Syntactic Sentence Fusion Techniques for Bengali

Abstract—The present paper describes various syntactic sentence fusion techniques for Bengali, belongs to Indo-Aryan group. Firstly a clause identification and classification system marks clause boundaries as well as principle clause and subordinate clause. A rule-based sentence classification has been developed to categorize sentences as simple, complex and compound. The final fusion system make use of sentence class and clause type and finally fuse two textually entailed sentences involving verb paradigm information and noun morphological generator. The output of the present sentence fusion technique has been tested with manually created gold standard. The result proves that the technology is promising and may be used for any other Indian languages too.

Keywords—Put your keywords here, keywords are separated by comma.

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” (Reiter and Dale, 2000). Recently, there is an increased interest in NLG applications that produce meaningful text from meaningful text rather than from abstract meaning representations. Such applications are sometimes referred to as text-to-text generation applications (Chandrasekar and Bangalore, 1997), (Knight and Marcu, 2002), (Lapata, 2003), and may be likened to earlier revision-based generation strategies (Robin, 1994) (Callaway and Lester, 1997). Text-to-text generation is often motivated from practical applications such as summarization (Barzilay et al., 1999), sentence simplification (McKeown et al., 2010), and sentence compression (Clarke and Lapata, 2006). One reason for the interest in such generation systems is the possibility to automatically learn text-to-text generation strategies from corpora of parallel text or semantically texts.

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

According to best of our knowledge there is no significant effort could be found for sentence fusion in Indian languages. Therefore it will be unfair to compare the present task with the existing technologies for other languages. As mentioned earlier that Bengali (may be other Indian languages too), has own peculiarity and thus it demand special problem definition and solution architecture.

The other aspect of challenges should be addressed that Bengali is a electronically resource constrain language. Therefore many linguistic analysis tools have to build in order to build the final fusion system. A clause analysis (include: clause boundary identification and classification system) and sentence categorization (classify sentences either as simple, complex or compound categories) have been built.

II. RESOURCE ACQUISITION

Resource acquisition is one of the most challenging obstacles to work with resource constrained languages like Bengali. Therefore some of the tools are collected from publicly available resources and some are developed. In this section we will describe particularly corpus acquisition and annotation. Additionally we will give necessary information about those publicly available tools, we have collected and used successfully.

A. Corpus

1) Corpus for Clause Analysis

The NLP TOOLS CONTEST: ICON 2009 dependency relation marked training dataset has been used. The corpus has been further annotated at the clause level: Principal clause and Subordinate clause. According to standard grammar (Chatterji, 1995) 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 Sanchay2. A brief statistics of the corpus are reported in Table 1.

<table>
<thead>
<tr>
<th>No of Sentences</th>
<th>TRAIN</th>
<th>DEV</th>
<th>TEST</th>
</tr>
</thead>
<tbody>
<tr>
<td>980</td>
<td>150</td>
<td>100</td>
<td></td>
</tr>
</tbody>
</table>

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>
<thead>
<tr>
<th>BOUNDARY</th>
<th>TYPE</th>
</tr>
</thead>
<tbody>
<tr>
<td>76.54%</td>
<td>89.65%</td>
</tr>
</tbody>
</table>

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

---

1 http://ltrc.iiit.ac.in/nlptools2009/CR/all-papers-toolscontest.pdf
2 http://ltrc.iiit.ac.in/nlpaicontest07/Sanchay/
such cases is Inquisitive semantic (Groenendijk, 2009) cases, ambiguous for human annotators 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.

3) Corpus for Sentence Fusion

For the sentence fusion task a NEWS corpus has been manually created. Two popular Bengali news papers have been chosen as the information resource. Total 50 parallel stories have been picked up randomly. As the present system works by analysing sentence type thus sentences are chosen based on their type (simple, complex or compound). Total 300 sentences have been chosen on the basis of sharing redundant information on common topic.

Three annotators have been asked to create manually fused sentences. As the present system is rule based in nature therefore 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 pair of the particular type has been chosen. Specifically 48 simple-simple, 53 simple-complex, 46 simple-compound, 58 complex-complex, 50 complex-compound and 45 compound-compound sentence pair have been chosen from both the newspaper. Row represents the newspaper1 and column depicts newspaper 2.

<table>
<thead>
<tr>
<th></th>
<th>Simple</th>
<th>Complex</th>
<th>Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>48</td>
<td>53</td>
<td>46</td>
</tr>
<tr>
<td>Complex</td>
<td>53</td>
<td>58</td>
<td>50</td>
</tr>
<tr>
<td>Compound</td>
<td>46</td>
<td>46</td>
<td>45</td>
</tr>
</tbody>
</table>

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). The dependency parser follows the tagset identified for Indian languages as a part of NLP TOOLS CONTEST 2009 as a part of ICON 2009.

III. Clause Analysis

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 (Chatterji, 1995) 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. 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.

A. Rule-based Clause Boundary Identification

Analysis of a Bengali corpus and standard grammar (Chatterji, 1995) 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 understand 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 of disambiguation. These types

3 http://www.anandabazar.com/
4 http://www.sangbadpratidin.net/
5 http://crf.sourceforge.net/
of language properties make the clause identification problem difficult.

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

1) Karaka relation

Dependency parsing assigns the inter chunk relationships 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, every chunk has a dependency relation with the verb chunk. These relations are called 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 Figure 1.

Animesh approached and saw that (among the girls) one was in front of him.

animeRa xeKala egiyZe (meyZexera maXye) ekajan wAra sAmne

Figure 1: Karaka Relations

2) Compound Verbs

In every Indian languages and especially 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.

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.

(If you look), how can I bath?

xeKala (karaba snAna) koWAvza?

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 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.

(Do you know) what is their news?

jAne wAxera Kabara ki?

Figure 3: Shasthi Relation
B. Case Grammar/Karaka Relations for Clause Boundary Identification

The classical Sanskrit grammar Astadhyayi (‘Eight Books’), written by the Indian grammarian 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 morphosyntax.

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-space orienting 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 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 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.
```

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.

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.

Rama came to meet when he was going to market.

In the previous case system marks “রামা / 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.

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.

যামার বাঁধায় সুর মাঝারুদক কর।
The music of Shyam’s messmerized me.

3) **Dative**

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 4. Still there are many difficulties to identify special cases of relative time expressions. As an example:

When moon rise we will start our journey.

In the previous example the relative time expression is “উর্ধ্বনীরিতন আমারা রামা হয়” (When rise/When stop…) 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.

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 an ambiguity among Agent, Dative and Locative case marker as they orthographically generates same type of surface form (using common suffixes as: ০, ০র 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.

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.

### Table 4: Categories of Time Expressions

<table>
<thead>
<tr>
<th>Bengali</th>
<th>English Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মুরনিরি, মুরনিরি</td>
<td>Morning/evening/night/dawn</td>
</tr>
<tr>
<td>নামনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>O clock/time/hour/minute/second</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Monday/Tuesday/Sunday…</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Bengali months…</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>January/February…</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Day/month/year…</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Long time/moment…</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Morning/evening/night/dawn</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Upcoming/</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>When rise/When stop…</td>
</tr>
</tbody>
</table>

### Table 5: Categories of Causative Expressions

<table>
<thead>
<tr>
<th>Bengali</th>
<th>English Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>ভয়ড়_নেব</td>
<td>Hence/Reason/Reason</td>
</tr>
<tr>
<td>ভয়ড়_নেব</td>
<td>If else</td>
</tr>
</tbody>
</table>

### Table 6: Categories of Locative Expressions

<table>
<thead>
<tr>
<th>Bengali</th>
<th>English Gloss</th>
</tr>
</thead>
<tbody>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Morning/evening/night/dawn</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>Upcoming/</td>
</tr>
<tr>
<td>মান্ত্রেরনিরি, মোরনিরি, মোরনিরি, মোরনিরি</td>
<td>When rise/When stop…</td>
</tr>
</tbody>
</table>

IV. 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.

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 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 "inmate" 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.

B. 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.

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 used here to statistically capture the prosodic structure of the language. The features experimentally found useful are chosen as 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 the any clause pattern.

3) Word

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

D. Performance of Clause Identification and Classification

<table>
<thead>
<tr>
<th>Type</th>
<th>Performance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>88.12%</td>
</tr>
<tr>
<td>Complex</td>
<td>78.07%</td>
</tr>
<tr>
<td>Compound</td>
<td>82.50%</td>
</tr>
</tbody>
</table>

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

### Table 9: Confusion Matrix

<table>
<thead>
<tr>
<th></th>
<th>Simple</th>
<th>Complex</th>
<th>Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>88.12%</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Complex</td>
<td>-</td>
<td>78.07%</td>
<td>8.02%</td>
</tr>
<tr>
<td>Compound</td>
<td>-</td>
<td>3.5%</td>
<td>82.50%</td>
</tr>
</tbody>
</table>

VI. MORPHOLOGICAL GENERATION

Two types of basic morphological generation techniques that mainly adopted for the present task. First is noun generation and second is verb generation.

A. Noun Phrase Generation

Noun could 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 role could be identify from various existing resources like Propbank, FrameNet or VerbNet. But unfortunately no such resources exist for Indian languages. Therefore the present task only based on the basic semantic role types of the classical theory, which is remaining useful for modern settings. Based on the Fillmore’s Case grammar and semantic roles different suffix list have been prepared. Depending upon the semantic role of a noun in the targeted sentence these suffix agglutinated with the root form, obtained from the output of the morphological analyzer. Few examples of such suffixes are reported in the Table 10.

### Table 10: Case-Wise Suffix List

<table>
<thead>
<tr>
<th>Case</th>
<th>Suffix</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Agent</td>
<td>ে,ের,ক,বা,বু</td>
<td>রাসের, শ্যামক</td>
</tr>
<tr>
<td>Instrument</td>
<td>পিটি,ধারা, চাঁচ,চিন</td>
<td></td>
</tr>
<tr>
<td>Dative</td>
<td>সম, দিনি,, পেশ,দিন</td>
<td></td>
</tr>
<tr>
<td>Factive</td>
<td>ে,ের</td>
<td>ভাসর</td>
</tr>
<tr>
<td>Locative</td>
<td>তে,র,</td>
<td>কলকাতায়,রাস্তায়</td>
</tr>
</tbody>
</table>

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 (apni), “তুমি গেল” (tumi gele), “তুই গেলি” (tui geli)/ and, likewise, plural form
New verbs are often translated to Bengali by adding a light verb like “করা” (kara[do]), “হওয়া” (neowa[v. take]) to the transliterated form of the source language word. Examples are, to fax - “ফেকার” (fax) and reported in Table 12.

<table>
<thead>
<tr>
<th>English verbs</th>
<th>Bengali verbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>Swim</td>
<td>স্নাতার” (santar [n. swimming]) “কাটা” (kata[v. cut])</td>
</tr>
<tr>
<td>Try</td>
<td>“চেষ্টা” (chesta [n. try]) “কারা” (kara[v. do])</td>
</tr>
<tr>
<td>rest</td>
<td>“বিশ্রাম” (bishram [n. rest]) “হওয়া” (neowa[v. take])</td>
</tr>
<tr>
<td>fail</td>
<td>“বায়া” (byartho [adj. futile]) “হওয়া” (bowa[v. bel])</td>
</tr>
<tr>
<td>rain</td>
<td>“বৃষ্টি” (brishti [n. rain]) “পড়া” (pora[v. fall])</td>
</tr>
</tbody>
</table>

Bengali verbs are morphologically very rich. A single verb has many morphological variants. To deal with these types of complex phenomena we simply believe in standard grammar. A separate morphological paradigm suffix table is maintained for each such spelling pattern category; and 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 groups 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 Table 13.

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

VII. SENTENCE FUSION

For the sentence fusion there is two basic steps. First information redundancy identification and the second is sentence wise generation.

A. Case Base Information Compression

The basic information of semantic constituents has been identified based on CASE grammar semantic roles. Case markers: Agent, Instrument, Dative, Factitive, Locative are serving as a good checklist for information redundancy among two given sentences. Finally a Case base information list has been made and the information has been generated through the sentence level generation process. An example could be seen in Figure 4.

B. Sentence Wise Generation

Final sentence level fusion mechanism works on type of sentences. Rules are made accordingly. The details could be found in the following sub sections.

1) Simple-Simple Fusion

There is only one principal clause in every simple sentence. Henceforth after sentence fusion the generated sentence may be complex or compound. If there is any indefinite verb exists in any of the sentences then a complex sentence has been produced by the system otherwise simple conjunct has been added between two simple sentences to produce compound sentence. Clause ordering is a vital issue for readability. For
complex sentence generation a simple sentence with no indefinite verb keep in the first position. If there is any Tense-Aspect-Modality mismatch between verbs of two sentences then it has been changed accordingly by the rules described in Verb Phrase Generation section.

Sentence 1: (rabinxranAWa Takura) (gLwAjFalIra) (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 for Gitanjali.

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

2) Simple-Complex Fusion

Sentence fusion among simple and complex sentence produces a complex sentence. Is there any principal clause exists in a complex sentence then it is treated as simple sentence and first simple-simple sentence fusion technique has been applied. Finally the dependant clause or complex predicate has been fused to the simple-simple adjoin.

3) Simple-Compound Fusion

Compound sentences are generally consists of two or more principal clauses. Case base information compression technique finds out the principal clause has any information redundancy 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 dependant (principal clause) and one independent clause. There are two possibilities that the two sentences have redundancy in their principal clause or the two sentences have redundancy in one’s principal clause and dependant clause of either. Therefore two set of rule has been defined. For dependant-dependant relation simply simple-simple fusion rules have been applied (as described in 7.2.1).

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 identify the non-redundant dependant predicate and add 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 added conjunct to produce a resultant compound sentence.

Sentence 1: (ye hewu) (bAmapanWIxera dAkA BArawa banXa) (wAi) (paScimabafgera bAmaPranta sarakAra sarakAri) (sArkulAre nirapekRawA) (bajAyZa rAKAra) (ceRtA kareCe).
English: As it is leftist holiday therefore the leftist government of West Bengal try to keep biasedness in government circular.

Sentence 2: (bAmPrantera 16 wAriKera banXa) (GoRanA hayZeCe) (ebaM) (sarakAra) (sarakAri sArkulAre nirapekRawA) (bajAyZa rAKAra) (ceRtA kareCe).
English: Left government declared 16th as a holiday the government try to keep biasedness in government circular.
English: As 16th is leftist holiday therefore the leftist government of West Bengal try to keep biasedness in government circular.

VIII. EVALUATION

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

<table>
<thead>
<tr>
<th>Table 14: Human Evaluation of the Present System</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
</tr>
<tr>
<td>Simple</td>
</tr>
<tr>
<td>Complex</td>
</tr>
<tr>
<td>Compound</td>
</tr>
<tr>
<td>Overall</td>
</tr>
</tbody>
</table>

The first one 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. QUALITY ANALYSIS

As we stated at very first section that all techniques reported in this paper is syntactic. What we feel during error analysis of the present system is that the main missing criteria are the semantic aspect of the problem.

X. 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. Although the readability issues are still there and demands more research to find out more sophisticated techniques. As research is an endless endeavour therefore our next study will be towards more readable outputs of a sentence fusion system.

. Table 15: BLEU Scores

<table>
<thead>
<tr>
<th>BLEU</th>
<th>Simple</th>
<th>Complex</th>
<th>Compound</th>
</tr>
</thead>
<tbody>
<tr>
<td>Simple</td>
<td>0.68</td>
<td>0.70</td>
<td>0.76</td>
</tr>
<tr>
<td>Complex</td>
<td>0.50</td>
<td>0.54</td>
<td>0.57</td>
</tr>
<tr>
<td>Compound</td>
<td>0.62</td>
<td>0.64</td>
<td>0.71</td>
</tr>
<tr>
<td>Overall</td>
<td>66.00</td>
<td>55.41</td>
<td>65.66</td>
</tr>
</tbody>
</table>

REFERENCES