English to Bangla Machine Translation System Using Context-Free Grammars

Shibli Syeed Ashrafi[1], Md. Humayun Kabir[2], Md. Musfique Anwar[3] and A. K. M. Noman[4]

[1] **United Commercial Bank Limited (UCBL)**
Gulshan 2, Dhaka, Bangladesh

[2,3] **Dept. of Computer Science and Engineering, Jahangirnagar University**
Savar, Dhaka, Bangladesh

[4] **Purbani Group**
Gulshan 1, Dhaka, Bangladesh

Abstract

We present a machine translation (MT) system which translates an English sentence into a Bangla sentence of equivalent meaning. We have defined context-free grammars (CFGs) for English simple assertive sentences to discover the syntactic structures of both English and Bangla correct and complete sentences. The CFGs provide the basis for parse rules for rewriting the syntactic structures during bottom-up parsing. A bi-lingual dictionary provides the morphological properties and contextual information of the English words with their corresponding Bangla meaning. The proposed MT system requires transferring English language structure to the corresponding Bangla language structure with an approximate lexical meaning mapping (ALMM) of the English words. The system incorporates a module to verify the accuracy of the constructed Bangla sentence, and allows manual refinement if required. We have implemented a prototype of this system, and have applied to a large number of simple assertive sentences, which shows promising results.

Keywords: *Syntactic Structure, Parse Rules, Bottom-up Parsing, Syntactic Transfer, Bi-lingual Dictionary, Approximate Lexical Meaning Mapping.*

1. Introduction

An MT system generates a target language output sentence by translating a source language sentence. The overall system involves analysis of the input sentence in the source language to discover its grammatical structure and transfer it to target language structure. A natural language translation system requires a source language module to analyze source language sentence, and a transfer module along with a generation procedure to obtain a target language sentence of equivalent meaning [1]. The component words in the source language sentence are translated to obtain the meaning in the target language based on the contextual information, e.g., person, number, auxiliary verb etc. appearing in the source language sentence. The meanings are then replaced in the target language structure to generate the output sentence of equivalent meaning. The generated output may need to be refined manually to obtain correct output sentence.

Various approaches exist for developing MT systems: i) syntactic ii) semantic, and iii) lexica list for natural language translation [1]. In MT system, every sentence is decomposed into multiple phrases of different categories based on the morphological and syntactic properties of different constituent words [2]. The quality and capability of an MT system largely depends on the size and quality of the bi-lingual dictionary [3]. The MT dictionary should contain a large collection of source language words with their meaning in the target language, and the morphological properties of the source language words. The selection of meaning of the source language word from the MT dictionary and to incorporate the meaning in the target language sentence requires intelligent algorithms. English to Bangla machine translation system employing heuristics or artificial intelligence in algorithms for word recognition and translation [4] has been developed to translate English sentence(s) to corresponding Bangla sentence(s). Mapping rules have also been defined and used in the development of MT system [5] to obtain the equivalent Bangla grammatical structure of the structure of an input English sentence. This MT system [5] can handle only affirmative sentence in the present indefinite and continuous forms of English sentences.

[3] Md. Musfique Anwar is on study leave for pursuing his MS program abroad.

In this paper, we propose a bi-lingual MT system for Bangla translation of an English simple assertive sentence employing structural analysis using grammatical rule-based approach in the form of context-free grammars (CFGs) [1,5,6]. The proposed MT system uses *approximate lexical meaning mapping* (ALMM) approach as the Bangla meaning of an English word is actually an approximation depending on the contextual information. There is no individual Bangla meaning of the English auxiliary verbs, e.g., am/is/was/will etc. The Bangla meaning of the English main verb is inflected by the presence of the auxiliary verbs. We use transfer (move) and translate (convert) for two different purposes. An input English sentence is parsed to extract its structural representation in its simple form. A prototype has been developed for computer evaluation of the proposed MT system. A bi-lingual dictionary provides the morphological properties of the English words and their meanings in Bangla. The sentence and phrase structures are represented in terms of CFG rules [1,2,5,6] both for English and Bangla languages.

2. Related Work

MT system can work between two or more natural languages spoken in the world by human being. For simplicity of presentation, we consider only Bangla and English natural languages. Development of MT systems for English to Bangla [4,5] and Bangla to English [1-3,7,8] machine translation employs a parsing technique. The parser module determines the grammatical structure of the source language sentence [6] in terms of parts of speech, e.g., noun (N), pronoun (PN) etc. [8,9] using CFG rules.

Definition 1 (CFG Rules for English). The CFG rules which are used to parse an input English sentence in the proposed MT system are defined in Table 1 [1,2,5,6].

Definition 2 (CFG Rules for Bangla). The CFG rules which are used to parse a Bangla sentence in the proposed MT system are defined in Table 2.

The CFG for Bangla language is defined to make it analogous to English CFG in terms of CFG variable naming. Bangla language has its own terminologies for each CFG component.

In the proposed system, Bangla to English character mapping technique is used to represent Bangla words in terms of fixed pattern of English character sequences. For example, the English character sequence *rdpckcY* is used to represent the Bangla word পড়িতেছে.

Table 1: Sample CFG Rules for English Language

CFG rules			CFG rules		
S	→	NP VP | NP VP ADV | …	PER-PN	→	I | you | he | she | we | …
NP	→	N | PN | CN | …	INTG-PN	→	who | what | whom | …
VP	→	MV | CV | AV CV | …	AV	→	am | is | was | will | …
CV	→	MV NP* | MV PP | …	MV	→	read | reading | drink | …
PP	→	PREP CN* | PREP PN | …	ADV	→	very | slowly | not | …
CN	→	DEF-ART N | DEF-ART ADJ N | INDEF-ART N | INDEF-ART ADJ N | …	ADJ	→	good | bad | red | one | …
			PREP	→	about | after | before | …
			DEF-ART	→	the
			INDEF-ART	→	a | an
			CONJ	→	and | or | …
PN	→	PER-PN | INTG-PN | …			
N	→	boy | girl | book | tea | …			

Abbreviations**:** S → Sentence, NP → Noun Phrase, VP → Verb Phrase, PP→ Prepositional Phrase, CN → Complex Noun, CV→ Complex Verb, DEF-ART→ Definite Article, INDEF-ART→ Indefinite Article, N → Noun, PER-PN → Personal Pronoun, INTG-PN → Interrogative Pronoun, AV → Auxiliary Verb, MV → Main Verb, ADV → Adverb, ADJ → Adjective, PREP → Preposition, CONJ → Conjunction. * indicates any number of occurrence(s).

Table 2 uses Bangla words directly instead of using the English character sequences for easier interpretation. Within the set of CFG rules, we removed the rule ART -> DEF-ART | INDEF-ART to eliminate the difficulty in handling definite and indefinite articles in Bangla. The given English CFG rule for CN involving DEF-ART is valid only for singular form of nouns, but heuristics are needed to apply for plural form of nouns to eliminate

DEF-ART as redundant term to obtain equivalent Bangla CFG rule for CN as the DEF-ART *The* does not contribute explicitly to the Bangla meaning of the plural noun.

Table 2: Sample CFG Rules for Bangla Language

CFG rules			CFG rules		
S	→	NP VP | NP ADV VP | …	PER-PN	→	আমি | তুমি | সে | আমরা | …
NP	→	N | PN | CN | …	N	→	বালক | বালিকা | ফুল | বই | চা | …
VP	→	MV | CV | …	INT-PN	→	কে | কি | কাহাকে | …
CV	→	NP* MV | PP MV | …	MV	→	পড়ি | পড়িতেছে | পান করিতেছে | …
PP	→	PREP NP | …	ADV	→	অত্যন্ত | ধীরে ধীরে | …
CN	→	N DEF-ART | ADJ N DEF-ART | INDEF-ART N | INDEF-ART ADJ N | …	ADJ	→	ভাল | মন্দ | লাল | একটি | …
			PREP	→	সম্পর্কে | পরে | পূর্বে | …
			DEF-ART	→	টি | টা | …
			INDEF-ART	→	একটি | …
PN	→	PER-PN | INTG-PN | …	CONJ	→	এবং | ও | আর | অথবা | …

During parsing, a token list is produced and a parse tree can be generated from the grammatical extracts of the parsing steps applied to the source language sentence. This token list is input to the MT system along with the grammatical structure to generate the output sentence in the target language [6].

2.1 Morphological Analysis

In the development of MT systems, morphological analysis of the component words of a source language sentence plays a vital role in the translation process [8,10]. Most of the words of a natural language have roots, and others are inflectional variants of the roots formed with affixes. This property allows the MT system to find a correspondence between the strings of two languages for the application of some heuristics for easier translation. Morphological study helps to determine the minimal unit of meaning of a word and identifies the structure and rules for formation of words [8,10]. As an example, consider the English words 'determines', 'determined' and 'determining', which are recognized as having the same stem *determine* and different endings with *s*, *d*, and *ing* with suppressing *e*.

2.2 Syntactic Analysis and Transfer

In any natural language, every sentence is formed with a single phrase or a combination of multiple phrases, e.g., NP, VP etc. An input sentence is analysed to discover the grammatical structure based on a formal grammar of the source language [2]. During syntactic analysis, different groups of related words called phrases are identified. The syntactic structure of different types of sentences can be described using a tree in which the root is the sentence S, the descendants of the root are the phrases, and the elementary words are at the leaf. During syntactic analysis, at first, the sentence type is identified, and different component phrases belonging to a sentence are separated by grouping different grammatical elements [2]. For example, a NP may be derived by bottom-up parsing of an English sentence from the combination DEF-ART N, or only PN.

Syntactic transfer techniques have been used to develop MT systems for Bangla-English machine translation [1,8]. Syntactic transfer MT system [1] applies a tree-to-tree transformation to the parse tree generated from the input language sentence to obtain the corresponding grammatical structure of the output sentence in target language. The Bangla translation of the English words of the input sentence is done at the phrasal level. The example depicted in Fig. 1 demonstrates English to Bangla translation of the English sentence *The boy is drinking tea*.

The example shows the switching of verb and object of the verb of the sentence *The boy is drinking tea* in the translated Bangla sentence বালকটি চা পান করিতেছে. In English, the object *tea* appears after the verb *drinking*, whereas in Bangla, the object চা appears before the verb পান করিতেছে, i.e., *Subject Verb Object* (S V O) is transformed to Subject Object Verb (S O V) [8]. Another important transfer of grammatical information occurs in head switching [1] while translating the English sentence to Bangla sentence. In the above example, the auxiliary verb *is* appears before the main verb *drinking* (i.e., verb root: *drink* + affix: *ing*), which is in the continuous form of present tense. As the subject *The boy* is non-honorific third person, singular number and the main verb *drinking* is in present continuous form, the main verb *drinking* is translated to পান করিতেছে using the bi-lingual dictionary. More examples on English to Bangla translation of simple, complex and compound sentences can be found in [11,12].

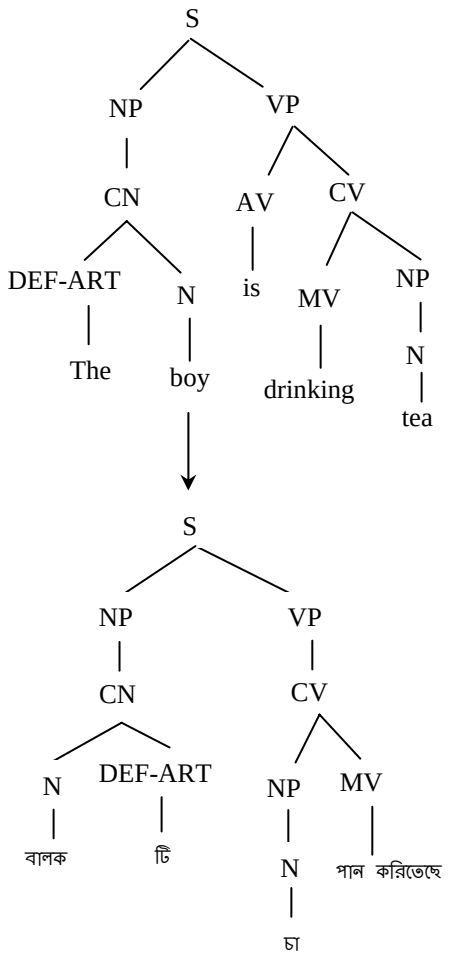

Fig. 1. English to Bangla Translation of *The boy is drinking tea*.

2.3 Role of Bi-lingual Dictionary in MT System

A bi-lingual dictionary [5] is a key component of any MT system. Each entry stored in a bi-lingual dictionary forms a part of a sentence, which helps to determine the grammatical structure and meaning of the input sentence. In English-Bangla MT system, the bi-lingual dictionary serves the purpose of a lexicon by storing the root words together with their derivatives, morphological properties and meaning of the English words in Bangla depending on the auxiliary verbs, tense, and persons used in the input English sentence. The proposed MT system maintains three separate storage structures for English words, which are verbs, nouns and other categories together. Table 3 represents the storage details of some English verbs with their Bangla meaning.

3. Generation and Transformation of Syntactic Structures

To obtain a Bangla output sentence from an input English sentence, the differences in the syntactic structures at the phrasal level of the two languages must be determined [1] using their CFGs. When an appropriate grammatical structure of the output sentence in the target language is obtained, the bi-lingual dictionary can be used to generate the output sentence.

3.1 Structural Differences and Similarities between English and Bangla Sentences

MT systems so far have been developed concerning English-Bangla machine translation [4,5,11,12] mostly deal with syntactic differences between these two languages for translation purposes. In an MT system, the analysis, transfer and translation algorithms detect and transform source language structures into target language structures identifying the structural differences [8].

In English, object appears after the occurrence of a verb [8] in the form S V O in simple sentences, whereas in Bangla, this becomes S O V. The definite and indefinite articles appear before nouns in English. In Bangla, indefinite articles precede nouns, but definite articles occur after nouns. Adjectives appear before nouns [8] both in English and Bangla as shown in the following examples.

English: I have a good book on data mining.

Bangla: আমার ডেটা মাইনিং এর উপর একটি ভাল বই আছে।

Adverbs appear after verbs in English, whereas they appear before verb in Bangla. The following examples demonstrate this fact.

English: She walks slowly.

Bangla: সে খুব আস্তে হাঁটে।

Prepositional phrase usually follows nouns [8] in English, which is the reverse of that used in Bangla as shown in the following examples.

English: This is the first world cup cricket tournament in Bangladesh.

Bangla: বাংলাদেশে এই প্রথম বিশ্বকাপ ক্রিকেট টুর্ণামেন্ট।

In English, auxiliary verb (e.g., am, is, are etc.) appears before the main verb. On the contrary, the auxiliary verb is removed in Bangla, and ক্রিয়া বিভক্তি [13] is added with the main verb instead.

Table 3: Typical Entries for English Verbs in a Sample Bi-lingual Dictionary

English Word	Root	POS Category	Auxiliary Verb	Tense	Affix	Person	Number	Bangla Meaning
read	read	MV	Null	Present Ind.	Null	1p	s	পড়ি
reads	read	MV	Null	Present Ind.	s	3p	s	পড়ে
reading	read	MV	am	Present Cont.	ing	1p	s	পড়িতেছি
reading	read	MV	were	Past Cont.	ing	3p	pl	পড়িতেছিল
drinking	drink	MV	is	Present Cont.	ing	3p	S	পান করিতেছে
…	…	…	…	…	…	…	…	…

Abbreviations**:** Null → empty, 1p → 1st person, s → Singular, pl → Plural.

The inflections of the Bangla verbs are caused by the various forms of tenses and persons used in Bangla sentences. The inflection of verbs in English is also caused by the various forms of tenses, numbers and persons used in the sentence [14]. Following are two examples to demonstrate these facts.

English: He is playing cricket.

Bangla: সে ক্রিকেট খেলিতেছে।

3.2 Generation of Syntactic Structure from English Sentence

The syntactic structure of an English sentence can be extracted by employing any of the two parsing methods: i) top-down parsing, and ii) bottom-up parsing [6]. Both of these methods apply CFG rules provided for the source language to the input sentence during parsing.

Bottom-up Parsing of English Sentence

In natural language processing (NLP), the major task in any MT system is to extract the phrase structure of the input sentence using bottom-up parsing [2,5,6]. The grammatical or syntactic structure is discovered by rewriting the input sentence using CFG rules employing leftmost derivation. In the proposed model, the bottom-up parsing is achieved in two phases. In the derivation phase, each of the words available in the token list of the input sentence is replaced with the variable appearing on the left side of the appropriate CFG rule during parsing if the word is found in the terminal list. An *abstraction phase* is applied to the generated grammatical structure to obtain the highest level of abstraction.

During abstraction, a single CFG variable is replaced with the same or its more higher level category, and a group of CFG variables is replaced with a more higher level variable as defined by the CFG rules until a sentence S is obtained [11,12]. The bottom-up parsing of the English sentence "*The boy is drinking tea*" is shown in Fig. 2.

Input:	The boy is drinking tea
→	DEF-ART boy is drinking tea
→	DEF-ART N is drinking tea
→	DEF-ART N AV drinking tea
→	DEF-ART N AV MV tea
→	DEF-ART N AV MV N

(a) Phase 1: Structure Derivation

Input:	The boy is drinking tea
Structure:	DEF-ART N AV MV N
→	CN AV MV N
→	NP AV MV N
→	NP AV MV NP
→	NP AV CV
→	NP VP
→	S

(b) Phase 2: Abstraction Phase

Fig. 2. Bottom-up Parsing of *The boy is drinking tea*.

The parsing proceeds by replacing each component word of the input sentence with its syntactic category (POS) from left to right resulting in an intermediate structure [11,12] as shown in Fig. 2(a). A parse tree can be constructed from the abstraction steps of Fig. 2(b) by rearranging the steps in reverse order from bottom to top as shown in the left parse tree of Fig. 1. In this parse tree,

the root S stands for the English sentence where each leaf contains a component word of the input sentence. The bottom-up parser will try to parse incorrect sentences as well but the parsing process will not produce S.

Top-Down Parsing of English Sentence

In top-down parsing [5,6,11,12], the parser starts with the highest abstract level of the CFG rules by assuming a sentence S, and tries to establish whether the desired sentence is derivable or not. The parser derives the required syntactic structure consisting of most specific CFG variables (i.e. POS) for the desired English sentence by applying the appropriate CFG rules using left most derivation. These variables are then replaced with the appropriate words from the terminal list of the CFG rules to generate the input English sentence. If all of the words are available in the terminal list, the parsing terminates successfully with the input sentence, otherwise the derivation results in an intermediate string. A parse tree can be constructed from the parsing steps from top to bottom in the left to right direction. The top-down derivation may be very expensive in the worst case computation where the derivation has to try every possible branch of computation that may result from the alternative CFG clauses.

3.3 Transformation of the Syntactic Structure of English Sentences

The extracted syntactic structure of an English sentence is transferred to a syntactic structure of an equivalent Bangla sentence. A set of mapping rules [5] in terms of syntactic structures at the POS, clause and phrase levels can be used for English to Bangla machine translation. A tree-to-tree transformation [1] can be applied recursively to a syntactic structure of an English sentence to obtain the corresponding Bangla syntactic structure by mapping surface structures of sentences. In the proposed method, the grammatical structure of the input English sentence is discovered using bottom-up parsing, and the generated syntactic structure is then abstracted to the highest level of a sentence S by rewriting the derived syntactic structure using the given English CFG rules. We obtain the top-down parsing equivalent of an English sentence in a deterministic way [12] by reorganizing the abstraction steps of the bottom-up parsing in reverse order from bottom to top. An English parse tree can be constructed from these parsing steps. This English parse tree can be transformed into an equivalent Bangla parse tree by applying equivalent Bangla CFG rules to this tree at the phrase level. The proposed bottom-up parsing method is a great improvement over the original top-down parsing in terms of computation steps. Fig. 3 shows an intermediate parse tree, which can be generated during the parsing of the sentence *The boy is drinking tea*.

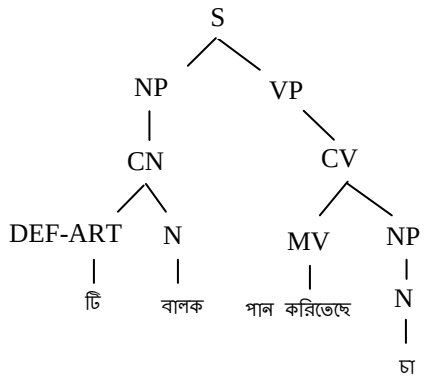

Fig. 3. Intermediate Tree Generated durir Translation of *The boy is drinking tea*.

In this sentence, the word '*is*' is of AV category, which is removed from this intermediate tree generated to construct an equivalent Bangla sentence, and the inflected English verb *drinking* is replaced with the Bangla verb পান করিতেছে. This intermediate tree is further processed to obtain the correct Bangla output sentence as shown in Fig. 1.

4. Proposed MT System

Development of a complete English to Bangla MT system [4,5,11,12] is really a challenging task. The proposed MT system is simple, well defined and modular in its architecture.

4.1 Architecture of the Proposed MT System

The architecture for English to Bangla MT system is shown in Fig. 4.

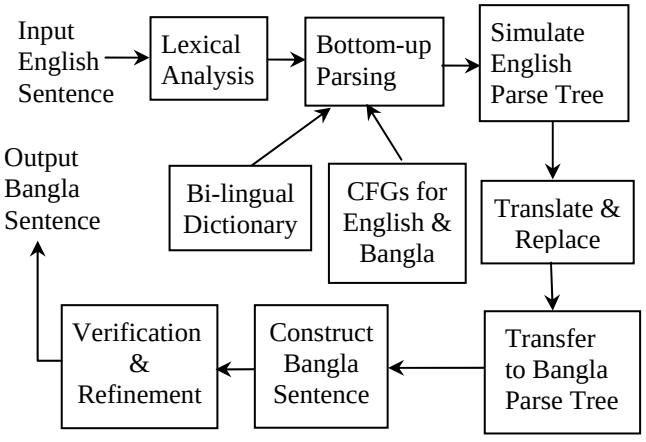

Fig. 4. Architecture for English to Bangla MT System [11,12].

The English sentence is input to the lexical analysis module, which converts this sentence into a list of tokens by reading individual group of character(s). The bi-lingual dictionary provides the morphological features and contextual information along with the corresponding Bangla meaning of each English word. The CFGs provide English and Bangla grammar rules at the phrase and sentence levels for machine translation purpose. The *Simulate English Parse Tree* module functions to simulate the computation history of parsing steps as in the form of a parse tree by reorganizing the structures obtained in the abstraction phase of bottom-up parsing in reverse order rather than constructing an actual parse tree. The *translate and replace* module replaces each English word with the corresponding Bangla meaning based on the tense, person and morphological properties appearing in the English sentence using the bi-lingual dictionary in the intermediate parse tree. The transfer process is then applied to this tree to obtain the equivalent Bangla parse tree by performing the changes required to the intermediate tree using the Bangla CFG rule for its English counterpart. The output Bangla sentence is constructed from the Bangla parse tree, which is verified and manually refined if required to obtain a more accurate translation.

4.2 English to Bangla Machine Translation Algorithm

The proposed MT system transforms an input English sentence into an equivalent Bangla sentence by using Algorithm 1 shown in Fig. 5. Considering auxiliary verbs as a special case, the MT system rewrites an English syntactic structure into an equivalent Bangla syntactic structure comparing equivalent phrases and syntactic components of both CFGs.

Algorithm 1: *TranslateEtoB*(S_E)
// S_E is the input English sentence
1: Tokenize the input English sentence S_E.
2: Parse the input sentence using bottom-up parsing technique.
3: Extract syntactic structure from the derivation phase of bottom-up parsing.
4: Apply abstraction phase on the extracted syntactic structure obtained in step 3.
5: Simulate English parse tree reorganizing the abstraction steps obtained in step 4 in reverse order.
6: Replace English words with their equivalent Bangla meanings in the simulated intermediate English parse tree using the bi-lingual dictionary.
7: Transfer intermediate English parse tree to equivalent Bangla parse tree by using equivalent Bangla CFG rule for each English CFG rule as required.
8: Extract the output Bangla sentence from the Bangla parse tree.
9: Verify the correctness of the grammatical structure of the output Bangla sentence, and refine the structure and meaning if necessary.

Fig. 5. English to Bangla Machine Translation Algorithm.

4.3 Sentence Classification Algorithm

The MT algorithm defined in Fig. 5 can be extended to deal with different types of English sentences using Algorithm 2 defined in Fig. 6. This algorithm can determine whether the input sentence is assertive, interrogative or imperative simple sentence [2,9]. The sentence classification algorithm searches for a fixed CFG component pattern within the parsed structure of the English sentence to determine the sentence type, which helps to guide the MT system to perform correct machine translation. This algorithm may incorrectly detect an interrogative sentence as an imperative sentence, and vice versa as both of these types of sentences may start with *do*. So, extra checking must be ensured to eliminate this deficiency. The sentence classification algorithm is a basic sentence *classifier*, which can be extended to deal with more complex sentences.

Algorithm 2: *ClassifySentence*(S_E)
// S_E is an input English sentence

1: If the first lexical entry is a NP followed by AV MV, or CONJ
2: then S_{type} = *Assertive*
3: Else
4: If the first lexical entry is an INTG-PN or AV
5: then S_{type} = *Interrogative*
6: Else if the first lexical entry is MV followed by NP or PP
7: then S_{type} = *Imperative*
8: Else
10: Ignore the input

Fig. 6. Sentence Classification Algorithm.

5. Implementation of the Proposed MT System

We have implemented a prototype of the proposed MT system using NetBeans IDE (Java) and *Bswing* classes [15] for English to Bangla character mapping. In this implementation, the parse tree concept has been simulated using an array data structure. We are improving the prototype following the algorithms and concepts presented in this paper. Current version can translate simple English assertive sentences in various tense forms. A very good

quality bi-lingual dictionary is designed and used in this prototype. Fig. 7 shows a snapshot of the prototype while translating the sentence *The healthy boys are playing football*.

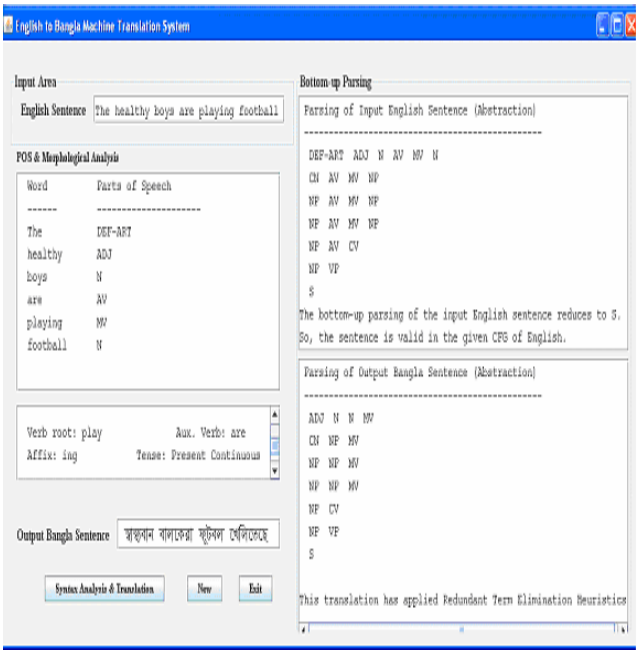

Fig. 7. Sample Output of the Prototype MT System.

6. Results

The prototype MT system has been tested with some English sentences as shown in Table 4. The corresponding parsed syntactic structures of both English and Bangla sentences using their CFGs are also given. In Table 4, the seventh English sentence contains a definite article (i.e., *The*) before the plural noun *boys*. The variable DEF-ART is removed from the parsed syntactic structure for the output Bangla sentence by the elimination rule option *redundant term elimination* using heuristics as *The* have no explicit contribution to the output Bangla meaning of the input English sentence.

7. Performance Analysis

The performance of an MT system largely depends on the size and quality of the MT dictionary and the CFG rules designed for the source and target languages. The proposed MT system is capable to translate English simple assertive sentence in various tense forms to its equivalent Bangla sentence providing all of the English words of the input sentence are known to the bi-lingual dictionary and the MT system.

Table 4: Some Examples of English to Bangla Translation

No.	Input and Output Sentences	Corresponding Syntactic Structures
1.	You were writing a letter তুমি একটি চিঠি লিখিতেছিলে	PER-PN AV MV INDEF-ART N PER-PN INDEF-ART N MV
2.	We are learning a new language আমরা একটি নতুন ভাষা শিখিতেছি	PER-PN AV MV INDEF-ART ADJ N PER-PN INDEF-ART ADJ N MV
3.	She was eating a mango সে একটি আম খাইতেছিল	PER-PN AV MV INDEF-ART N PER-PN INDEF-ART N MV
4.	He drinks tea সে চা পান করে	PER-PN MV N PER-PN N MV
5.	He will go home সে বাড়ী যাবে	PER-PN AV MV N PER-PN N MV
6.	I gave him a flower আমি তাকে একটি ফুল দিলাম	PER-PN MV PER-PN INDEF-ART N PER-PN PER-PN INDEF-ART N MV
7.	The healthy boys are playing football স্বাস্থ্যবান বালকেরা ফুটবল খেলিতেছে	DEF-ART ADJ N AV MV N ADJ N N MV

The developed MT system shows 100% translation accuracy for the sentences shown in Table 4, but it may fail to translate even a very simple sentence if it does not belong to the class of syntactic structures implemented in the prototype. The current version cannot deal with sentences containing multiple auxiliary verbs. If any proper noun [9] is not available in the bilingual dictionary, the MT system will generate a Bangla character sequence for the English characters of the noun. For example, the noun *Karim* will be replaced with কঅরইম using character mapping technique which is under implementation. This mapping of proper nouns into Bangla requires user interaction for refinement to obtain correct Bangla word. With sound and complete CFG rules for English and Bangla languages, the proposed MT system can achieve more translation power using the proposed MT dictionary storage structure. The termination of the bottom-up parsing technique resulting in the highest abstraction level S (i.e. sentence) ensures the correctness of the English and Bangla sentences as well as guarantees the soundness of the MT system. A lot of heuristic guidance must be incorporated in the MT system to maintain the translation accuracy in obtaining a correct Bangla sentence.

8. Termination

The proposed MT system rewrites a group of CFG variables into a higher level variable in the abstraction phase using the CFG rules. This clearly shows that the size *n* of the total number of terms occurring in the syntactic structure of a sentence is reduced during abstraction phase of bottom-up parsing to obtain the highest abstraction level S. This reduction ↓ in size is sufficient to ensure the successful termination of bottom-up parsing algorithm of any syntactically correct sentence in any of the source and target languages. More about the termination of transformation algorithms can be found in [16].

9. Conclusions

In this paper, we have presented a MT system for translating an English sentence to obtain a Bangla sentence of equivalent meaning. The proposed system uses sentence construction rules in the form of context-free grammars both for English and Bangla languages. We observed that the need for top-down parsing can be eliminated by rearranging the steps obtained in the abstraction phase of bottom-up parsing in reverse order, which results deterministic computation. This parsing method eliminates unnecessary computation which cannot be avoided in top-down parsing. The proposed system is more transparent as it directly uses the CFG rules for both English and Bangla languages. We have presented a complete architecture for English to Bangla MT system, and several algorithms for this system. More versatile English to Bangla MT dictionary has been designed and used to support efficient translation using contextual information. We are improving the prototype so that more difficult sentences can be translated. Research is going on how to handle sentences consisting of idioms and phrases, complex, compound and other forms of sentences in perfect and perfect continuous forms of tenses.

References

[1] M. M. Asaduzzaman and Muhammad Masroor Ali. *Transfer Machine Translation – An Experience with Bangla English Machine Translation System*. In the Proceedings of the International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 265-270, 2003.

[2] M. Reza Selim and M. Zafar Iqbal. *Syntax analysis of Phrases and Different Types of Sentences in Bangla.* In the Proceedings of International Conference on Computer and Information Technology (ICCIT 1999), Sylhet, Bangladesh, pp. 175-186, 1999.

[3] Mortuza Ali and Muhammad Masroor Ali. *Development of Machine Translation Dictionaries for Bangla Language.* In the Proceedings of International Conference on Computer and Information Technology (ICCIT 2002), Dhaka, Bangladesh, pp. 267-271, 2002.

[4] S. A. Rahman, K. S. Mahmud, B. Roy and K. M. A. Hasan. *English to Bengali Translation Using a New Natural Language Processing Algorithm*. In the Proceedings of International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 294-298, 2003.

[5] Sabbir Ahmed, Md. Obaidur Rahman, Saifur Rahman Pir, M. A Mottalib and Md. Saiful Islam. *A New Approach towards the Development of English to Bangla Machine Translation System.* In the Proceedings of International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 360-364, 2003.

[6] Mohammed Moshiul Hoque and Muhammad Masroor Ali. *A Parsing Methodology for Bangla Natural Language Sentences.* Proceedings of International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 277-282, 2003.

[7] S. K. Chakravarty, K. Hasan, A. Alim. *A Machine Translation (MT) Approach to Translate Bangla Complex Sentences into English*. Proceedings of International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 342-346, 2003.

[8] K. D. Islam, M. Billah, M. R. Hasan and M. M. Asaduzzaman. *Syntactic Transfer and Generation of Complex-Compound Sentences for Bangla-English Machine Translation.* Proceedings of International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 321-326, 2003.

[9] Md. Yusuf Ali Chowdhury and Md. Mofazzel Hossain. *Advanced Learner's Communicative English*, Paper I, Advanced Publications, Dhaka-1100.

[10] M. M. Asaduzzaman, Muhammad Masroor Ali. *Morphological Analysis of Bangla Words for Automatic Machine Translation*. In the Proceedings of the International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 271-276, 2003.

[11] Shibli Syeed Ashrafi. *Bangla Translation of English Simple Sentences Using MT System.* MS Project Report, Department of CSE, Jahangirnagar University, Savar, Dhaka-1342, 2009.

[12] A. K. M. Noman. *Bangla Translation of English Complex and Compound Sentences Using MT System.* MS Project Report, Department of CSE, Jahangirnagar University, Savar, Dhaka-1342, 2009.

[13] Lenin Mehedy, S. M. N. Arifin and M. Kaykobad. *Bangla Syntax Analysis: A Comprehensive Approach*. Proceedings of the International Conference on Computer and Information Technology (ICCIT 2003), Dhaka, Bangladesh, pp. 287-293, 2003.

[14] B. Hettige and A. S Karunananda. *Theoretical based Approach to English to Sinhala Machine Translation.* Fourth International Conference on Industrial and

Information Systems, ICIIS December 2009, Sri Lanka. IEEE XPlore.

[15] M. Shahriar. *bswing: a Unicode Based Bangla Java Package.*

[16] Md. Humayun Kabir. *Automatic Inductive Theorem Proving and Program Construction Methods Using Program Transformation*. PhD Thesis, School of Computing, Dublin City University, Ireland, September 2007.

Shibli Syeed Ashrafi graduated from the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. He is working at UCBL, Gulshan 2, Dhaka, Bangladesh. His research interests include Machine Translation, Database Systems and so on.

Md. Humayun Kabir graduated from Dhaka University, Bangladesh in Applied Physics and Electronics in 1991, and completed his Postgraduate degree in Computer Science in 1992 from the same University. He received his Ph.D from the School of Computing, Dublin City University, Ireland in 2007 in the area of formal software development. He is currently the Chairman of the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. He is also an Associate Professor in the same Department. His research interests include Program Transformation, Program Verification, Automatic Program Construction, OO Software Engineering, Machine Translation, Data Mining and so on.

Md. Musfique Anwar completed his B. Sc (Hons.) in Computer Science and Engineering from the Department of Computer Science and Engineering (CSE), Jahangirnagar University, Savar, Dhaka, Bangladesh in 2006. He is now a Lecturer (on study leave) in the Dept. of CSE, Jahangirnagar University. He is doing his MS in Japan. His research interests include Natural Language Processing, Artificial Intelligence, Image Processing, Pattern Recognition, Software Engineering and so on.

A. K. M. Noman graduated from the Department of Computer Science and Engineering, Jahangirnagar University, Savar, Dhaka, Bangladesh. He has completed his Postgraduate degree from the same Department. He is working at Purbani Group, Gulshan 1, Dhaka, Bangladesh. His research interests include Machine Translation, Database Systems and so on.