Clause Identification and Classification in Bengali

Aniruddha Ghosh[1] **Amitava Das**[2] **Sivaji Bandyopadhyay**[3]
Department of Computer Science and Engineering
Jadavpur University
arghyaonline@gmail.com[1] amitava.santu@gmail.com[2] sivaji_cse_ju@yahoo.com[3]

Abstract

This paper reports about the development of clause identification and classification techniques for Bengali language. A syntactic rule based model has been used to identify the clause boundary. For clause type identification a Conditional random Field (CRF) based statistical model has been used. The clause identification system and clause classification system demonstrated 73% and 78% precision values respectively.

1 Introduction

The clause identification is one of the shallow semantic parsing tasks, which is important in various NLP applications such as Machine Translation, parallel corpora alignment, Information Extraction and speech applications. Grammatically a clause is a group of words having a subject and a predicate of its own, but forming part of a sentence. Clause boundary identification of natural language sentences poses considerable difficulties due to the ambiguous nature of natural languages. Clause classification is a convoluted task as natural language is generally syntactically rich in formation of sentences or clauses.

By the classical theory of Panini (Paul and Staal, 1969) a clause is the surface level basic syntactic element which holds the basic dependent semantics (i.e. lexical semantic have no dependency) to represent the overall meaning of any sentence. This syntactic to semantic derivation proceeds through two intermediate stages: the level of *karaka* relations, which are comparable to the thematic role types and the level of inflectional or derivational morphosyntax.

Fillmore's Case Grammar (Fillmore et. al, 2003), and much subsequent work, revived the Panini's proposals in a modern setting. A main objective of Case Grammar was to identify syntactic positions of semantic arguments that may have different realizations in syntax.

In the year of 1996 Bharati et al. (1996) defines the idea of Chunk or local word group for Indian languages. After the successful implementation of Shakti[1] , the first publicly available English-Hindi machine translation system the idea of chunk became the most acceptable syntactic/semantic representation format for Indian languages, known as Shakti Standard Format (SSF).

In 2009 Bali et al. (2009) redefines the idea of chunk and establishes that the idea of chunking varies with prosodic structure of a language. Boundary of chunk level is very ambiguous itself and can differ by writer or speaker according to their thrust on semantic.

Therefore it is evident that automatic clause identification for Indian languages needs more research efforts. In the present task, clause boundary identification is attempted using the classical theory of Panini and the Case Grammar approach of Fillmore on the shallow parsed output in SSF structure. It may be worth mentioning that several basic linguistic tools in Indian languages such as part of speech tagger, chunker, and shallow parser follow SSF[2] as a standard.

Previous research on clause identification was done mostly on the English language (Sang and Dejean, 2001). There have been limited efforts on clause identification for Indian languages. One such effort is proposed in Ram and Devi,

[1] http://shakti.iiit.ac.in/
[2] http://ltrc.iiit.ac.in/MachineTrans/research/tb/shakti-analy-ssf.pdf

(2008) with statistical method. The idea of generative grammar based on rule-based descriptions of syntactic structures introduced by Chomsky (Chomsky, 1956) points out that every language has its own peculiarities that cannot be described by standard grammar. Therefore a new concept of generative grammar has been proposed by Chomsky. Generative grammar can be identified by statistical methods. In the present task, conditional random field (CRF) [3] -based machine learning method has been used in clause type classification. According to the best of our knowledge this is the first effort to identify and classify clauses in Bengali.

The present system is divided into two parts. First, the clause identification task aims to identify the start and the end boundaries of the clauses in a sentence. Second, Clause classification system identifies the clause types.

Analysis of corpus and standard grammar of Bengali revealed that clause boundary identification depends mostly on syntactic dependency. For this reason, the present clause boundary identification system is rule based in nature. Classification of clause is a semantic task and depends on semantic properties of Bengali language. Hence we follow the theory of Chomsky's generative grammar to disambiguate among possible clause types. The present classification system of clause is a statistics-based approach. A conditional random field (CRF) based machine learning method has been used in the clause classification task. The output of the rule based identification system is forwarded to the machine learning model as input.

The rest of the paper is organized as follows. In section 2 we elaborate the rule based clause boundary identification. The next section 3 describes the implementation detail with all identified features for the clause classification problem. Result section 4 reports about the accuracy of the hybrid system. In error analysis section we reported the limitations of the present system. The conclusion is drawn in section 5 along with the future task direction.

2 Resource Acquisition

Bengali belongs to Indo-Aryan language family. A characteristic of Bengali is that it is under-resourced. Language research for Bengali got attention recently. Resources like annotated corpus and linguistics tools for Bengali are very rarely available in the public domain.

2.1 Corpus

We used the NLP TOOLS CONTEST: ICON 2009[4] dependency relation marked training dataset of 980 sentences for training of the present system. The data has been further annotated at the clause level. According to the standard grammar there are two basic clause types such as Principal clause and Subordinate clause. Subordinate clauses have three variations as Noun clause, Adjective clause and Adverbial clause. The tagset defined for the present task consists of four tags as Principal clause (PC), Noun clause (NC), Adjective clause (AC) and Adverbial clause (RC). The annotation tool used for the present task is Sanchay[5]. The detailed statistics of the corpus are reported in Table 1.

	Train	Dev	Test
No of Sentences	980	150	100

Table 1: Statistics of Bengali Corpus

2.1.1 Annotation Agreement

Two annotators (Mr. X and Mr. Y) participated in the present task. Annotators were asked to identify the clause boundaries as well as the type of the identified clause. The agreement of annotations among two annotators has been evaluated. The agreements of tag values at clause boundary level and clause type levels are listed in Table 2.

	Boundary	Type
Percentage	76.54%	89.65%

Table 2: Agreement of annotators at clause boundary and type level

It is observed from the Table 2 that clause boundary identification task has lower agreement value. A further analysis reveals that there are almost 9% of cases where clause boundary has nested syntactic structure. These types of clause boundaries are difficult to identify. One of such cases is Inquisitive semantic (Groenendijk, 2009) cases, ambiguous for human annota-

[3] http://crf.sourceforge.net/

[4] http://ltrc.iiit.ac.in/nlptools2009/

[5] http://ltrc.iiit.ac.in/nlpai_contest07/Sanchay/

tors too. It is better to illustrate with some specific example.

```
If John goes to the party,
will Mary go as well?
```

In an inquisitive semantics for a language of propositional logic the interpretation of disjunction is the source of inquisitiveness. Indicative conditionals and conditional questions are treated both syntactically and semantically. The semantics comes with a new logical-pragmatically notion that judges and compares the compliance of responses to an initiative in inquisitive dialogue (Groenendijk, 2009). Hence it is evident that these types of special cases need special research attention.

2.2 Shallow Parser

Shallow parser[6] for Indian languages, developed under a Government of India funded consortium project named Indian Language to Indian Language Machine Translation System (IL-ILMT), are now publicly available. It is a well developed linguistic tool and produce good credible analysis. For the present task the linguistic analysis is done by the tool and it gives output as pruned morphological analysis at each word level, part of speech at each word level, chunk boundary with type-casted chunk label, vibhakti computation and chunk head identification.

2.3 Dependency parser

A dependency parser for Bengali has been used as described in Ghosh et al. (2009). The dependency parser follows the tagset[7] identified for Indian languages as a part of NLP TOOLS CONTEST 2009 as a part of ICON 2009.

3 Rule-based Clause Boundary Identification

Analysis of a Bengali corpus and standard grammar reveals that clause boundaries are directly related to syntactic relations at sentence level. The present system first identifies the number of verbs present in a sentence and subsequently finds out dependant chunks to each verb. The set of identified chunks that have relation with a particular verb is considered as a clause. But some clauses have nested syntactic formation, known as inquisitive semantic. These clauses are difficult to identify by using only syntactic relations. The present system has limitations on those inquisitive types of clauses.

Bengali is a verb final language. Most of the Bengali sentences follow a Subject-Object-Verb (SOV) pattern. In Bengali, subject can be missing in a clause formation. Missing subjects and missing keywords lead to ambiguities in clause boundary identification. In sentences which do not follow the SOV pattern, chunks that appear after the finite verb are not considered with that clause. For example:

```
wAra AyZawana o parimANa
xeKe buJawe asubiXA hayZa ei
paWa hAwi geCe.

After seeing the size and
effect, it is hard to under-
stand that an elephant went
through this way.
```

In the above example, there is hardly any clue to find beginning of subordinate clause. To solve this type of problem, capturing only the tree structure of a particular sentence has been treated as the key factor to the goal of disambiguation. One way to capture the regularity of chunks over different sentences is to learn a generative grammar that explains the structure of the chunks one finds. These types of language properties make the clause identification problem difficult.

3.1 Karaka relation

Dependency parsing generates the inter chunk relation and generates the tree structure. The dependency parser as described in Section 2.3 used as a supportive tool for the present problem.

In the output of the dependency parsing systems, most of the chunks have a dependency relation with the verb chunk. These relations are called as *karaka* relation. Using dependency relations, the chunks having dependency relation i.e. *karaka* relation with same verb chunk are grouped. The set of chunks are the members of a clause. Using this technique, identification of chunk members of a certain clause becomes independent of SOV patterns of sentences. An example is shown in Figure 1.

[6] http://ltrc.iiit.ac.in/analyzer/bengali/

[7] http://ltrc.iiit.ac.in/nlptools2009/CR/intro-husain.pdf

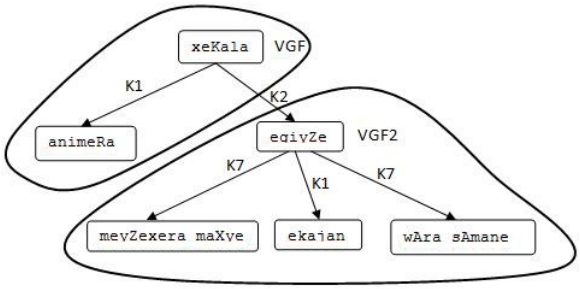

Figure 1: Karaka Relations

3.2 Compound verbs

In Bengali language a noun chunk with an infinite verb chunk or a finite verb chunk can form a compound verb. An example is shown in Figure 2.

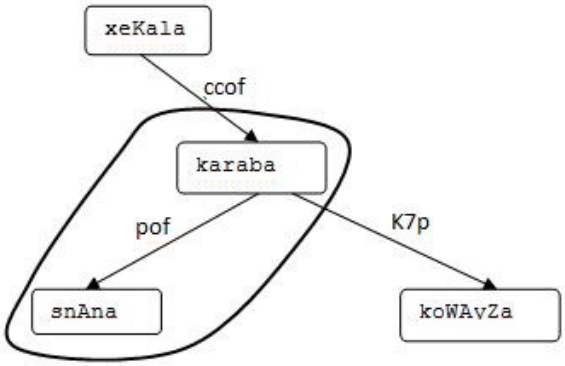

Figure 2: Compound Verb

In the above example, the noun chunk and the VGF chunk form a compound verb. These two consecutive noun and verb chunks appearing in a sentence are merged to form a compound verb. These chunks are connected with a part-of relation in Dependency Parsing. The set of related chunks with these noun and verb chunks are merged.

3.3 Shasthi Relation (r6)

In dependency parsing the genitive relation are marked with *shasthi* (r6) relation. The chunk with *shasthi* (r6) (see the tagset of NLP Tool Contest: ICON 2009) relation always has a relation with the succeeding chunk. An example is shown in Figure 3.

In the example as mentioned in Figure 3, the word "*wadera*"(their) has a genitive relation with the word in the next chunk "*manera*"(of mind). These chunks are placed in a set. It forms a set of two chunks members. The system generates two different types of set. In one forms a set of members having relation with verb chunks. Another set contains two noun chunks with genitive relation. Now the sets containing only noun chunks with genitive relation does not form a clause. Those sets are merged with the set containing verb chunk and having dependency relation with the noun chunks. An example is shown in Figure 3.

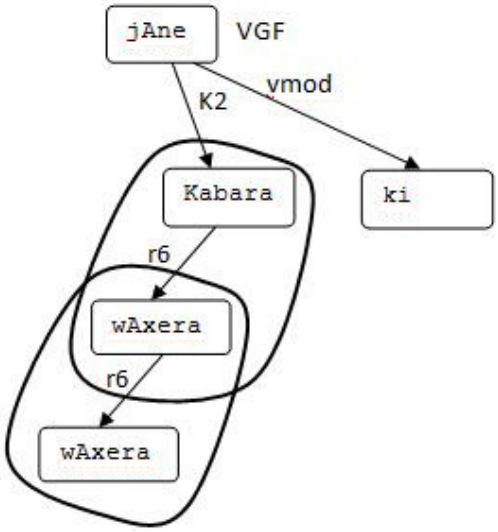

Figure 3: Shasthi Relation

Consider ω is set of all sets containing two chunk members connected with genitive marker. Consider β is a set of all sets consisting of related chunks with a verb chunk. λ is a element of ω. α is a element of β. Now, If a set λ which can have common chunks from a α set then λ set is associated with the proper α set. So, $\lambda \cap \alpha \neq$ Null then $\alpha = \alpha \cup \lambda$. If a set λ which can have common chunks from two α sets which leads to ambiguity of associability of the λ set with the proper α set. If $\lambda \cap \alpha$ = verb chunk, then λ set will be associated with α set containing the verb chunk. From the related set of chunk of verb chunks, system has identified the clauses in the sentence. Afterwards, the clauses are marked with the B-I-E (Beginning-Intermediate-End) notation.

4 Case Grammar-Identification of Karaka relations

The classical Sanskrit grammar Astadhyayi [8] ('Eight Books'), written by the Indian gramma-

[8] http://en.wikipedia.org/wiki/P%C4%81%E1%B9%87ini

rian Panini sometime during 600 or 300 B.C. (Robins, 1979), includes a sophisticated theory of thematic structure that remains influential till today. Panini's Sanskrit grammar is a system of rules for converting semantic representations of sentences into phonetic representations (Kiparsky, 1969). This derivation proceeds through two intermediate stages: the level of *karaka* relations, which are comparable to the thematic role types described above; and the level of morpho-syntax.

Fillmore's Case Grammar (Fillmore, 1968), and much subsequent work, revived the Panini's proposals in a modern setting. A main objective of Case Grammar was to identify semantic argument positions that may have different realizations in syntax. Fillmore hypothesized 'a set of universal, presumably innate, concepts which identify certain types of judgments human beings are capable of making about the events that are going on around them'. He posited the following preliminary list of cases, noting however that 'additional cases will surely be needed'.

- **Agent:** The typically animate perceived instigator of the action.
- **Instrument:** Inanimate force or object causally involved in the action or state.
- **Dative:** The animate being affected by the state or action.
- **Factitive:** The object or being resulting from the action or state.
- **Locative:** The location or time-spatial orientation of the state or action.
- **Objective:** The semantically most neutral case, the concept should be limited to things which are affected by the action or state.

The SSF specification handles this syntactic dependency by a coarse-grain tagset of Nominitive, Accusative, Genitive and Locative case markers. Bengali shallow parser identifies the chunk heads as part of the chunk level analysis. Dependency parsing followed by a rule based module has been developed to analyze the inter-chunk relationships depending upon each verb present in a sentence. Described theoretical aspect can well define the problem definition of clause boundary identification but during practical implementation of the solution we found some difficulties. Bengali has explicit case markers and thus long distant chunk relations are possible as valid grammatical formation. As an example:

```
bAjAre yAoyZAra samayZa xeKA
kare gela rAma.

bAjAre yAoyZAra samayZa rAma
xeKA kare gela.

rAma bAjAre yAoyZAra samayZa
xeKA kare gela.

Rama came to meet when he
was going to market.
```

In the above example *rAma* could be placed anywhere and still all the three syntactic formation are correct. For these feature of Bengali many dependency relation could be missed out located at far distance from the verb chunk in a sentence. Searching for uncountable numbers of chunks have dependency relation with a particular verb may have good idea theoretically but we prefer a checklist strategy to resolve the problem in practice. At this level we decided to check all semantic probable constituents by the definition of universal, presumably innate, concepts list. We found this is a nice fall back strategy to identify the clause boundary. Separately rules are written as described below.

4.1 Agent

Bengali is a verb final language. Most of the Bengali sentences follow a Subject-Object-Verb (SOV) pattern. In Bengali, subject can be missing in a clause formation. Missing subjects and missing keywords lead to ambiguities in clause boundary identification.

দরজাটা বন্ধ করো।
Close the door.

In the previous case system marks "*দরজাটা/door*" as an "*Agent*" whereas the "*Agent*" is "*you*" (2nd person singular number), silent here.

We developed rules using case marker, Gender-Number-Person (GNP), morphological feature and modality features to disambiguate these

types of phenomena. These rules help to stop false hits by identifying no 2nd person phrase was there in the example type sentences and empower to identify proper phrases by locating proper verb modality matching with the right chunk.

4.2 Instrument

Instrument identification is ambiguous for the same type of case marker (nominative) taken by agent and instrument. There is no animate/inanimate information is available at syntactic level.

শ্যামের বাঁশির সুর মন্ত্রমুগ্ধকর।

The music of Shyam's messmerized me.

সুমির ছাতা।

The umbrella of Sumi.

Bengali sentences follow a Subject-Object-Verb (SOV) pattern. Positional information is helpful to disambiguate between agent and instrument roles.

4.3 Dative

	Bengali		English Gloss
General	সকাল/সন্ধ্যে/রাত/ভোর...		Morning/evening/night/dawn…
	_টার সময়/সময়/ঘটিকায়/মিনিট/সেকেন্...		O clock/time/hour/minute/second…
	সোমবার/মঙ্গলবার/রবিবার...		Monday/Tuesday/Sunday…
	বৈশাখ/জৈষ্ঠ/...		Bengali months…
	জানুয়ারী/ফেব্রুয়ারী		January/February…
	দিন/মাস/বছর...		Day/month/year…
	কাল/ক্ষন/পল...		Long time/moment…
Relative	আগে/পরে...		Before/After…
	সামনে/পেছনে...		Upcoming/
	Special Cases	উঠলে/থামলে...	When rise/When stop…

Table 3: Categories of Time Expressions

Time expression identification has a different aspect in NLP applications. People generally studied time expression to track event or any other kind of IR task. Time expressions could be categorized in two types as General and Relative.

In order to apply rule-based process we developed a manually augmented list with pre defined categories as described in Table 3. Still there are many difficulties to identify special cases of relative time expressions. As an example:

চাঁদ উঠলে আমরা রওনা হবো।

When moon rise we will start our journey.

In the previous example the relative time expression is "*উঠলে/when rise*" is tagged as infinite verb (for Bengali tag level is VGNF). Statistics reveals that these special types of cases approximately are only 1.8-2% in overall corpus.

These types of special cases are not handled by the present system.

4.4 Factitive

The particular role assignment is the most challenging task as it separately known as argument identification. To resolve this problem we need a relatively large corpus to learn fruitful feature similarities among argument structures.

A manually generated list of causative postpositional words and pair wise conjuncts as reported in Table 4 has been prepared to identify argument phrases in sentences.

	Bengali	English Gloss
General	জন্য/কারনে/হেতু...	Hence/Reason/Reason
Relative	যদি_তবে	If_else
	যদিও_তবুও	If_else

Table 4: Categories of Causative Expressions

4.5 Locative

Rules have been written using a manually edited list as described in Table 5. Morphological locative case marker feature have been successfully used in identification of locative marker. There is an ambiguity among Agent, Dative and Locative case marker as they orthographically generates same type of surface form (using common

suffixes as: ে, ের etc). There is less differences we noticed among their syntactic dependency structure throughout the corpus. Positional information helps in many cases to disambiguate these cases.

দেশে কাজ নেই বাবু।

There is unemployment in country side.

A different type of problem we found where verb plays locative role. As an example:

লোকে যেখানে কাজ করে সেখানে।

Where people works there.

Here *"যেখানে কাজ করে/Where people works"* should be identified as locative marker. But this is a verb chunk and leads difficulty. Corpus statistics reveals that this type of syntactic formation is approximately 0.8-1.0% only and not been handled by the present system.

	Bengali	English Gloss
General	মাঠে/ঘাটে/রাস্তায়	Morning/evening/night/dawn…
Relative	আগে/পরে...	Before/After…
	সামনে/পেছনে...	Front/Behind

Table 5: Categories of Locative Expressions

4.6 Objective

The concept of objectivity is limited to things or human which are affected by the action or state. Statistical parser is a best way out for the present problem. The *karma karaka* (k2) identified by the dependency parser is simply the objective constituent of any clause.

5 Identification the Type of Clauses

After marking of the clause boundaries, clause types are identified. According to the clause structure and functions in a sentence, clauses are classified in to four types as principal clause, noun clause, adverbial clause and adjective clause. To identify the clause types, a CRF based statistical approach has been adopted.

5.1 Generative Grammar

In theoretical linguistics, generative grammar refers to a particular approach to the study of syntax. A generative grammar of a language attempts to give a set of rules that will correctly predict which combinations of words will form grammatical sentences. Chomsky has argued that many of the properties of a generative grammar arise from an "innate" universal grammar. Proponents of generative grammar have argued that most grammar is not the result of communicative function and is not simply learned from the environment. Strongly motivated by Chomsky's generative grammar we adopt the CRF based machine learning to learn the properties of a language and apply the knowledge to typecast clause classification as well.

5.2 Conditional Random Fields (CRF)

CRFs are undirected graphical models which define a conditional distribution over a label sequence given an observation sequence. CRF usually trained based on input features. Maximum likelihood is being calculated on chosen features for training.

5.2.1 Features

The vitality of using any machine learning approach is in identification of proper feature set. Conditional Random Field (CRF) works on a conditional distribution over a label sequence given an observation sequence. Hence CRF used here to statistically capture the prosodic structure of the language. The features experimentally found useful are chosen as listed below.

5.2.2 Chunk Label

An *n*-gram chunk label window has been fixed to capture internal arrangement of any particular clause type.

5.2.3 Chunk Heads

Chunk head pattern is the vital clue to identify the any clause pattern.

5.2.4 Word

In the clause type identification task words play a crucial part as word carries the information of the clause type.

From the input file in the SSF format, all the morphological information like root word, chunk heads are retrieved. The clause type identification depends on the morphological information along with the position in the sentences and also the surrounding chunk labels. Therefore the CRF based statistical tool calculates the probability of

the morphological information along with the dependency relations of the previous three and next three chunks. For the present task a quad-gram technique is used as most of the sentences have around 10 chunks.

The input file in the SSF format includes Chunk labels and word. The clause information in the input files are in B-I-E format so that the begin (B) / inside (I) / end (E) information for a clause are associated as a feature. The chunk heads, words are identified from the training file and noted as an input feature in the CRF based system. Each sentence is represented as a feature vector for the CRF machine learning task. The input features associated with each word in the training set are the word, clause boundary tags, chunk tag and clause type tags.

6 Error Analysis

During the development stage of the system we had studied the various clause boundary labeling errors committed by the system. In the above examples, the system faces ambiguity to derive the rules for the identification of the clause members when a noun chunk acts as a noun modifier of a clause. In complex sentences, the verb chunk of the subordinate clause may have noun modifier relation with the principal clause. As System forms the groups the chunks with dependency relation, system merges the subordinate clause with principal clause. An example is shown in Figure 4.

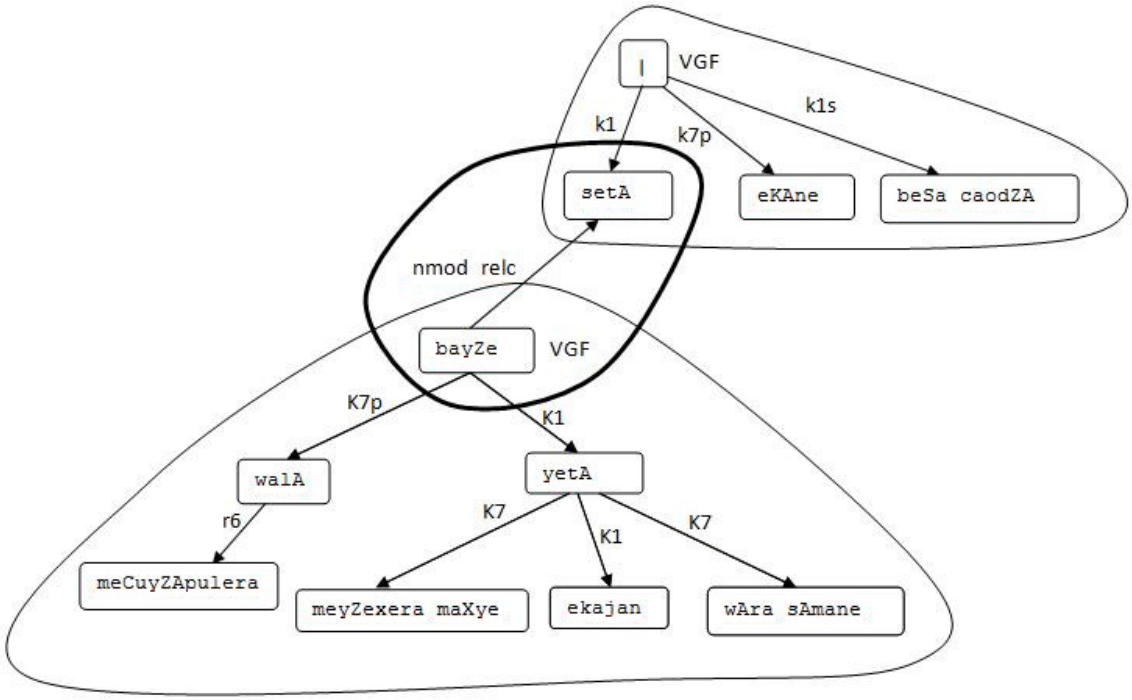

Figure 4: Shasthi Relation

7 Experimental results

System	Precision	Recall
Boundary	73.12%	75.34%
Classification	78.07%	78.92%

Table 6: Performance of present System

The accuracy of the rule-based clause boundary identification system is 73.12% and 78.07% is the accuracy clause type classification system as reported in Table 6.

8 Conclusion

This paper reports about our works on clause identification and classification in Bengali language. We have used the rule based system to identify clause boundary and a statistical CRF based model is used to decide the type of a clause.

In future we would like to study different semantic relations which can regulate clause type and boundary.

References

A. Ghosh, A. Das, P. Bhaskar, S. Bandyopadhyay. Dependency Parser for Bengali: the JU System at ICON 2009, In NLP Tool Contest ICON 2009, December 14th-17th, 2009, Hyderabad.

Akshar Bharati, Vineet Chaitanya , Rajeev Sangal. Natural Language Processing A Paninian Perspective. Prentice Hall of India (1995).

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. International Journal of Lexicography, 16:235–250.

Chomsky, Noam (1956). "Three models for the description of language". IRE Transactions on Information Theory 2: 113–124.

Erik F. Tjong kim sang and Herve Dejean Introduction to CoNLL-2001 shared task: clause identification.

Groenendijk, J.: (2009), 'Inquisitive Semantics: Two Possibilities for Disjunction'. In Lecture Notes in Computer Science. ISBN- 978-3-642-00664-7. Volume- 5422/2009. Berlin, Heidelberg. Pages-80-94.

Kalika Bali, Monojit Choudhury, Diptesh Chatterjee, Arpit Maheswari, Sankalan Prasad. Correlates between Performance, Prosodic and Phrase Structures in Bangla and Hindi: Insights from a Psycholinguistic Experiment. In Proceeding of ICON 2009. Hyderabad. India.

Kiparsky, Paul and J. F. Staal (1969). 'Syntactic and semantic relations in Panini.' Foundations of Language 5, 83-117.

Robins, R. H. (1979). A Short History of Linguistics (2nd Edition). London: Longman.

Vijay Sundar Ram. R and Sobha Lalitha Devi, 2008 Clause Boundary Identification Using Conditional Random Fields.