MODELLING A QUESTION-ANSWERING SYSTEM USING STRUCTURED REPRESENTATION OF ASSAMESE TEXT

Rita Chakraborty and Shikhar Kr. Sarma

Department of Information Technology, Gauhati University
Guwahati, Assam, India 781014

ABSTRACT

Information or knowledge contained by the texts is structured in a language specific syntactic form. They are neither understood nor processed by the computers. They must be organized in a structured form. Structured representation of sentences written in a particular language enables a computer to have a good understanding of the knowledge they contain. Such type of representation is also useful for modelling systems like information retrieval, question-answering, machine learning etc. This paper focuses on extracting knowledge from sentences in a form convenient for manipulation through computers. We are basically concerned with representing these structures of sentences written in Assamese language. Moreover, we have also tried to extract answers to questions from these generated structures. Our primary goal of research is to design and develop a question-answering system for Assamese language which is expected to bring a new era of digital renaissance in the field of Artificial Intelligence for this language.

KEYWORDS

Structured Representation, Question- Answering System, Machine learning, Information Retrieval, POS tagger, Assamese language.

1. INTRODUCTION

Written languages contain information or knowledge which is not suitable for processing through computers. Therefore, they must be organized in a manner so as to achieve direct manipulation. There are several techniques through which knowledge extraction is possible. One such technique is the structured representation of texts contained in the form of written documents [1][2]. Moreover, such kind of representation enables handling, processing and understanding the meanings of sentences [3].

Our paper aims at building the structured representation of texts written in Assamese language only. Such kind of representation will hopefully be able to pave the way for new research areas like question answering for this language [4]. Our work also provides a way of getting an insight into how sentences of Assamese language are constructed as well as, a better understanding of the syntax and semantics of the sentences.

DOI : 10.5121/ijaia.2015.6601

There is no common structure present for sentences in Assamese. They may vary according to the sentence form. The more information each sentence contains within it, the corresponding structure also increases. To analyze those sentences, one must be well versed with the vocabulary and grammar of Assamese [5][6]. Finding the correct structure for each sentence is a complex task, because lots of aspects regarding the language must be considered. First and the most important aspect is that, structured representation requires the text document to be tagged with a Parts-of-Speech Tagger (POS tagger). A tagged corpus enables a user to find out the definitive parameters that are actually occurring in the sentence and whose existence can influence the interpretation of the sentence. Apart from this, the tagged words can determine the syntactic structure of a sentence. Moreover, they may be attached with inflections and affixes [5][6]. These are also required to be found out and removed so as get the actual words taking part in the sentence. Apart from definitive parameters, the structure should also contain an extra but a very important property of a sentence- the tense information of the sentence. Existence of such type of information in a structure makes the semantic interpretation of the sentence almost complete. We have discussed briefly a little bit of the sentential structure of Assamese sentences and some of its important issues. We also have discussed the way we have done our work to find the structured representation of Assamese texts.

2. ASSAMESE SENTENCE STRUCTURE

In contrast to English sentences, which follow the Subject+Verb+Object format, sentences in Assamese follow the Subject+Object+Verb format [6]. Assamese sentences are basically simple in nature. We assume the following simple sentence-

তাই কিতাপ পঢ়ে । (*Tai kitap porhe*) (In English, She reads books)

The sentence contains subject as তাই, object as কিতাপ and verb as পঢ়ে. However, two or more simple sentences may be combined to form either compound or complex sentences. An example of such kind may be-

ৰাজেনক এটা ৰঙা আৰু এটা নীলা কলম লাগে । (*Rajenok eta ronga aru eta nila kolom lage*) (In English, Rajen needs a red pen and a blue pen)

A compound sentence in Assamese can be constructed with the help of some connectors like – আৰু, বা, নাইবা, কিন্তু etc.

A structured representation of a simple Assamese sentence resembles the meaning of the sentence at the semantic level. Simple sentence does not require much overhead to find the structure. However, complex sentences require a great amount of effort to be incorporated. As we already have mentioned that to find the structured text of a sentence, each word must be POS tagged properly. Let us see this with help of the following sentence-

মনোজ কলিকতালৈ গল । (*Manoj kolikotaloi gol*) (In English, Manoj went to Kolkata)

Now, the corresponding tagged sentence would be-

মনোজ<NP> কলিকতালৈ<NP> গল <VM>

The actual morphemes are মনোজ, কলিকতা and যা [5]. During sentence formation, the root morphemes are combined with case markers and person markers. The noun morphemes are attached with the case markers and verb roots are combined with person markers. The person markers also indicate which tense form the sentence belongs to. Verbs play a very important role in all Assamese sentences [7].

3. RELATED TOPIC

3.1 Structured Text

Structured text describes each and individual object occurring in a sentence. This representation attempts to catch the internal knowledge contained in a text. Structured representation tries to capture the context or meaning of a sentence. It also tries to determine the object behind the inexplicit things. For example, pronouns contained in a sentence are basically represented as agents in the corresponding structure. However, the value for the agent part is determined by a sentence that precedes the sentence which contains the pronoun as an agent. Research works have been done to find the structured representation of English text [4]. Let's see this with the help of an English sentence-

Sam loved the blue shirt.

When converted to structured form, the representation would be-

Agent - Sam

Object - Shirt

Instance - Love

Modifier - Blue

Tense - Past

The above structure contains agent as the subject of the sentence, which is Sam in this case. The object parameter contains shirt as its value. Similarly, the instance part contains the verb occurring in this sentence. Modifier generally takes the value which qualifies something; in this case it is the shirt. A very important aspect that the structure has taken care of is the tense form of the sentence. This sentence is a past tense sentence and the structure has held it.

4. OUR WORK

The goal of our project work is to achieve semantic interpretation of Assamese sentences. In this paper, we have tried to generate some structures representing the meanings of sentences. The structures generated may be used for modeling a question-answering system and in this paper we

have tried to model that [3][4]. We have considered some Assamese sentences and tried to convert them into structural form. We have named the definitive parameters subject(s), object(s) and verb(s) occurring in the sentence as agent(s), object(s) and instance(s) respectively. We also have introduced other semantic information like tense, number and location. As we already have mentioned, sentences must be POS-tagged prior to generating the structure [7]. In our model, we have proposed our own POS tags to annotate the Assamese words. Some of them are- Noun Common (<NC>), Noun Proper (NP), Verb Main (<VM>), Verb Auxiliary (VA), Adjective (JJ). Now, let's consider the sentence-

1. মানুহজনে<NC> গাড়ী<NC> কিনিলে<VM> । (*Manuhjone gari kinile*)(In English ,The man bought a car)

Structure:

Agent - মানুহ
Object – গাড়ী
Instance – কিন্
Tense - Past

The root words constituting this sentence are – মানুহ, গাড়ী and কিন্ [5]. Apart from these, the tense information has also been kept in the structure. This sentence is in past tense form. Now, if a question is asked like –

কোনে গাড়ী কিনিলে? (*Kone gari kinile?*) (In English, Who bought the car?)

Then, the reply to the question-word "কোনে" (That is, Who in English) is returned by the Agent part of the structure. Thus, the answer would be মানুহজনে.

Again, if a question is asked like –

মানুহজনে কি কিনিলে ? (*Manuhjone ki kinile?*) (In English, What did the man buy?)

The answer to the question-word "কি" (That is, What in English) is provided by the object parameter of the structure. Therefore, the answer would be গাড়ী. Though the corresponding English sentence contains the information about the number of cars that the man has bought, the Assamese sentence does not hold that information. Therefore, by looking at the Assamese sentence, there is no way of knowing how many cars the man has bought. Now, we consider a sentence written in the following manner.

2. মানুহজনে <NC> এখন<JJ> গাড়ী<NC> কিনিলে<VM> । (*Manuhjone ekhon gari kinile*)(In English ,The man bought a car)

Structure:

Agent - মানুহ
Object – গাড়ী
Instance - কিন্
Number - এখন
Tense - Past

This sentence contains one more information - the number information. The number parameter determines the quantity of a particular thing. It also provides the knowledge about the number of cars that the person has bought. The sentence contains এখন as an adjective. But, since we are concentrating on finding the semantic interpretation of the sentence, therefore we can put it into the number parameter. Now, if a question contains the word "কিমান" (That is, How many in English), then definitely the number information is going to provide the answer. Now, let us consider the following question –

মানুহজনে কিমানখন গাড়ী কিনিলে ? (*Manuhjone kimankhon gari kinile?*) (In English, How many cars did the man buy?)

Then the answer would be – এখন

Now, if the same question is asked-

মানুহজনে কি কিনিলে ? (*Manuhjone ki kinile?*) (In English, What did the man buy?)

Then the number and object parameters jointly will provide the answer. Thus the answer would be এখন গাড়ী. We have seen that the number parameter gives more accurate information about the number of cars bought which was not available in the previous structure.

3. মানুহজনে<NC> এখন<JJ> নতুন<JJ> গাড়ী<NC> কিনিলে<VM> । (*Manuhjone ekhon notun gari kinile*) (In English ,The man bought a new car)

Structure:

Agent - মানুহ
Object – গাড়ী
Instance - কিন্
Tense - Past
Modifier - Object1

Object1:

Number - এখন
Modifier - নতুন

Sentences of above type which contain two continuous adjectives can be encapsulated into a single object. If we look this sentence from semantic view point, then definitely এখন provides number information and নতুন acts as a qualifier of the car [4][6]. Now if the following question is asked-

মানুহজনে কি কিনিলে ? (*Manuhjone ki kinile?*) (In English, What did the man buy?)

Definitely, the answer returned for the question-word "কি"(What) is contained by the combined object and object parameter of the structure. Therefore, the answer returned is – এখন নতুন গাড়ী.

4. মানুহজনে<NC> যোৱাকালি<JJ>এখন<JJ> নতুন<JJ> গাড়ী<NC> কিনিলে<VM> । (*Manuhjone jowakali ekhon notun gari kinile*)(In English ,The man bought a new car yesterday)

Structure:

Agent	- মানুহ
Object	– গাড়ী
Instance	- কিন্
Time	- যোৱাকালি
Modifier	- Object1
Tense	- Past

Object1:

Number	- এখন
Modifier	- নতুন

This sentence contains knowledge about the time, when the man had bought the car. Therefore, if the following question is asked-

মানুহজনে কেতিয়া গাড়ী কিনিলে ? (*Manuhjone ketia gari kinile?*) (In English, When did the man buy a car?)

The answer returned for the question-word কেতিয়া (That is, When in English) is represented by the Time parameter of the structure.

Let's assume a different sentence-

5. ৰীমাই <NP> বাহিৰত<NC> ফুৰি <VA> ভাল<JJ> পায়<VM> । (*Rimai bahirot furi bhal pai*) (Rima loves to travel outside)

Structure:

Agent	- ৰীমা
Instance1	- পা
Modifier	- ভাল
Instance2	- ফুৰ্
Location	- বাহিৰ

The structure representing this sentence contains a parameter location, which designates a particular place. Now, if a query contains a question-word কত (That is, Where in English), the location information will give the answer. Thus, if a query is of the following form-

ৰীমাই কত ফুৰি ভাল পায় ? (*Rimai kot furi bhal pai?*) (In English, Where does Rima love to travel?), then the answer would be – বাহিৰত

This sentence contains multiple instances. Similarly, sentences may contain multiple agents or multiple objects. Thus, there may be various types of sentences and their corresponding structures also vary. Considering all those sentences and finding their corresponding structures are a complicated task. In this paper, we have tried to model the structures representing some Assamese sentences, so that knowledge about the sentences could be achieved in an organized manner. As well as we have tried to perceive the internal meaning of those sentences could be gained which in turn is understood by the computer.

5. PROPOSED MODEL

This paper proposes a model which accesses the Assamese language sentences and tries to extract the structured representation of those sentences. Our model works on an annotated text corpus. The model is shown below.

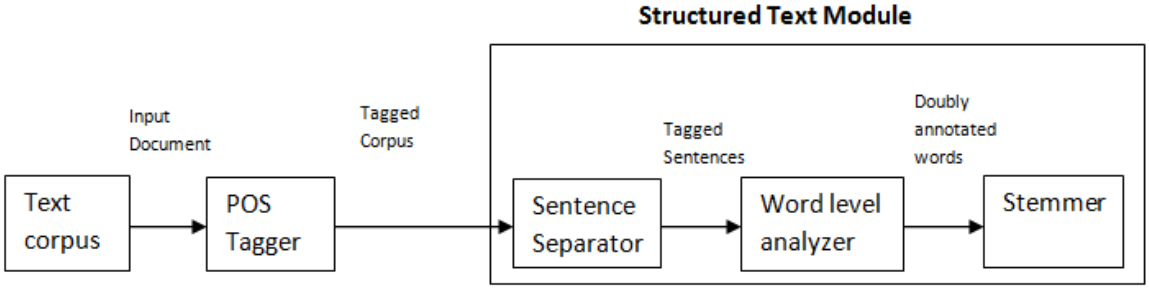

Figure: Structured Text Model

The text corpus acts as an input to the system. It is passed through a POS tagger which makes each sentence of the document properly annotated. The annotated document is now fed as an input to the sentence separator module. This module separates each sentence of the tagged corpus and makes the analysis of these sentences convenient in the word level. Now, the tagged sentences are inputted into the next level of analysis called the word level analyzer. This part takes out the actual words influencing the meaning of a sentence. Also, it deals with finding the

stop words like আৰু, বা, নাইবা, কিন্তু, অথবা, যদি etc and other punctuation markers. The word level analyzer not only deals with finding but removing those stop words also. After removing these words, we get those words whose meanings as a whole can help in finding the interpretation of the whole sentence. In this level, we will try to manually add one more parameter to each these words. Since the sentences are already tagged with a POS tagger, after passing through this module, they will become doubly annotated. The parameters are those which we have discussed in section 4 of this paper. The annotated words generated in this way will form the basis for training a bigger corpus. We have planned to train the system with the annotated words generated as such. Now, the root morphs for these words need to be extracted since they take part in the structure. For this, the actual words must be fed into the stemmer, which stems and generates the root morphs. Now they will form the structure for the sentence. The Sentence Separator, Word level analyzer and stemmer as a whole form the structured text module.

6. DISCUSSIONS AND RESULTS

As mentioned earlier, written language documents or text corpus do not have any specific structure. Computational accesses to such documents allow them to be processed and generated in a structure so that they can be understood in terms of syntactic semantic means. Structured data enables information retrieval easier. It forms the heart of many Artificial Intelligence research areas like question-answering, knowledge representation etc.

The key challenge to our project is the extraction of most relevant information. We already have a properly annotated text corpus which we need to process to make it structured. We have to give each word our own annotation which will form the key input to the actual question-answering task.

Let us make it clearer with the help of an example sentence as mentioned in section 4. The following discussion shows how the sentence flows through the given modules of the proposed system and finally provides a doubly annotated structured format of the sentence. Therefore, considering the following sentence-

মানুহজনে যোৱাকালি এখন নতুন গাড়ী কিনিলে ।

While passing through the POS tagger module, each of the sentence word is tagged by the module and it becomes-

মানুহজনে <NC>

যোৱাকালি <JJ>

এখন <JJ>

নতুন <JJ>

গাড়ী <NC>

কিনিলে<VM>

। <PUN>

Since, this is the only sentence we have assumed, it is passed through the Word level analysis module. This module extracts the stop words first. There is no such occurrence here except the punctuation marker ।, which is removed from this sentence. Therefore, the next level of annotation can now begin. It is important to be mentioned that the annotation must be done manually. After manual annotation, the words will become-

মানুহজনে <NC><Agent>

যোৱাকালি <JJ><Time>

এখন <JJ><Number>

নতুন <JJ><Modifier>

গাড়ী <NC><Object>

কিনিলে<VM><Instance>

Such kind of doubly annotated words will be used as a training data for construction of structured text representation for other tagged sentences. This annotated representation is passed through the stemmer so as to get the root morphs actually occurring in the sentence. Now, they will be as follows-

মানুহ <NC><Agent>

যোৱাকালি <JJ>< Time>

এখন <JJ><Number>

নতুন <JJ><Modifier>

গাড়ী <NC><Object>

কিন্<VM><Instance>

Such a format can construct the structured representation of the sentence. Different structures can be created for different types of sentences. Since, our example sentence modifies the information of গাড়ী (car) as এখন নতুন (a new), therefore, the modified value must be considered as a separate structure [4]. This structure is nested within the original structure (as shown in section 4) representing the whole sentence. In this way, we can create separate structures depending on the type of sentence.

7. CONCLUSION

Natural language processing is an area where numerous research works are going on now a day. It is a significant area of research in Artificial Intelligence. Assamese is a new language in this field where lots of research works are going on. Developments of tools and techniques have started as a mark of digital revolution for this language also. Our work aims at finding the structured text format of Assamese sentences so that an internal representation can be gained. These facilitate computers to process and manipulate them. Our project is the first ever intended work for finding the structured representation of Assamese text. We have tried to work at the semantic level of text analysis. As this language is becoming richer for digital revolution, we visualize our work will bring possibilities for new kinds of research areas such as machine learning, information retrieval or question-answering etc.

REFERENCES

[1] 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

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

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

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

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

[6] Goswami Golak C. (2003). "Asomia Byakoron Prabesh", Bina Library, Guwahati.

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