ISSN:0975-9646

Syntactic Sentence Fusion Techniques for Bengali

Amitava Das[1] and Sivaji Bandyopadhyay[2]

Department of Computer Science and Engineering
Jadavpur University

***Abstract*— The present paper describes various syntactic sentence fusion techniques for Bengali language that belongs to the Indo-Aryan language family. Firstly a clause identification and classification system marks clause boundaries and classifies them as principle clause and subordinate clauses. A rule-based sentence classification system has been developed to categorize sentences as simple, complex and compound. The final syntactic sentence fusion system makes use of the sentence class and the clause types and finally fuses two textually entailed sentences using verb paradigm information and noun morphological information. The system outputs are compared with a gold standard data set using manual evaluation and BLEU techniques. The evaluation results yield good accuracy scores. The syntactic sentence fusion technique developed in the present work may be applied for other Indian languages.**

***Keywords*—Clause Identification and Classification, Sentence Type, Syntactic Sentence Fusion, Evaluation.**

I. INTRODUCTION

Traditionally, Natural Language Generation (NLG) is defined as the automatic production of "meaningful texts in human language from some underlying non-linguistic representation of information" [1]. Recently, there is an increased interest in NLG applications that produce meaningful text from other meaningful texts rather than from abstract meaning representations. Such applications are sometimes referred to as text-to-text generation applications [2], [3], [4], and may be likened to earlier revision-based generation strategies [5], [6]. Text-to-text generation is often motivated from practical applications such as summarization [7], sentence simplification [8] and sentence compression [9]. One reason for the interest in such generation systems is the possibility of automatically learning text-to-text generation strategies from corpora of parallel text or semantically entailed texts.

Our endeavor was to develop a syntactic sentence fusion system for Bengali language that belongs to the Indo-Aryan language family. Bengali is the sixth [1] highest speaking language round the globe, second in India and is the national language in Bangladesh. Indian languages and especially Bengali is morpho-syntactically rich and highly inflective in nature. The properties of the language make the language generation problem itself very hard.

According to the best of our knowledge no significant effort could be found for sentence fusion techniques in Indian languages. Therefore it will be unfair to compare the present task with the existing technologies for other Indian languages. Bengali is a resource constrained language for natural language processing activities. Therefore many linguistic analysis tools have to be built in order to develop the final fusion system. A clause analysis module that includes clause boundary identification and classification and a sentence categorization module that classifies sentences into simple, complex or compound categories have been built. The rest of the paper is organized as follows, Section II describes resource acquisition that includes corpus collection and annotation and various necessary linguistic tools. Section III detailed clause boundary identification module and Section IV describes clause type classification system. Section V describes the sentence type identification module, which classify sentences into simple, complex or compound categories. Morphological generation for Bengali noun and verb has been described in Section VI and Section VII details the syntactic sentence fusion techniques separately for each sentence category. Section VIII describes the various evaluation strategies. The conclusion is drawn in Section IX.

II. RESOURCE ACQUISITION

Resource acquisition is one of the most challenging tasks while working with resource constrained languages like Bengali. Some of the linguistic tools have been collected from publicly available resources and some other tools have been developed like the clause analysis module that includes clause boundary identification and classification. In this section we will describe particularly corpus acquisition and annotation. Additionally we will give necessary information about the publicly available tools that have been collected and used successfully.

A. Corpus

1) Corpus for Clause Analysis

The NLP TOOLS CONTEST: ICON 2009 [2] dependency relation marked training dataset has been used. The corpus has been further annotated at the clause level marking the Principal clause and the Subordinate clauses. According to standard Bengali grammar [10] subordinate clauses have three variations as Noun clause, Adjective clause and Adverbial

[1] http://en.wikipedia.org/wiki/List_of_languages_by_number_of_native_speakers

[2] http://ltrc.iiit.ac.in/nlptools2009/CR/all-papers-toolscontest.pdf

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[3]. A brief statistics of the corpus is reported in Table 1.

TABLE 1: STATISTICS OF BENGALI CORPUS

	TRAIN	DEV	TEST
No of Sentences	980	150	100

2) Annotation and Challenges

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.

TABLE 2: AGREEMENT OF ANNOTATORS AT CLAUSE BOUNDARY AND TYPE LEVEL

	BOUNDARY	TYPE
PERCENTAGE	76.54%	89.65%

It is observed from the Table 2 that clause boundary identification task has relatively 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 [11], present in the following example sentence:

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-pragmatic notion that judges and compares the compliance of responses to an initiative in inquisitive dialogue [11]. Hence it is evident that these types of special cases need special research attention.

3) Corpus for Sentence Fusion

For the sentence fusion task a NEWS corpus has been manually collected. Two popular Bengali news papers[4,5] have been chosen as the information resource. Total 50 parallel stories have been picked up randomly. As the present system works by analyzing sentence types thus sentences are chosen based on their type (simple, complex or compound). In total 300 sentences have been chosen that share redundant information on a common topic in the two Bengali newspapers.

Three annotators have been asked to create manually fused sentences. As the present system is rule based in nature the gold standard data has been used for the evaluation purpose only and not for training. A brief statistics of the sentence fusion corpus are reported in the Table 3. The number reported in the table simply depicts how many sentence pairs of the particular type have been chosen. Specifically 48 simple-simple, 53 simple-complex, 46 simple-compound, 58 complex-complex, 50 complex-compound and 45 compound-compound sentence pairs have been chosen from both the newspapers.

TABLE 3: SENTENCE FUSION CORPUS

	Simple	Complex	Compound
Simple	48	53	46
Complex	53	58	50
Compound	46	46	45

B. Linguistics Tools

Sentence fusion is a natural language generation problem, therefore the basic language analysis tools are required.

1) Shallow Parser

Publicly available shallow parsers for Indian languages (specially for Bengali) has been used 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) Dependency Parser

A dependency parser for Bengali has been used as described in Ghosh et al., 2009 [12]. The dependency parser follows the tagset identified for Indian languages as part of NLP TOOLS CONTEST 2009 held in ICON 2009.

III. CLAUSE BOUNDARY IDENTIFICATION

The clause analysis 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 [10] of Bengali reveals 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 clauses is a semantic task and depends on the semantic properties of Bengali language. The present clause classification system follows a statistical approach. A conditional random field (CRF[6]) 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.

A. Rule-based Clause Boundary Identification

[3] http://ltrc.iiit.ac.in/nlpai_contest07/Sanchay/
[4] http://www.anandabazar.com/
[5] http://www.sangbadpratidin.net/
[6] http://crf.sourceforge.net/

Analysis of a Bengali corpus and standard grammar [10] 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 dependent 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 these 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 understand that an elephant went through this way].

In the above example, there is hardly any clue to find the beginning of the subordinate clause. To solve this type of problem, capturing only the tree structure of a particular sentence has been treated as the key factor of disambiguation. These types of language properties make the clause identification problem difficult.

Every language has some peculiarity or in other words some unique distinguishing feature. Therefore some rules developed to deal with these peculiarities are described in the following sub sections.

1) Karaka relation

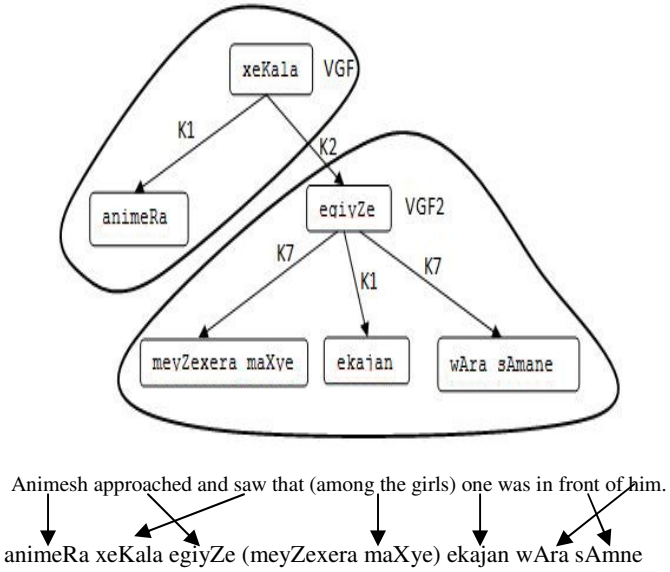


Figure 1: Karaka Relations

Dependency parsing assigns the inter chunk relationships and generates the tree structure. The dependency parser is used as a supportive tool for the present problem.

In the output of the dependency parsing systems, every chunk has a dependency relation with the verb chunk. These relations are identified as *karaka* relations for Indian languages. Using dependency relations, the chunks having dependency relation, i.e., *karaka* relations 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 the Figure 1.

2) Compound Verbs

In every Indian languages and especially in Bengali language a noun chunk with an infinite or a finite verb chunk can form a compound verb. An example is shown in the Figure 2.

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 during Dependency Parsing. The set of related chunks with these noun and verb chunks are merged.

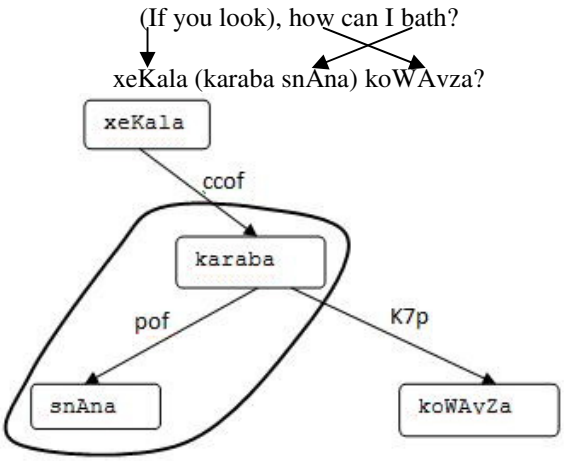


Figure 2: Compound Verb

3) Shasthi Relation (r6)

Two nouns connected with genitive relations are marked as *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 as members. The system generates two different types of sets. The first one is a set of members having relation with verb chunks. The other set contains two noun chunks with genitive relation. Now the set containing only noun chunks with genitive relation does not form a clause. These sets are merged with the set containing

verb chunk having dependency relation with the noun chunks. An example is shown in the Figure 3.

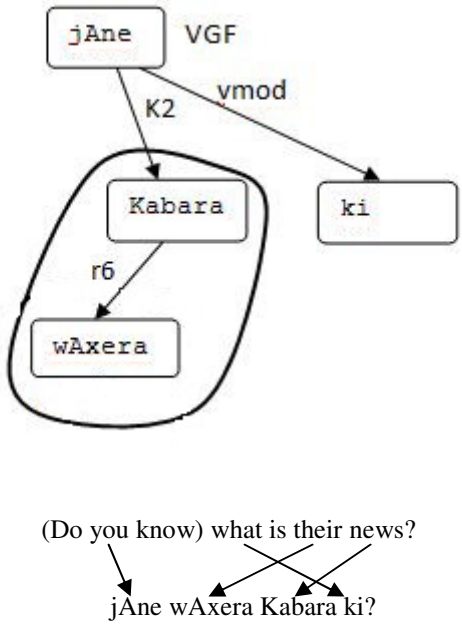


Figure 3: Shasthi Relation

B. Case Grammar/Karaka Relations for Clause Boundary Identification

The classical Sanskrit grammar Astadhyayi [7] ('Eight Books'), written by the Indian grammarian Panini sometime during 600 or 300 B.C. [13], 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 [14]. 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 morphosyntax.

Fillmore's Case Grammar [15], 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 Shakti Standard Format (SSF) specification handles this syntactic dependency by a coarse-grain tagset of Nominative, 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 analyse the inter-chunk relationships depending upon each verb present in a sentence. Some difficulties were faced during the implementation of the clause boundary identification. Bengali has explicit case markers and thus long distant chunk relations are possible as valid grammatical forms. 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 formations are correct. For this feature of Bengali many dependency relations could be missed out that are located at a distance from the verb chunk in a sentence. Searching for all chunks in a sentence that have dependency relations with a particular verb takes time and space. A checklist is preferred to resolve the issue in practice. At this level we check all semantic probable constituents following the definition of universal, presumably innate, concepts list. We found this is a nice fall back strategy to identify the clause boundary. Separate rules are written as described below.

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) that is silent in the sentence.

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 that no 2nd person phrase was there in the example type sentences and identifies appropriate

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

phrases by locating proper verb modality matching with the right chunk.

2) Instrument

Instrument identification is ambiguous for the same type of case marker (nominative) taken by agent and instrument. 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.

3) Dative

Time expression identification has a different aspect in NLP applications. Time expressions are studied to track events or various kinds of Information Retrieval (IR) tasks. Time expressions could be categorized into two types: General and Relative.

TABLE 4: CATEGORIES OF TIME EXPRESSIONS

	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…

In order to apply rule-based process we developed a manually augmented list with pre-defined categories as described in Table 4. Still there are many difficulties to identify special cases of relative time expressions. Consider the example sentence:

চাঁদ উঠলে আমরা রওনা হবো।

[We will start our journey when the moon rises.]

In the previous example the relative time expression is "*উঠলে/when rise*" is tagged as infinite verb (for Bengali tag level is VGNF). Statistics reveal that these special types of cases are approximately 1.8-2% in the overall corpus. These types of special cases are not handled by the present system.

4) Factitive

This particular role assignment is the most challenging task, also 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 5 has been prepared to identify argument phrases in sentences.

TABLE 5: CATEGORIES OF CAUSATIVE EXPRESSIONS

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

5) Locative

Rules have been written using a manually edited list as described in Table 6. Morphological locative case marker feature have been successfully used in identification of locative marker. There is ambiguity among Agent, Dative and Locative case markers as they orthographically generate same type of surface form (using common suffixes as: ে, েের etc). It has been observed that there is minimal differences among their syntactic dependency structures 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 is observed where verb plays the locative role. As an example:

লোকে যেখানে কাজ করে সেখানে।

[Where people works there.]

TABLE 6: CATEGORIES OF LOCATIVE EXPRESSIONS

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

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

6) Objective

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

IV. CLAUSE TYPE IDENTIFICATION

After marking of the clause boundaries, clause types are identified. According to the clause structure and functions in a sentence, clauses are classified into 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.

A. 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 correctly predicts the combinations of words to form grammatical sentences. Chomsky [16] 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 grammars are not the result of communicative functions and are not simply learned from the environment. Strongly motivated by Chomsky's generative grammar we adopt the CRF based machine learning technique to learn the properties of a language and apply the knowledge to typecast clause classification as well.

B. Conditional Random Fields (CRF)

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

C. 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 is used here to statistically capture the prosodic structure of the language. The features that are experimentally found as useful are chosen and are listed below.

1) Chunk Label

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

2) Chunk Heads

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

3) Word

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

D. Performance of Clause Identification and Classification

TABLE 7: PERFORMANCE OF PRESENT SYSTEM

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

The accuracy of the rule-based clause boundary identification system is 73.12% while the accuracy of the clause type classification system is 78.07%, as reported in Table 7.

E. Error Analysis

During the development stage of the system we have studied various clause boundary labelling errors committed by the system. The system faces ambiguity to identify 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 the chunks are grouped with dependency relations, system merges the subordinate clause with the principal clause. An example is shown in the Figure 4.

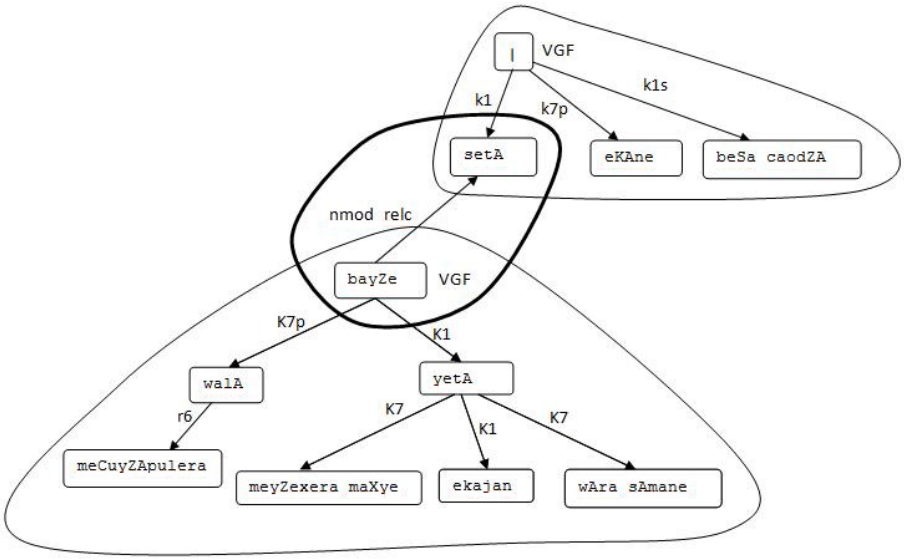


Figure 4: Shasthi Relation

V. SENTENCE TYPE IDENTIFICATION

The system is fully rule based and depends on the type of clauses present in a sentence. If there is only one principal clause present then it is treated as a simple sentence. Both complex and compound sentences consist of at least one principal clause and one dependant clause. A few simple rules easily help to distinguish among these two sentence types. If any conjunct is left out after clause boundary identification then the sentence is treated as a compound sentence otherwise the sentence is treated as a complex sentence.

As this is a rule based system therefore no gold standard data has been prepared. Evaluation has been carried out on a randomly selected set of 100 sentences. The system yields good accuracy for simple sentences but faced minor difficulties for identification for complex and compound sentences as reported in Table 7. The present system assigns type information for all the sentences present in a corpus and hence the precision and recall value for the system is same. Only accuracy score is reported in the Table 8.

TABLE 8: PERFORMANCE OF PRESENT SYSTEM

Type	Performance
Simple	88.12%
Complex	78.07%
Compound	82.50%

Looking at the accuracy figures it could be easily inferred that system needs more rules for disambiguation of complex and compound sentences. A confusion matrix helps to understand the detail of the problem. The confusion matrix is reported in the Table 9.

TABLE 9: CONFUSION MATRIX

	Simple	Complex	Compound
Simple	88.12%	-	-
Complex	-	78.07%	8.02%
Compound	-	3.5%	82.50%

VI. MORPHOLOGICAL GENERATION

Two types of morphological generation techniques have been adopted for the present task: noun generation and verb generation.

A. Noun Phrase Generation

Noun words have different semantic roles in a sentence. Based on the semantic roles of a noun in any sentence the target language noun has been generated. Various fine-grained semantic roles could be identified from various existing resources like Propbank, FrameNet or VerbNet. But unfortunately no such resources exist for Indian languages. Therefore the present task is only based on the basic semantic role types. Based on the Fillmore's Case grammar and semantic roles different suffix lists have been prepared. Depending upon the semantic role of a noun in the targeted sentence these suffixes are attached with the root form, obtained from the output of the morphological analyser. Some examples of such suffixes are reported in the Table 10.

TABLE 10: CASE-WISE SUFFIX LIST

Case	Suffix	Example
Agent	ে,ের,কে,বাবু	রামের, শ্যামকে
Instrument	_দিয়ে,দ্বারা	চামচ_দিয়ে
Dative	_সময়ে,_দিনে,	শেষ_দিনে
Factitive	ে,ের	ভাতের
Locative	তে,য়,	কলকাতাতে,রাস্তায়

B. Verb Phrase Generation

English verbs have to agree with the subject in person and number information. But in contrast, Bengali verbs have to agree with the subject in person and formality. Since second person singular personal pronoun 'you' have three forms in Bengali /"আপনি" (apni), "তুমি" (tumi), and "তুই" (tui)/, depending on formality information, "you went" has the singular form /"আপনি গেলেন" (apni gelen), "তুমি গেলে" (tumi gele), "তুই গেলি" (tui geli)/ and, likewise, plural form /"আপনারা গেলেন" (apnara gelen), "তোমরা গেলে" (tomra gele), "তোরা গেলি" (tora geli)/, as shown in the Table 11. Similarly, 'he came' has the meaning /"তিনি আসলেন" (tini aslen), "সে আসল" (se aslo)/.

TABLE 11: SINGULAR-PLURAL BASED BENGALI VERB FORMS

"you went"	Singular	Plural
Formal	"আপনি গেলেন" (apni gelen)	"আপনারা গেলেন" (apnara gelen)
Neutral	"তুমি গেলে" (tumi gele)	"তোমরা গেলে" (tomra gele)
Intimate	"তুই গেলি" (tui geli)	"তোরা গেলি" (tora geli)

TABLE 12: BENGALI COMPOUND VERBS

English verbs	Bengali verbs
Swim	"সাঁতার" (santar [n. swimming]) "কাটা" (kata [v. cut])
Try	"চেষ্টা" (chesta [n. try]) "করা" (kara[v. do])
rest	"বিশ্রাম" (bishram [n. rest]) "নেওয়া" (neowa [v. take])
fail	"ব্যার্থ" (byartho [adj. futile]) "হওয়া" (howa [v. be])
rain	"বৃষ্টি" (brishti [n. rain]) "পড়া" (pora [v. fall])

Auxiliaries and root verbs form English verb phrases. There can be zero or more auxiliaries in a verb phrase. The root verb is either suffixed or it takes any of the three forms (present, past, past participle). But in Bengali, verb phrases are formed by appending appropriate suffixes to the verb stem.

TABLE 13: BENGALI VERB PARADIGM LIST

Tense	3PN	3/2PF	2PN	2PI	1P
Pres Simp	+y	+n	+o	+s	+I
Pres Prog	+cche	+cchen	+ccho	+cchis	+cchi
Pres Perf	-1+ eyeche	-1+ eyechen	-1+ eyecho	-1+ eyechis	-1+ eyechi
Imper Pres	+k	+n	+o	+φ	--
Past Simp	-1+elo	-1+elen	-1+ele	-1+eli	-1+elam
Past Prog	+cchilo	+cchilen	+cchile	+cchilis	+cchilam
Past Perf	-1+ eyechilo	-1+ eyechilen	-1+ eyechile	-1+ eyechilis	-1+ eyechilam
Past Hab	-1+eta	-1+eten	-1+ete	-1+etis	-1+etam
Fut Simp	+be	+ben	+be	+bi	+bo
Imper Fut	+be	+ben	-1+eyo	+s	--

Again, some verbs in Bengali use a combination of a semantically "light" verb and another meaning unit (a noun, generally) to convey the appropriate meaning. New verbs are often translated to Bengali by adding a light verb like "করা" (kara[do]), "নেওয়া" (neoa [take]) to the transliterated form of the source language word. An example is, to fax - "ফ্যাক্স করা". Examples of Bengali compound verbs are listed in the Table 12.

Bengali verbs are morphologically very rich. A single verb has many morphological variants. A separate morphological paradigm suffix table is maintained for each spelling pattern of the Bengali root verb; generally all Bengali root verbs, belonging to the same spelling pattern category, follow the morphological paradigm suffix table corresponding to that particular spelling pattern category.

There are 20 such spelling pattern categories for Bengali. These suffixes also vary in the Classical and Colloquial form of Bengali. In the present work, the Bengali root verb classification proposed by (Chatterjee, 1992) has been maintained. The desired verb phrase (for the new fused sentence) has been obtained by simple agglutination rule as described in the Table 13.

The '-' symbol, followed by a number, at the start of an entry in the morphological paradigm suffix table indicates that the specified number of characters have to be trimmed from the right of the stem, and then the suffix which follows the symbol '+' is added to it.

VII. SENTENCE FUSION

The sentence fusion task involves two basic steps. The first step is the information redundancy identification and the second step is sentence type wise generation.

A. Case Base Information Compression

The basic information of semantic constituents has been identified based on CASE grammar semantic roles. Case markers like Agent, Instrument, Dative, Factitive and Locative serve as a good checklist for information redundancy check between two given sentences. Finally a Case based information list has been made and the information has been generated through the sentence level generation process. An example is shown in Figure 5.

B. Sentence Type Wise Generation

Final sentence level fusion mechanism is based on the type of sentences. Rules have been developed accordingly. The details are described in the following sub sections.

1) Simple-Simple Fusion

There is only one principal clause in a simple sentence. Hence after sentence fusion the generated sentence may be complex or compound. If any infinite verb is present in any of the simple sentences then a complex sentence is generated by the system otherwise simple conjunct is placed between the two simple sentences to generate a compound sentence. Clause ordering is a vital issue for readability. For complex sentence generation a simple sentence with no infinite verb is kept in the first position. If there is any Tense-Aspect-Modality mismatch between verbs of the two sentences then it is changed accordingly by the rules described in Verb Phrase Generation section.

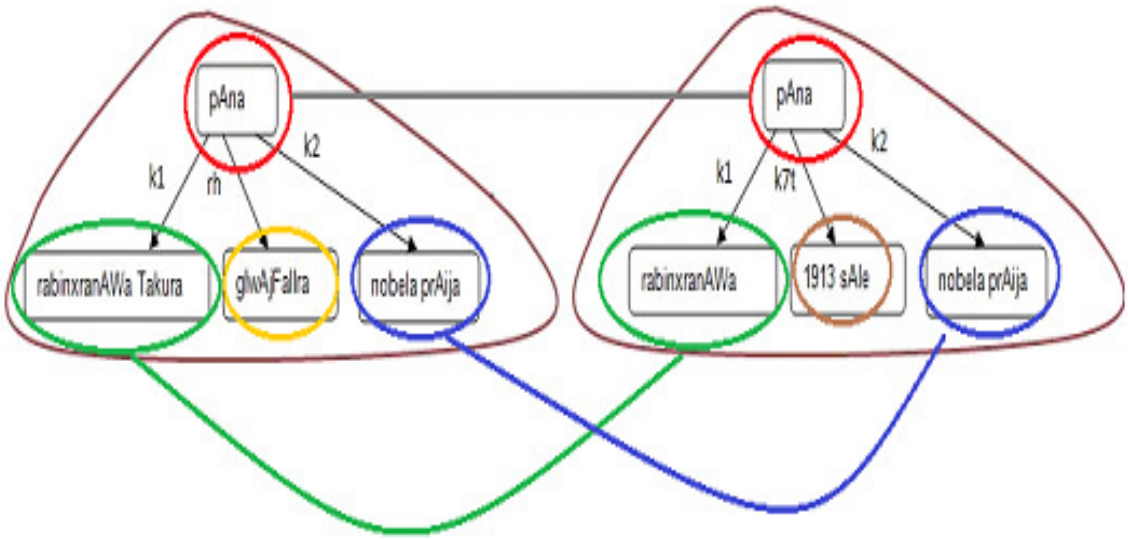
Sentence 1: (rabinxranAWa Takura) (gIwAjFalIra jANya) (nobela prAija)(pAna).
English: Rabindranath Thakur got Noble prize for Gitanjali.

Sentence 2:(rabinxranAWa) **(1913 sAle)**(nobela prAija) (pAna).
English: Rabindranath got Noble prize in 1913.

Fused Sentence: (rabinxranAWa Takura)**(1913 sAle)** (gIwAjFalira jANya)(nobela prAija) (pAna).
English: Rabindrantha Thakur got Noble prize in 1913 for Gitanjali.

2) Simple-Complex Fusion

Sentence fusion among simple and complex sentences produces a complex sentence. The principal clause in the complex sentence is treated as a simple sentence and first simple-simple sentence fusion technique is applied. Finally the dependant clause or the complex predicate is fused to the simple-simple adjoin.



Agent:	rabinxranAWa (Rabindranath), rabinxranAWa Takura (Rabindranath Thakur)
Instrument:	gIwAjFalIra jANYa (for Gitanjali),-
Dative:	-, 1913 sAle (in 1913)
Factitive:	pAna (got), pAna (got)
Locative:	-,-
Objective:	nobela prAija (Noble prize), nobela prAija(Noble prize)

Figure 5: Case Based Information Compression

3) Simple-Compound Fusion

Compound sentences generally consist of two or more principal clauses. Case based information compression technique finds out the principal clause that has any redundant information with the simple sentence. Then the sentences are merged.

4) Complex-Complex Fusion

Complex-complex fusion leads to various challenges during the development stage. Generally complex sentence consists of one independent (principal clause) and possibly more than one independent clause. The possibilities are that the two sentences have redundancy in their principal clauses or the two sentences have redundancy in one's principal clause and dependent clause of the other. Therefore two sets of rules have been defined. For dependent-dependent relation simply simple-simple fusion rules have been applied.

waKana BebeCi, era parera yAwrA habe xakRiNamerura xike, yeKAne SuXu baraPa, sArA xina rAwa, sArA baCara.

English: Then I thought that next trip will be towards South Pole, where there is only frost, whole day night, whole year.

5) Complex-Compound Fusion

For complex compound fusion the system first identifies the non-redundant dependant predicate and adds it to the principal clause of the complex sentence.

6) Compound-Compound Fusion

During compound-compound fusion system first identifies the redundant clause and then adds a conjunct to generate the resultant compound sentence.

Sentence 1: (ye hewu) (bAmapanWIxera dAkA BArawa banXa) (wAi) (paScimabafgera bAmaPanWI sarakAra) (sarakAri sArkulAre) (nirapekRawA) (bajAyZa rAKAra) (ceRtA kareCe).

English: As it is Bharat Bandh called by leftists therefore the leftist government of West Bengal is trying to be impartial in the government circular.

Sentence 2: (bAmPanWIra 16 wAriKera banXa) (GoRanA hayZeCe) (ebaM) (sarakAra) (sarakAri sArkulAre) (nirapekRawA) (bajAyZa rAKAra) (ceRtA kareCe).

English: Leftists have declared the Bandh on 16th and the Government is trying to be impartial in the government circular.

Fused Sentence: (ye hewu) (16 wAriKa) (bAmapanWIxera banXa) (wAi) (paScimabafgera bAmaPanWI sarakAra) (sarakAri sArkulAre) (nirapekRawA) (bajAyZa rAKAra) (ceRtA kareCe).

English: As leftist called Bandh is on 16th therefore the leftist government of West Bengal is trying to be impartial in the government circular.

VIII. EVALUATION

Generally natural language generation techniques always face the readability issues. Instead of one evaluation methodology we use two techniques to evaluate the performance of the present system.

TABLE 14: HUMAN EVALUATION OF THE PRESENT SYSTEM

	Simple	Complex	Compound
Simple	5	3	5
Complex	3	2	4
Compound	4	4	5
Overall	4	3	4.6

The first technique is based on standard BLEU (Bilingual Evaluation Understudy) score and the second one is direct human evaluation score based technique. We use a 1-5 scoring technique in human evaluation whereas 1 denotes very poor, 2 denotes poor, 3 denotes acceptable, 4 denotes good and 5 denotes excellent. The results are reported in Table 14 and 15 respectively. Table 14 is similar to the Table 3 as the gold standard has been made sentence type pair wise.

IX. CONCLUSION

The present paper describes the overall process of sentence fusion techniques for Bengali. The process could be replicated for other Indian languages too. The result section shows the effectiveness of the proposed techniques. The readability issues are still there and demands more research to find out more sophisticated techniques. Future works will focus towards the generation of more readable outputs from the sentence fusion system.

All the techniques reported in this paper are syntactic in nature. It is felt during error analysis of the present system that the semantic aspect of the sentence fusion techniques has to be looked into.

ACKNOWLEDGMENT

The work reported in this paper is supported by a grant from the India-Japan Cooperative Programme (DST-JST) 2009 Research project entitled “Sentiment Analysis where AI meets Psychology” funded by Department of Science and Technology (DST), Government of India.

. TABLE 15: BLEU SCORES

	Simple					Complex					Compound				
BLEU	**1**	**2**	**3**	**4**	**Avg**	**1**	**2**	**3**	**4**	**Avg**	**1**	**2**	**3**	**4**	**Avg**
Simple	68	70	76	81	73.75	50	54	57	61	55.50	62	64	71	78	68.75
Complex	50	54	57	61	55.50	42	46	54	56	49.50	55	58	63	69	61.25
Compound	62	64	71	78	68.75	55	58	63	69	61.25	61	65	69	73	67.00
Overall	66.00					55.41					65.66				

REFERENCES

[1] E. Reiter and R. Dale. Building Natural Language Generation Systems. Cambridge University Press, Cambridge, 2000.

[2] R. Chandrasekar and S. Bangalore. Automatic induction of rules for text simplification. Knowledge-based Systems, 10(3):183–190, 1997.

[3] K. Knight and D. Marcu. Summarization beyond sentence extraction: A probabilistic approach to sentence compression. Artificial Intelligence, 139(1):91–107, 2002.

[4] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. In Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, pages 545–552, Sapporo, 2003.

[5] J. Robin. Revision-based generation of Natural Language Summaries Providing Historical Back-ground. Ph.D. Thesis, Columbia University, 1994.

[6] C. Callaway and J. Lester. Dynamically improving explanations: A revision-based approach to explanation generation. In Proceedings of the 15th Inter-national Joint Conference on Artificial Intelligence (IJCAI 1997), pages 952–958, Nagoya, Japan, 1997.

[7] Barzilay, Regina, Kathleen R. McKeown, and Michael Elhadad. 1999. Information fusion in the context of multi-document summarization. In Proceedings of the 37th Annual Conference of the Association for Computational Linguistics, College Park, MD. Association for Computational Linguistics, New Brunswick, NJ.

[8] Kathleen McKeown, Sara Rosenthal, Kapil Thadani and Coleman Moore. Time-Efficient Creation of an Accurate Sentence Fusion Corpus. In Proceeding of Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the ACL, pages 317–320, Los Angeles, California, June 2010.

[9] James Clarke and Mirella Lapata. Models for sentence compression: a comparison across domains, training requirements and evaluation measures. In Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics. Pages: 377 - 384. 2006.

[10] Chatterji, S. K. 1995. Bhasa Prakash Bangla Vyakaran. Kolkata: Rupa Publications.

[11] 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.

[12] 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.

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

[14] Kiparsky, Paul and J. F. Staal (1969). ‘Syntactic and semantic relations in Panini.’ Foundations of Language 5, 83-117.

[15] Fillmore, Charles (1968). The Case for Case. In Universals in Linguistic Theory. New York: Holt, Rinehart, and Winston. 1-88.

[16] Chomsky, Noam (1956). "Three models for the de-scription of language". IRE Transactions on In-formation Theory 2: 113–124.