Structured and Logical Representations of Assamese Text for Question-Answering System

Rita CHAKRABORTY Shikhar Kr. SARMA
DEPARTMENT OF INFORMATION TECHNOLOGY, GAUHATI UNIVERSITY
Guwahati, Assam, India 781014
ritachk@rediffmail.com, sks001@gmail.com

ABSTRACT

Written documents contain information in a language specific syntax form. Computational processing of such information demands representation in a structured form suitable for handling, processing, and analyzing. Such structured representation of documents enables extraction of knowledge through computational means. Once the textual data are represented in structured form, logical representation also becomes easier. This paper discusses our work on analyzing texts in Assamese language and processing those texts in terms of converting into structured and logical representations. Our emphasis is on the Structured Representation of texts and current study focuses on providing the architecture and processing workflow of the system to output structured form of Assamese text. It also includes system design discussions on how these representations of texts in Assamese language will contribute towards building a Question-Answering system.

KEYWORDS: Structured Representation, Logical Representation, Question- Answering System

Proceedings of the Workshop on Question Answering for Complex Domains, pages 27–38,
COLING 2012, Mumbai, December 2012.

1. Introduction:

Written documents such as text documents, web pages and books contain information in a language specific syntactic form, not suitable for automatic processing through computers. Therefore, this textual information must be represented in such a manner so that analyzing and processing of these texts becomes easier [5][6]. This representation also makes automatic knowledge extraction possible. Assamese is a new language for which NLP research works have recently started analyzing and building various automated computational models. Many research works are going on for other naturally occurring languages like English, Bodo etc. Various automated tools and techniques have also been developed for these languages. The proposed research work is a new domain of research in Assamese language for getting an insight into how sentences in Assamese can be analyzed, parsed and interpreted. The structured text representation shows the relationship among the constituents of a sentence. It will provide the basis for doing higher level projects such as building Question-answering systems [1]. The information gained from this kind of representation will also pave the way for further research works related to cross-lingual information system, automatic text extraction, mining information etc [3]. This research work is expected to give syntactic and semantic characteristics of Assamese language in the perspective of computational linguistics [1][5]. We are also expecting to get a POS tagged Assamese corpus as well as computational modules for building structured and logical representations of Assamese text corpus.

Development of regional languages has been a great concern now a days. This is due to the fact that these languages are getting more demands for putting them as a medium of communication of the digital world. Researchers and government agencies have started their effort to design and develop different technologies for putting those regional languages into the digital world. Research and development of various language technologies have also started for Assamese language, which is recognized as a scheduled language of Indian constitution. Technologies like UNICODE compliant fonts and keyboards, automated spell checker have already been developed for this language [9]. Research works are also going on for developing various technologies for Assamese language as part of the efforts on technology development of Indian languages.

Assamese is a new language to digital revolution. This language is also used as a medium of communication within the states of North-East India, especially in Arunachal Pradesh and Nagaland. A huge population of India speaks Assamese in different parts of the nation who originally belong to Assam. Assamese speaking people can also be found in some nations like, Bhutan and Bangladesh. Tentatively, about 14 million people speak Assamese in the state of Assam and its neighbouring states and about 14.3 million Assamese speaking people can be found in all over India [9].

The origin of Assamese language can be derived from its relation with Indo-Aryan group of languages and a little bit with Sino-Tibetan group of languages. Apart from Assamese, languages like Bangla, Oriya, Hindi etc. also fall into the category of Indo-Aryan group[9]. These languages have similarity with Assamese language. However, differences may also exist among these languages; still, the technologies developed for Assamese are expected to be able to provide an insight into the development of similar

kind of technologies for these languages. In this paper, we are focusing on the representations of Assamese text in terms of the structured and logical formalisms. The technologies built in this regard will hopefully be able to represent knowledge in structured and logical forms for other Indo-Aryan languages. This will happen due to the fact that these languages have similarity with Assamese language. Therefore, we expect to design and develop an automated model which will also be used for developing technologies for other similar languages.

In order to analyze a corpus based text in terms of computational linguistics, it must pass through different phases like- morphological, syntactical, semantics, pragmatic etc. The morphological analysis analyzes individual word and non-word tokens such as punctuation markers. These non-word special markers must be separated in this phase. In the syntax analysis part, the tokens generated from the morphological analysis are transformed into structures showing the relationship among the tokens [1][4]. This relationship must follow the grammatical rules of the language. If a combination does not follow any rule, that sequence must be rejected. The structures created by the syntactic analyzer are assigned meanings at the semantic level. The ambiguity of sentences must be resolved in this phase [1][4][3]. We are basically concentrating on syntactic level analysis and partially on semantics. The extracted knowledge using these two representations provides information in terms of grammatical structures of the language. They also provide the scope of doing analysis in semantic levels so that the meaning of the sentences can be interpreted. This paper is organized in the following way- section 2 gives an idea of related topics which outlines the idea of structured and logical representations. Section 3 describes the overall planning of the project work. Section 4 provides an outline of analysis of sentences written in Assamese. Section 5 describes the proposed model for question-answering system which also outlines an idea of Assamese question pattern and section 6 is the proposed conclusion.

2. Related topics:

2.1 Structured Text:

The context of the input text must be represented in structured text format. Structured text describes the individual objects occurring in the sentences. It attempts to capture the knowledge contained in the text essential for doing various kinds of operations. Things that are not mentioned explicitly in the given text such as references to pronouns are made explicit here. As a whole, it can be said that the context of the sentences are represented using structured text [1]. To show this, Let us consider the following English sentence-

I got the red ball that I wanted.

This sentence can be represented in structured form in the following manner-

Event-1

Instance -	Get
Tense-	Past
Agent-	I
Object-	Thing

Thing
Instance- Ball
Color- Red
Event-2
Instance - want
Tense- Past

One of the key ideas of such kind of representation is to find out the meanings of the objects with reference to their connections to other objects. Such kind of representation can also be termed as slot-and-filler structure [1]. The information gained from such kind of structure represents knowledge in terms of syntactic level. It operates as a mechanism to see whether these structures conform to the rules or syntax of Assamese language.

2.2 Logical Representation:

The structured information forms the basic building block for knowledge acquisition formalism. The information acquired in structured text representation can be used to represent knowledge in logical formalism also. Logical representation can also be implemented to acquire new knowledge from old. It guarantees that a new statement can be proved to be true because the statement follows from some already proved statements [1]. Such kind of representation can be gained through First Order Predicate Calculus (FOPC) [2]. Basically the facts and rules in logical representation can be expressed in FOPC using PROLOG. The well formed formulas representing the facts and rules should be written in Clause form only. The clausal notations can be implemented in question answering systems also. Such representation can be used to answer not only yes-no questions but also fill-in-the blank questions [1]. To show this, let us consider a sentence in English – "Rina eats mango" can be represented in PROLOG as –

eat (Rina,mango).

Now, if we have a query like –

?eat (Rina,X) gives the answer **X=mango**.

The proof procedure of PROLOG follows Resolution principle where the proof is generated through backward chaining process [1][2]. Actually, the process of deriving answers to questions using logic is based on Matching technique. Matching takes two terms as input and checks to see whether they match. If they match, the process produces a success signal. Using matching process, variables can also be bound to values if necessary [2]. For e.g. the terms date(1, may, 2005) and date(D,M,Y) match. The results of this matching process is –

D=1
M=may
Y=2005

Both structured text and logical representations for Assamese language have been discussed later in this paper.

3. Proposed System:

The proposed research work is just the preprocessing phase of doing NLP research in Assamese language. We have planned to divide the whole project work into two primary modules- The Preprocessor module and the Structured Text Generator module. These two modules are again subdivided into some sub modules. The following diagram is a structural representation of the proposed system-

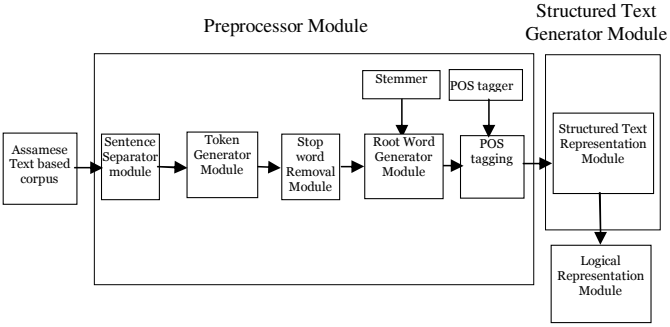


FIGURE: Structural / Diagrammatic representation of the proposed system

As shown in the diagram, an Assamese text based corpus will be taken as the input to the system. The Preprocessor module begins with the *Sentence Separator* module which separates the sentences basic to the text corpus. After that the *Token Generator* module begins which extracts individual tokens from the sentences. Next step is to remove the stop words such as the punctuation markers or conjunctives from the tokens generated so far using the *Stop Word Removal* module. At this stage, we get the tokens which are actually taking part in that particular context. Then we pass these tokens through an automated stemmer which generates the root morphemes behind every token. This is done by the *Root Word Generator* module [7]. After this, the root morphemes are annotated in terms of Parts-of-Speech (POS) tagging. Using a POS tagger, each morpheme will be tagged and these annotated morphemes will be used in the later processing of the text. These whole set of operations may be regarded as the morphological analysis in terms of computational linguistics.

Next module is the Structured Text Generator module which generates structured text from the outputs of the Preprocessor module. The annotated lexicons provide the information about subject(s), verb(s), instance(s) and object(s) in the sentences. Structured text representation will be generated based on this information. This same information can be used in generating logical representation also.

Automated validators may be required for testing these representations in terms of linguistics.

4. Analysis of Sentences Written in Assamese:

Sentences in Assamese are the well-organized sequence of parts-of-speech and inflections. Therefore, the syntax of a sentence can be termed as the rule-based implementation of inflections to the parts-of-speech of the sentence. The structure of sentences can also be termed as the structure of the language [8].

Sentences in Assamese basically follow the Subject+Object+Verb form. Actually this is the structure of a simple sentence. For e.g.

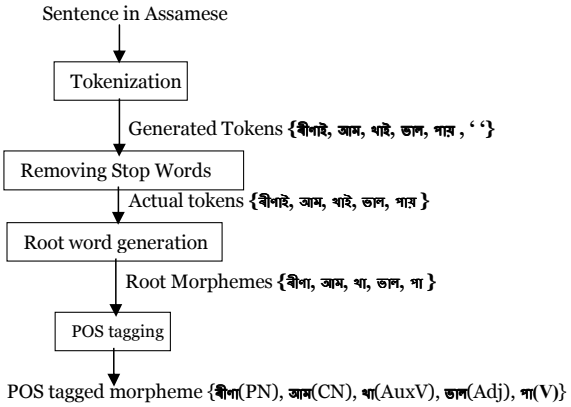
মই কিতাপ পঢ়োঁ (I read book)

Sentences may be compound or complex also. Verbs play an important role in such sentences because they help in achieving the idea about the meaning of the sentences. They have direct relationship with all the cases except case 6, that is, genitive case. The association between subject and verb make an Assamese sentence complete. Verb along with all cases except the subjective case occur in predicate part of an Assamese sentence [8]. The inflections should be incorporated into the words in a proper manner so that the actual interpretation of the sentences can be gained.

In order to analyze sentences written in Assamese, one must possess the knowledge of parts-of-speech of this language. In this section, we are producing a sentence level analysis in Assamese which has relevance with our work model.

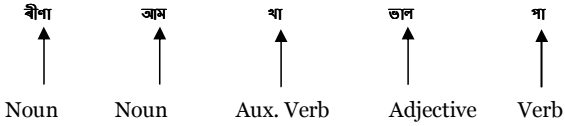
Let us consider the same sentence- **বীণাই আম খাই ভাল পায়**

The dataflow model for parsing this sentence is as follows-



The outputs of POS tagging phase are the tagged morphemes of the root words [7].

The tagged morphemes of the assumed sentence are as follows-



The POS tagged morphemes are passed through the syntax analysis phase where a graphical representation or a parse tree will be constructed on these morphemes. The hierarchical structure must obey the grammatical rule of the language. The output of the syntax analysis phase i.e. the parse tree for the particular sentence is as follows-

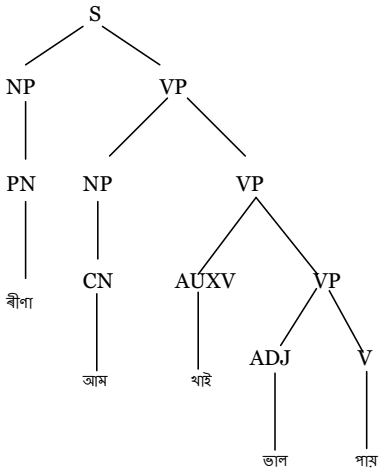


FIGURE : Parse tree for the Assamese sentence "বীণাই আম খাই ভাল পায়"

In this way, the sentences in Assamese will be analyzed and parsed to examine whether they follow the correct syntactic structure of the language. Any sentence structure which does not follow any syntax of the language should never be accepted.

5. Model for Question-Answering System:

5.1 Assamese Question Pattern:

Interrogative sentence or question pattern is one of the types of sentences in Assamese language. There are two factors playing important role in Assamese question patterns-parts-of-speech and rhythm in which the question is being asked. The rhythm of the question defines the kind and manner of the response being sought [8]. Some questions

may be asked using do-verb in different forms. An example of a simple interrogative sentence using do-verb in present tense, second person singular number is given below-

তুমি সদায় পঢ়া লে? (Do you read always?)

Questions may be formed using কি (What), কিয় (Why), ক'ত (Where), কেতিয়া (When), কাৰ (Whose), কেনেকৈ (How), কিমান (How Much) etc. Examples of simple question pattern of such kind may be-

তোমাৰ নাম কি ? (What is your name?)

তেওঁ ক'ত থাকে ? (Where does he live?)

Questions patterns may be complex also. These types of interrogative sentences may be split up into more than one simple questions or sentences. For e.g.

তুমি ভাত খালানে নাই মই নাজানোঁ । (I do not know whether you have eaten rice or not)

This sentence can be split up into-

মই নাজানোঁ । (I do not know)

তুমি ভাত খালানে ? (Have you eaten rice?)

In Assamese, the same question pattern may be asked in different forms. The pattern may be different, but the meaning of the question remains same. Let us consider the following two examples to understand this.

ৰীণাই কি খাই ভাল পায় ?

কি খাই ভাল পায় ৰীণাই ?

In this way, different question patterns generate the same semantic structure as well as same answer set.

We can also cite the example of rhetorical question where the answer is implicit in the question itself. Although asked in the form of a question, but the semantic structure generates the implicit answer. Such question patterns are basically used in Assamese to enhance the literary quality of Assamese language.

5.2 Proposed Model:

The proposed model for question-answering based on structured representation works in the following way.

Suppose we have the following Assamese sentence -

ৰীণাই আম খাই ভাল পায় (In English, Rina likes to eat mango)

This sentence can be represented in terms of structured text as given below-

Event-1

Instance -	খা	(To eat)
Tense-	বৰ্তমান	(Present)
Agent-	ৰীণা	(Rina)
Object-	আম	(Mango)

Event-2

Instance -	পা	(To get)
Tense-	বৰ্তমান	(Present)
Modifier-	ভাল	(Like)

Again suppose, we want a response to the question-

ঝীণাই কি খাই ভাল পায়? (What does Rina like to eat?)

The answer should be - **আম** (Mango)

To get the answer, we again have to convert the question into structured form [1]. The structured text for the question is represented as follows-

Event-1

Instance -	খা	(To eat)
Tense-	বর্তমান	(Present)
Agent-	ঝীণা	(Rina)
* Object-	কি	(What)

Event-2

Instance -	পা	(To get)
Tense-	বর্তমান	(Present)
Modifier-	ভাল	(Like)

The part of the structure serving as the answer should be marked. Often these markers correspond to the question words "who" or "what" in the sentence [1]. This structured text for the question will be matched against the structured text generated above. The response is generated based on the segments of the structured text that match the segments of the structured question being asked.

Similarly, the same sentence (as above) can be considered to represent knowledge using logical formalism also. For this, we have to take into consideration the parse tree generated (as in P7) for that sentence. Using this tree structure, logical rules can be derived. These rules of inference can then be used perceive answers to questions. Basically, the representation of rules can be done using First Order Predicate Logic. According to the parse tree, the logical rules of the sentence would be-

S -> NP VP

NP -> PN | CN

VP -> NP VP | AUXV VP | ADJ V

PN -> ঝীণাই

CN -> আম

AUXV -> খাই

ADJ -> ভাল

V -> পায়

As PROLOG structure, the tree can be represented as follows-

S (NP(PN(ঝীণাই)), VP(NP(CN(আম)), VP(AUXV(খাই), VP(ADJ(ভাল), V(পায়)))))

Similarly, if the same question (as above) is asked, then the question may also be transformed into a similar logical representation as shown above. It would be like as follows-

?- S (NP(PN(ৰীণাই)), VP(NP(X), VP(AUXV(খাই), VP(ADJ(ভাল), V(পায়)))))
Then the answer to the question representing the value of X is returned as - আম (Mango)
In this way, we can generate answers to questions from a given set of predicate logic statements using matching process [2].

The structures generated using these two representations may pave the way for doing similar kind of research in other languages also. Languages like Bangali, Oriya or Hindi fall into the category of Indo-Aryan group of languages to which Assamese language also belongs. Therefore, tools developed for one language may help in generating similar kind of tools for other related languages also.

6. Conclusion:

Natural Language Processing has been a significant area of research in recent years. Digital revolution is penetrating in the grassroots level facilitating social development in a faster way. Assamese is a new language for digital revolution. Research works have started for design and development of tools and technologies for this language. Our proposed work will facilitate the preprocessing phase of NLP research in Assamese language. Basically, we have planned to work on syntactic level analysis which will help in automatic knowledge acquisition in terms of linguistics. Our project is the first ever intended work for giving structured and logical representations in Assamese language. As the language is becoming richer for digital revolution, newer applications are becoming possibilities for future understanding of the unexplored areas as well as intricacies of Assamese language. I visualize this work will also pave the way for Artificial Intelligence research works in this language.

References:

[1] Rich Elain, Knight Kevin (1991*). Artificial Intelligence*, Tata McGraw Hill, New Delhi.

[2] Bratko Ivan. *PROLOG Programming for Artificial Intelligence*, Pearson Education.

[3] Chowdhury Gobida G. "*Natural Language Processing*". http://www.cis.strath.ac.uk/cis/research/publications/papers/strath_cis_publication_320.pdf

[4] http://www.cnlp.org/publications/03nlp.lis.encyclopedia.pdf.

[5] Stanojevic Mladen, Vranes Sanja. "*Representation of Texts in Structured Form*". http://www.comsis.org/archive.php?show=ppr275-1009

[6] Costantini Stefania, Florio Niva, Paolucci Alessio. "*A framework for structured knowledge extraction and representation from natural language via deep sentence analysis*". *ceur-ws.org/Vol-810/paper-l18.**pdf***

[7] Bora Lilabati S.(2006). "*Asomia Bhasar Rupatattwa*", Banalata, Panbajar.

[8] Goswami Golak C. (2008). "*Asomia Byakoronor Moulik Bisar*", Bina Library, Guwahati.

[9] Sarma Kr. Shikhar, Gogoi Moromi, Saikia Utpal, Medhi Rakesh. "Foundation and Structure of Developing an Assamese Wordnet" http://www.cfilt.iitb.ac.in/gwc2010/pdfs/50_Assamese_Wordnet__Sarma.pdf