probabilities of the Bi-grams as shown in fig 6.



```
1. Probability of Nloc after N_NN is 0.00093196644920783

2. Probability of N_NN after N_NN is 0.53448275862069

3. Probability of CC_CCD after N_NN is 0.025629077353215

4. Probability of N_NN after CC_CCD is 0.70526315789474

5. Probability of JJ after N_NN is 0.046598322460391
```

Figure 6: Probabilities of individual Bi-grams

These probabilities can be depicted through graphs. We will take some samples of Bi-gram probabilities and then try to exhibit them through a graph. The next diagram shows the graphical depiction of the probabilities of the occurrences N_NN after other PoS tags.

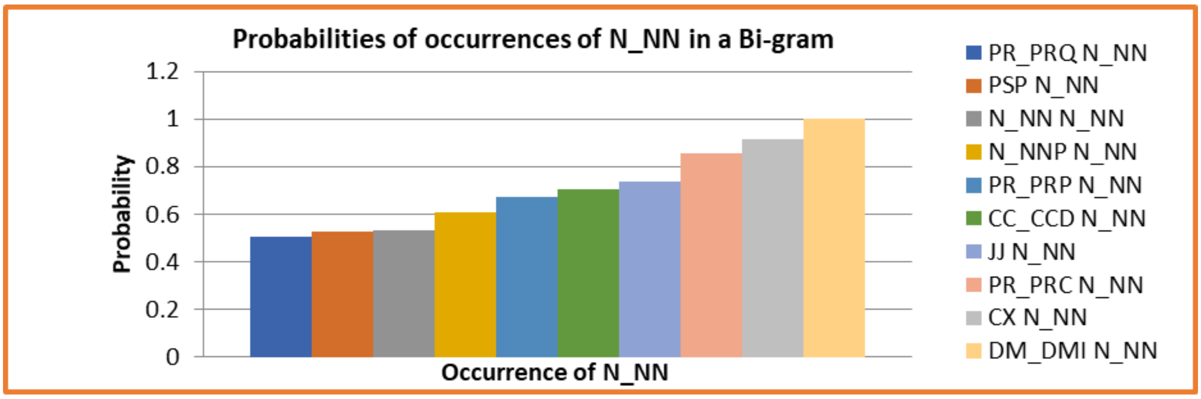
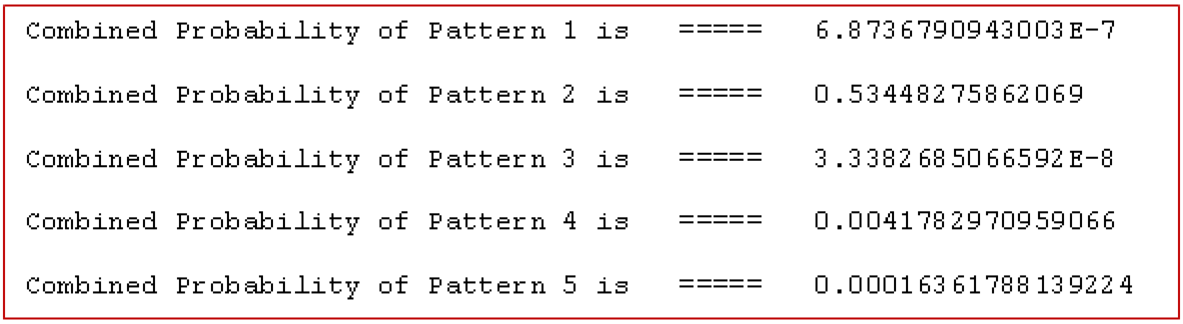


Fig 7: Graphical representation of Bi-gram Probability

Probabilistic representation provides the likelihood in which a Bi-gram occurs. A probability of 1 indicates that an event will certainly occur and a 0 indicates that an event will certainly not occur. Any number between 0 and 1 indicates some degree of uncertainty.

1) *Combined Bi-gram Probability*

Finding the most likely PoS sequence of a sentence in terms of the Bi-gram probabilities plays a crucial role. We also have calculated the combined Bi-gram probabilities of all the Bi-grams occurring in each question pattern. It is generated by multiplying the probabilities of each Bi-gram. A depiction for the same is shown in fig 8:



```
Combined Probability of Pattern 1 is   =====   6.873679094300 3E-7

Combined Probability of Pattern 2 is   =====   0.53448275862069

Combined Probability of Pattern 3 is   =====   3.3382685066592E-8

Combined Probability of Pattern 4 is   =====   0.0041782970959066

Combined Probability of Pattern 5 is   =====   0.00016361788139224
```

Figure 8: Combined Bi-gram probabilities of question patterns

Given the Bi-gram probabilities, we can generate the approximation of the probabilities of the sequences of question patterns. Later, these values can be fed into a machine learning system which would work as an approximation to deduce other question patterns.

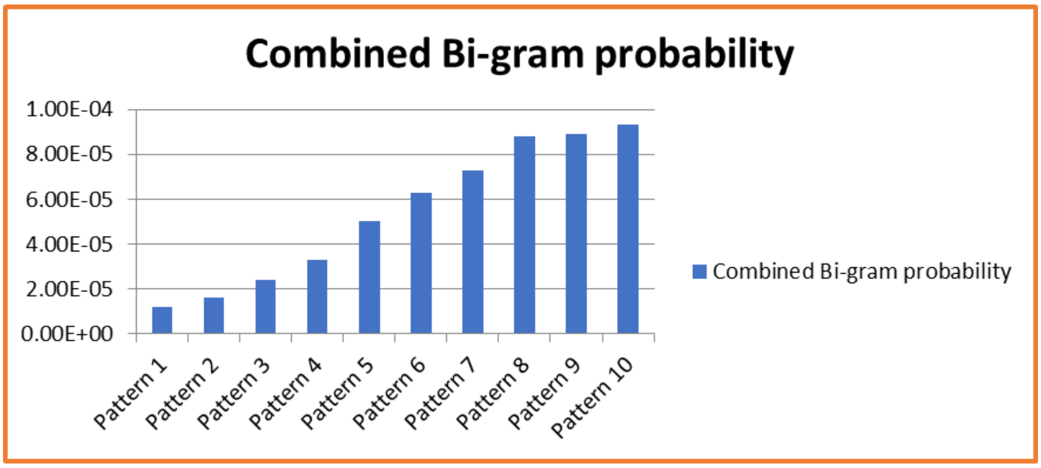
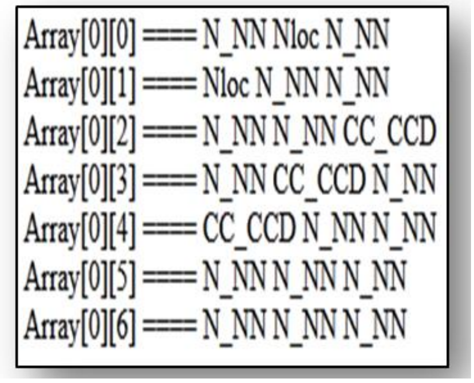


Figure 9: Combined Bi-gram Probabilities of question patterns

Thus, a graphical representation well establishes how likely a Bi-gram would happen in a sequence of patterns. Similarly, we can also showcase the combined Bi-gram probabilities of question patterns occurring in a text corpus. Such representation exhibit the likelihood that a question pattern occurs in a corpus.

B. Tri-gram Probability

Bi-gram information can be useful for deriving Tri-gram probabilities. We have derived the unique Tri-grams appearing in all positions. For instance, considering the Tri-gram pattern of fig 10, we can determine the probability in which N_NN occurs after N_NN Nloc.



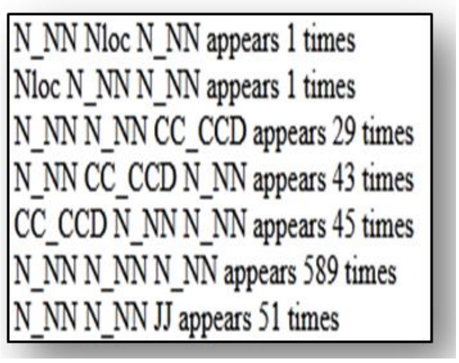
```
Array[0][0] ==== N_NN Nloc N_NN
Array[0][1] ==== Nloc N_NN N_NN
Array[0][2] ==== N_NN N_NN CC_CCD
Array[0][3] ==== N_NN CC_CCD N_NN
Array[0][4] ==== CC_CCD N_NN N_NN
Array[0][5] ==== N_NN N_NN N_NN
Array[0][6] ==== N_NN N_NN N_NN
```

Figure 10: A Tri-gram pattern

We apply the approximation rule 2 in order to find the probability. The probability function would be:

P (N_NN | N_NN Nloc) = C (N_NN Nloc N_NN) / C (N_NN Nloc)

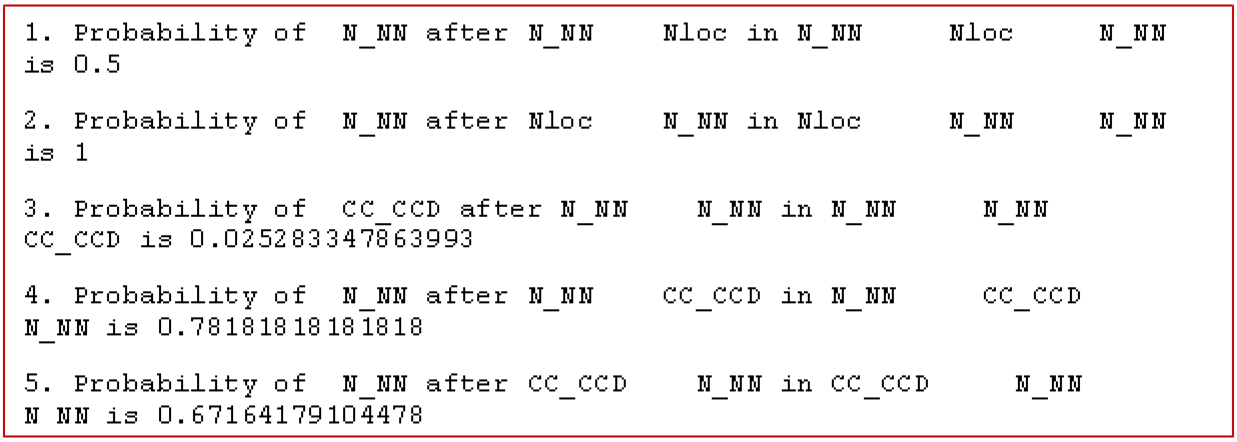
We move on approximating the probabilities of each unique pattern. We have got 500 unique Tri-gram patterns with N_NN N_NN N_NN appearing the highest 589 number of times. The next figure showcases this:



```
N_NN Nloc N_NN appears 1 times
Nloc N_NN N_NN appears 1 times
N_NN N_NN CC_CCD appears 29 times
N_NN CC_CCD N_NN appears 43 times
CC_CCD N_NN N_NN appears 45 times
N_NN N_NN N_NN appears 589 times
N_NN N_NN JJ appears 51 times
```

Figure 11: Unique Tri-grams with their frequecies

We now have to find out the occurrences of unique Bi-grams in order to execute the Tri-gram probability function. We already have derived that in Bi-gram probability. The approximation function derives the probabilities of all Tri-grams occurring throughout the question sequences. Fig 12 showcases the probabilities of Tri-grams.



```
1. Probability of  N_NN after N_NN    Nloc in N_NN     Nloc     N_NN
is 0.5

2. Probability of  N_NN after Nloc    N_NN in Nloc     N_NN     N_NN
is 1

3. Probability of  CC_CCD after N_NN    N_NN in N_NN     N_NN
CC_CCD is 0.025283347863993

4. Probability of  N_NN after N_NN    CC_CCD in N_NN     CC_CCD
N_NN is 0.781818181818 18

5. Probability of  N_NN after CC_CCD    N_NN in CC_CCD     N_NN
N NN is 0.67164179104478
```

Figure 12: Probabilites of individual Tri-grams

We again would like to depict the Tri-gram probability using a graph. It is an easier way to understand and illustrate a huge volume of data through graphs. Data can be compared and visually analyzed through graphical depiction. The following graph exhibits the occurrence after other PoS tags.

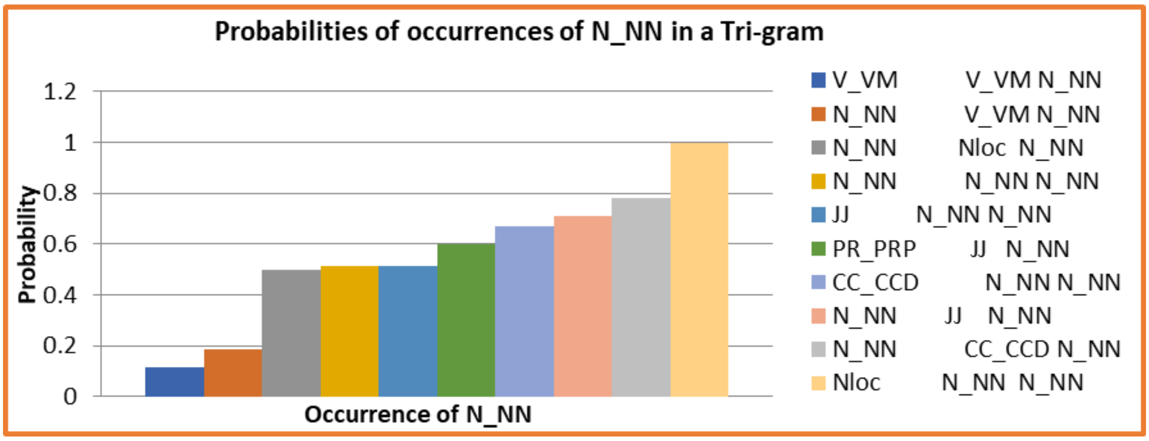
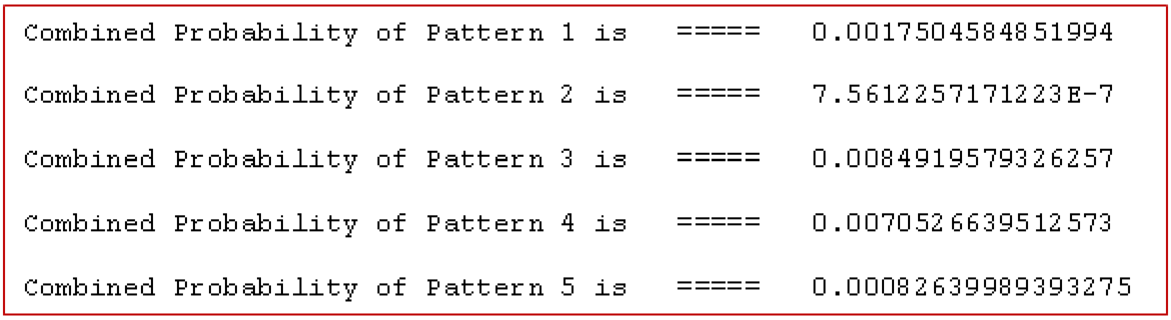


Figure 13: Graphical representation of Tri-gram Probability

Such kind of demonstration will surely provide an insight into how likely a Tri-gram may occur in a particular sequence of Assamese question patterns. These will also form a valuable input into various machine learning based activities.

1) Combined Tri-gram Probabilty

Finally, like Bi-grams, we have generated the combined Tri-gram probabilities of all Tri-grams existing throughout the question sequences. The combined Tri-gram is useful for predicting whether an inputted pattern is assertive or a question. Fig 14 showcases the combined Tri-gram probabilities of question sequences.



```
Combined Probability of Pattern 1 is   =====   0.0017504584851994

Combined Probability of Pattern 2 is   =====   7.5612257171223E-7

Combined Probability of Pattern 3 is   =====   0.0084919579326257

Combined Probability of Pattern 4 is   =====   0.0070526639512573

Combined Probability of Pattern 5 is   =====   0.00082639989393275
```

Figure 14: Combned Tri-gram probabilities

Similar to Bi-grams, the combined Tri-gram probabilities of all question patterns has been depicted in fig 15. This graphical representation is expected to provide some idea like in what probability a question pattern occurs in Assamese text.

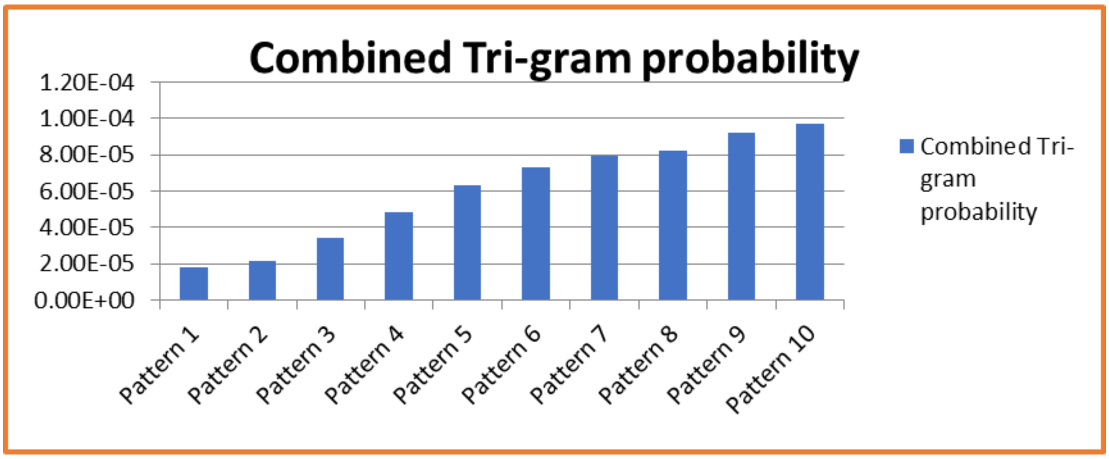


Figure 15: Combined Tri-gram Probabilities of question patterns

Such graphical demonstration establishes the likelihood that a particular Tri-gram would happen in a sequence of Assamese question pattern. Similarly, we can also showcase the combined Tri-gram probabilities of question patterns occurring in a text corpus. Representation of such kind exhibits the probability that a question pattern occurs in a corpus.

V. Novelty of our work

Though Assamese is a linguistically rich language, tools and technologies developed for this language have been very few. Our intention is to make the language digitally rich. Our work is a work of its own kind. Researches on Assamese question patterns were never done before. We have an Assamese corpus with mixed sentences containing both assertive and question. The corpus is tagged with BIS Tagsets. We have generated 511 unique PoS tagged question sequences.

We have derived the probabilities of Bi-grams and Tri-grams of question patterns in Assamese language. These are useful to build applications based on probabilities. They can provide information about the most probable N-grams (Bi-grams and Tri-grams) occurring in a question. They also give an insight into the completeness of sentences; that is, how sentences could be formed. Based on this probabilistic statistical model, it can be determined whethera sequence of words may occur after another sequence. Predictions of such kind may be useful in improving the performances text completion system, speech recognition system and auto correct system. Apart from improvement, the N-gram model has the advantage of being scalable. We can build a larger system which can store more probabilities as well as contexts. As the knowledge base gets bigger, more accurate predictions could be achieved. Subsequently, a more accurate and larger test set could also be gained.

Apart from these, our work has derived the combined N-gram (Bi-gram and Tri-gram) probabilities of the patterns. This is an approximation which measures possibility of occurrence of a question sequence. This in turn enriches the training corpus. Given a test set of mixed sequence of sentences, question patterns are generated if there are matches. Otherwise, the question pattern will be pushed into the training set. Eventually, the training size will get increased and we will get more probable patterns. Resources of such kind may be useful to implement a question-answering system. Since, we are working on patterns; these may be useful for generating answers to questions through some phenomenons like pattern-substitution method or structured text.

We are expecting that our work would revolutionize the domain of natural language processing in Assamese language. Since, we do not have any system which mainly deals with question patterns; we believe that the novelty of our work would form the foundation for doing more complex researches in Assamese language.

VI. Future scope and Conclusion

Our next plan is to apply machine learning approachesfor generation of question patterns. We have a training set containing the patterns we have generated so far. We also have the probabilities for each N-gram separately and combinedfor each pattern. Our intention is to find out more question structures based on the probabilities we have. We plan to apply machine learning practices which would best approximate those question patterns. Our project mainly emphasizes on finding the question structures in Assamese language as the structures are PoS tagged.

NLP has been a significant area of research in recent times. Our work aims at doing NLP research in Assamese language. Faster digital revolution facilitates faster social development. Assamese is a new language for digital revolution. The language is on high demand for putting it on digital platform since it is understandable to all as a medium of communication. Our work is a first ever intended work of doing NLP researches on Assamese question patterns. We hope that our work would provide newer tools and technologies which might eventually help in exploring the unexplored areas of Assamese language. We also expect that our work would form the foundation for doing Artificial Intelligence based research works.

References

[1] Stanojević, M., & Vraneš, S. (2012). Representation of texts in structured form. *Computer Science and Information Systems*, *9*(1), 23-47.

[2] Liddy, E. D. (2001). Natural language processing.

[3] Chakraborty R. (2017). *Structured representation of Assamese text* [Doctoral thesis, Gauhati University].http://hdl.handle.net/10603/200080.

[4] Roy, B. Parsing and Part of Speech Tagging For Assamese Texts, http://hdl.handle.net/10603/369761, (2017).

[5] Sarma, Bharali, Gogoi, Deka, Barman., A Structured Approach for Building Assamese Corpus: Insights, Applications and Challenges. ALR@COLING 2012: 21-28.

[6] Jurafsky, D., & Martin, J. H. (2018). N-gram language models. *Speech and language processing*, *23*.

[7] Fragos, K., & Skourlas, C. (2006). An N-gram Based Distributional Test for Authorship Identification. In *NLUCS* (pp. 139-148).

[8] Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175).

[9] Brill, E., Dumais, S., & Banko, M. (2002, July). An analysis of the AskMSR question-answering system. In *Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing (EMNLP 2002)* (pp. 257-264).

[10] Gómez Soriano, J. M., Montes y Gómez, M., Sanchis Arnal, E., & Rosso, P. (2005). A passage retrieval system for multilingual question answering. In *Text, Speech and Dialogue: 8th International Conference, TSD 2005, Karlovy Vary, Czech Republic, September 12-15, 2005. Proceedings 8* (pp. 443-450). Springer Berlin Heidelberg.

[11] Brown, P. F., Della Pietra, V. J., Desouza, P. V., Lai, J. C., & Mercer, R. L. (1992). Class-based n-gram models of natural language. *Computational linguistics*, *18*(4), 467-480.

[12] Lioma, C., & van Rijsbergen, C. K. (2008). Part of speech n-grams and information retrieval. *Revue française de linguistique appliquée*, (1), 009-022.

[13] Buscaldi, D., Rosso, P., Gómez-Soriano, J. M., & Sanchis, E. (2010). Answering questions with an n-gram based passage retrieval engine. *Journal of Intelligent Information Systems*, *34*, 113-134.

[14] McDonald, G., Macdonald, C., & Ounis, I. (2015, September). Using part-of-speech n-grams for sensitive-text classification. In *Proceedings of the 2015 International conference on the theory of information retrieval* (pp. 381-384).

[15] Sidorov, G., Gupta, A., Tozer, M., Catala, D., Catena, A., & Fuentes, S. (2013, August). Rule-based system for automatic grammar correction using syntactic n-grams for english language learning (l2). In *Proceedings of the Seventeenth Conference on Computational Natural Language Learning: Shared Task* (pp. 96-101).

[16] Ferrandez, A. (2011). Lexical and syntactic knowledge for information retrieval. *Information processing & management*, *47*(5), 692-705.

[17] Cavnar, W. B., & Trenkle, J. M. (1994, April). N-gram-based text categorization. In *Proceedings of SDAIR-94, 3rd annual symposium on document analysis and information retrieval* (Vol. 161175, p. 14).

[18] Sarma, S. K., Bharali, H., Gogoi, A., Deka, R., & Barman, A. (2012, December). A structured approach for building Assamese corpus: insights, applications and challenges. In *Proceedings of the 10th workshop on Asian language resources* (pp. 21-28).

[19] Barman, A. K., Sarmah, J., & Sarma, S. K. (2013, April). Pos tagging of Assamese language and performance analysis of CRF++ and FNTBL approaches. In *2013 UKSim 15th International Conference on Computer Modelling and Simulation* (pp. 476-479). IEEE.

[20] Sarmah, J., Saharia, N., & Shikhar, K. (2012). A novel approach for document classification using Assamese wordnet. In *6th International Global Wordnet Conference* (pp. 324-329).

[21] Gogoi, A., Baruah, N., Sarma, S. K., & Phukan, R. D. (2021). Improving stemming for Assamese information retrieval. *International Journal of Information Technology*, *13,* 1763-1768.

[22] Talukdar, G., Borah, P. P., & Baruah, A. (2018). Assamese Named Entity Recognition System Using Naive Bayes Classifier. In *Advances in Computing and Data Sciences: Second International Conference, ICACDS 2018, Dehradun, India, April 20-21, 2018, Revised Selected Papers, Part I 2* (pp. 35-43). Springer Singapore.

[23] Allen, J. (1995). *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc..