Bangla to English Machine Translation using Fuzzy Logic

Md. Musfique Anwar
Computer Science and Engineering Department, Jahangirnagar University, Bangladesh
Email: manwar@juniv.edu

***Abstract*- Transfer in machine translation (MT) plays an important role for producing correct output. This paper presents a technique to address about structural and lexical mappings from different types sentences of Bangla language for machine translation. Machine translation requires analysis, transfer and generation steps to produce target language output from a source language input. This paper deals with the syntactic transfer and generation for Bangla simple, complex and compound sentences into English. Structural representation of Bangla sentences encodes the information of Bangla sentences and a transfer module has been designed that can generate the English sentences from a corpus based automatic Bangla machine translator using Fuzzy logic. The effectiveness of this method has been justified over the demonstration of different Bangla sentences and the success rates in all cases are over 90%.**

Keywords- ***Machine Translation, Structural representation, Fuzzy Logic, Corpus***.

1. INTRODUCTION

Machine translation (MT) refers the translation from one natural (source) language to another (target language). It is an important area of Natural Language Processing (NLP). MT is a challenging job due to building up a successful translator for producing exact target language output from a source language. At a minimum, transfer systems require monolingual modules to analyze and generate sentences, and transfer modules to relate equivalent translation representations of those sentences.

To interpret language we need to determine a sentence structure. To do this we must know the rules of how language is organized and have an algorithm to analyze language given on those rules. Parsing serves in language to combine the meanings of words and phrases. A grammar captures the legal structure in a language and thus allows a sentence to be analyzed. Parsing a sentence then involves finding a possible legal structure for sentence. The result is usually a tree (referred to as parse tree) or structural representation (SR) [1].

Analysis and generation are two major phases of machine translation. There are two main techniques concerned in analysis phase. These are:

Morphological Analysis

Morphological analysis is the determination of the grammatical categories (noun, verb, adjective, adverb, etc) of the words of sentences. That means, it incorporates the rules by which the words are analyzed. To give an English example, the words analyzes, analyzed and analyzing might all be recognized as having the same stem analyze and the common endings –s, -ed, -ing. The result of morphological analysis then is a representation that consists of both the information provided by the dictionary and the information contributed by the affixes. Morphological information of words are stored together with syntactic and semantic information of the words.

Syntactic Analysis

Syntactic Analysis involves the inclusion of a few rearrangement rules in the basic word by word approach such as the inversion of ‘noun-adjective’ to ‘adjective-noun’. Rearrangement rules may take into account fairly long sequences of grammatical categories, but they do not imply any analysis of syntactic structure like the identification of a noun phrase. Complete syntactic analysis involves the identification of relationships among phrases and clauses within sentences. Syntactic analysis aims to identify three basic types of information about sentence structure:

1) The sequence of grammatical elements, e.g. sequences of word classes: article + verb + preposition ……., or of functional elements: subject + predicate. These are linear (or precedence) relations.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

2) The grouping of grammatical elements, e.g. nominal phrases consisting of nouns, articles, adjectives and other modifiers, prepositional phrases consisting of prepositions and nominal phrases etc. up-to the sentence level. These are constituency relations.

3) The recognition of dependency relations, e.g. the head noun determines the form of its dependent adjectives in inflected languages such as Bangla, German and Russian etc. These are hierarchical (or dominance) relations.

Actually each sentence is composed of one or more phrases. So if we can identify the syntactic constituents of sentences, it will be easier for us to obtain the structural representation of the sentence [2]. This paper implements a technique to perform syntactic analysis of Bangla sentences using grammatical rule-bases approach that accept almost all types of Bangla sentences. All rule-bases, namely inflection rule-bases, preposition mapping rule-bases and conjuncts mapping rule-base are designed to be declarative, rather than procedural which enables updating rule in a simple and easy manner [2].

A formal language is a set of *words*, i.e. finite strings of *letters*, *symbols*, or *tokens*. The set from which these letters are taken is called the *alphabet* over which the language is defined. A formal language is often defined by means of a formal grammar (also called its formation rules); accordingly, words that belong to a formal language are sometimes called *well-formed words* (or well-formed formulas). Natural languages such as Bangla, English, Chinese, have no strict definition but are used by a community of speakers [3].

2. PHRASES

Most grammar rule formalisms are based on the idea of phrase structure – that strings are composed of substrings called phrases, which come in different categories. For example, the phrases “the cow”, “the king”, “the agent in the corner”, are all examples of the category noun phrase or NP [3]. A sentence must have a subject phrase and a predicate phrase. The subject is the part, which names the person or thing we are speaking about. And the predicate is the part, which tells something about the subject.

Phrases form the building blocks for the syntactic structure of a sentence. In English, commonly used phrases are Noun phrase, Adjective phrase, Adverbial phrase and Prepositional phrase. Within the early standard transformational models it is assumed that basic phrase markers are generated by phrase structure rules (PS rules) of the following sort [4]:

S → NP AUX VP
NP → ART N
VP → V NP

Each rule is essentially a formula, or specification, called the production rules used of grammars by the parser to parse sentences. For example, the PS rules given above tell us that an S (sentence) can consist of, or can expanded as, the sequence NP (noun phrase) AUX (auxiliary verb) VP (verb phrase). The rules also indicate that NP can be expanded as ART N and that VP can be expressed as V NP.

There are three types of phrases in Bangla- Noun phrase, Adjective Phrase and Verb Phrase. Simple sentences are composed of these phrases. Complex and compound sentences are composed of simple sentences [5].

2.1 Analysis

Sentence can be analyzed into three main parts: (i) Syntactic interpretation (or Parsing), (ii) Semantic interpretation and (iii) Pragmatic interpretation.

Parsing is the process of building a parse tree for an input string [2][8][9]. The interior nodes of the parse tree represent phrases and the leaf nodes represent words. Semantic interpretation is the process of extracting the meaning of a sentence as an expression in some language representation. In this analysis phase, certain checks are made to ensure that the discrete input components fit together meaningfully. Pragmatic interpretation takes into account the fact that the same words can have different meaning in different situations [2] [5].

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

3. BANGLA SENTENCES STRUCTURE

A simple sentence is formed by an independent clause or principal clause. Example: বালকটি চা পান করে
Now, a simple sentence can have 1) Subject (উদ্দেশ্য) and 2) Predicate (বিধেয়) parts. Again subject are of two types a. Simple Subject (সরল উদ্দেশ্য), b. Expanded Subject (সম্প্রসারিত উদ্দেশ্য). And predicate are also of two types: a. Simple Predicate (সরল বিধেয়), b. Expanded Predicate (সম্প্রসারিত বিধেয়).

In Bangla language, complex sentence consists of one or more subordinate clause within a principle clause [2]. As for example, *ami Jakhon Dhaka gelam takhon se asustha chilo* (আমি যখন ঢাকা গেলাম তখন সে অসুস্থ ছিল)

From the above examples, we see that the prepositions are placed in free order but fixed order in English. Here in the first example, the first preposition in the preposition pair is used after noun whereas the one is used before the pronoun. In the second example, both the prepositions are placed before the pronouns respectively. The following prepositions (অব্যয়) are generally used to connect the principle clause with the subordinate clause. After analyzing the Bangla complex sentences, the following syntactic structure can be established: S → Conj* + DC + Conj + IC, where, S-sentence, Conjconjunctive word, DC-dependent clause, IC-independent clause.

When two or more independent clauses are connected by preposition (অব্যয়) and thus construct a single sentence, then the sentence is said to be compound sentence [3]. As for example, *se kal asbe ebong ami Jabo* (সে কাল আসবে এবং আমি যাব)

The following prepositions (অব্যয়) are generally used to connect the independent clauses: *o* (ও), *ebong* (এবং), *noile* (নইলে), *fale* (ফলে) etc. The syntactic structure for compound sentence can be established as:

sentence → clause + preposition (অব্যয়) + sentence|clause

3.1 Structural Transfer

Structural transfer for compound-complex sentences has two levels. Firstly, we have to take various clauses with respect to conjunctive word if available. Secondly we perform structural transformation of every clause. Every conjunctive word has expectations in terms of clauses and an associated rule for structural transfer. Conjunctive word mapping rule-base contains expectations and structural transfer rule for every conjunctive word [6]. Example includes, "and" expects two clauses and structural transfer rule for "and" sentences would be clause1 + and (এবং) + clause2 [2].

It is seen that there are two simple sentences in the primitive complex sentence. One is the principle simple sentence and the other is the subordinate simple sentence. To translate the complex Bangla sentence, it is necessary to separate these two simple sentences. To do this, first a given complex sentence is scanned and then searches to know which type of subordinate and/or subordinate complement is in the sentence. For example, a complex sentence "*jadi tumi paro tahole pas koriba* (যদি তুমি পড় তাহলে তুমি পাস করিবে)". Here,

Principal simple sentence: *tumi pas koriba* (তুমি পাস করিবে)
Subordinate simple sentence: *tumi paro* (তুমি পড়)
Subordinator: *jadi* (যদি)
Subordinator complement: *tahole* (তাহলে)

In complex sentences, above subordinators are usually come before the subordinate simple sentence and subordinator complements are added in the various positions in the principal simple sentence [4].

4. BANGLA STRUCTURE ANALYSIS

4.1 Proposed Model

The model proposed model for structural analysis of Bangla sentences is shown in **Fig. 1**.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

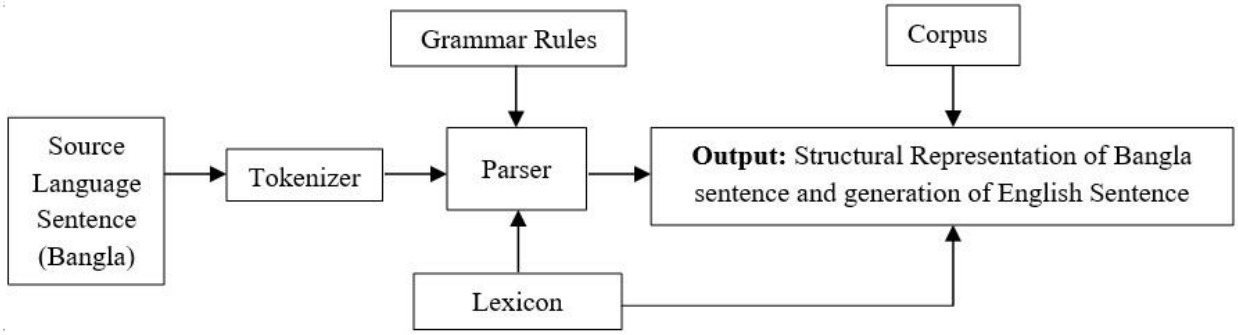

Fig. 1 Block diagram of Bangla parser

4.2 Description of the Proposed Model

For parsing, we take a Bangla natural sentence as input. In next phase the stream of characters are sequentially scanned and grouped into tokens according to lexicon. The words having a collective meaning are grouped together in a lexicon. The output of the Tokenizer of the input sentence **"**নির্বাচন হবে এবং গণতন্ত্র প্রতিষ্ঠিত হবে**"** is as follows [1] [5]: Token = ("নির্বাচন", "হবে", "এবং", "গণতন্ত্র", "প্রতিষ্ঠিত", "হবে").

The parser is the most important tool of this phase. To ensure its validity within the underlying grammar, every sentence must be checked by the parser. The parser involves grouping of tokens into grammatical phrases that are later used to synthesize the output. Usually, the phrases are represented by a parse tree that depicts the syntactic structure of the input. The most common way to represent grammar is as a set of production rules which says how the parts of speech can put together to make grammatical, or "well- formed" sentences.

Lexicon is a list of allowable words. The words are grouped into the categories or parts of speech. A lexicon can also be defined as a dictionary of words where each word contains some syntactic, semantic, and possibly some pragmatic information. This is the largest components of an MT system in terms of the amount of information they hold. The information in the lexicon is needed to help determine the function and meanings of the words in a sentence. Each entry in a lexicon will contain a root word called head. The entries in a lexicon could be grouped and given by word category (by specifier, nouns, verbs, and so on), and all words contained within the lexicon listed within the categories to which they belong [1] [3] [4] [5]. **Fig. 2** illustrates a sample lexicon for Bangla parsing.

A corpus is a large and structured set of texts. They are used to do statistical analysis and hypothesis testing, checking occurrences or validating linguistic rules on a specific universe. This is the basic training corpus used to train the alignment template Language Model.

```
Noun → শিক্ষা education | জাতি nation | মেরুদন্ড backbone | ভাত rice | মাঠ field | ফুটবল
football | রহিম rahim |
Pronoun → তুমি you | সে he | তোমার you |
Inflection → ই is the | এর of the | র of the | কে to | টি the | এ to |
Finite verb → হয় happen | যায় goes | যায় go | খায় eats | কর do | খেলে plays |
Infinite verb → করা to do | বলা to say |
Indeclinable → না no |
Adverb → পরাজিত loser | কোথায় where |
Conjuction → কিন্তু but | এবং and |
Additive word → যেখানে where | সেখানেই there | যাহা what is | তাহা that |
Interogative → কী what |
Adjective →সুন্দর beautiful | মনোরম wonderful | সহজ easy | ভাল good |
Preposition → জন্য for |
```

Fig. 2 Sample lexicon for Bangla parsing.

Structural Representation is a process of finding a parse tree for a given input string. That is, a call to the parsing function PARSE, such as PARSE ("The dog is dead") should return a parse tree with root S whose leaves are "The dog is dead" and whose internal nodes are non- terminal symbols [3]. In linear text, we write the tree as:
[S : [NP : [Article : **the**] [Noun : **dog**]] [VP : [Verb : **is**] [Adjective : **dead**]]] **Fig. 3** shows the parse tree for this sentence.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

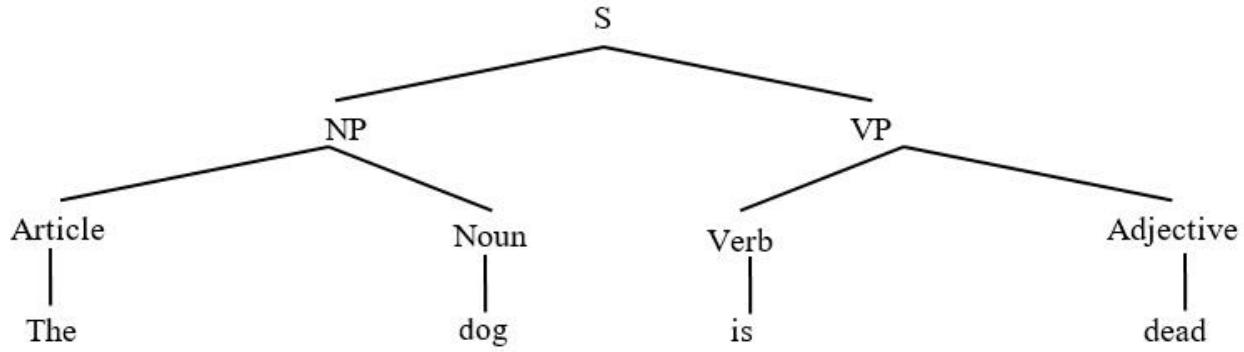

Fig. 3 Parse tree for the sentence "The dog is dead"

For conversion, we applied Fuzzy logic for the interpretation of the input Bangla sentence to English sentence as output; where fuzzy membership values are calculated with the help of probability distribution of the words in corpus [7]. This paper focuses on the formation and use of the grammar rules to be used by the parser in the syntax analysis phase. Structurally, there are three types of Bangla sentences:

a. Simple Sentence (সরল বাক্য),
b. Complex Sentence (জটিল বাক্য),
c. Compound Sentence (যৌগিক বাক্য).

4.3 Basic Rules to Parse a Sentence*

1. Sentence → Simple sentence | Complex sentence | Compound sentence; 2. Simple sentence → Principle clause; 3. Complex sentence → Subordinate part + Additive word + Principal clause | Principal clause + Additive word + Subordinate part; 4. Subordinate part → Subordinate clause | Subordinate clause + Additive word + Subordinate part; 5. Subordinate clause → Additive word + Principal clause; 6. Additive word → Indeclinable | Null; 7. Compound sentence → Principal clause + Additive word + Compound part; 8. Compound part → Principal clause | Compound sentence; 9. Principal clause → Subject + Predicate; 10. Subject → Simple subject | Expanded subject; 11. Predicate → Simple predicate |
Expanded predicate; 12. Simple Subject → Actor (কর্তৃপদ); 13. Actor → Noun + Inflection | Pronoun + Inflection | Implicit (উহ্য) Actor; 14. Pronoun → Person ; 15. Person → FP | SP | TP ; (Example -- FP - aami , aamraa) 16. SP → SPH | SPNH | SPP ; (Example -- SPH- aapni , aapnaara ; SPNH - tumi , tomraa ; SPP - tui , toraa) 17. TP → TPH | TPNH ; (Example -- TPH - tini , taaraa ; TPNH - shey , taaraa) 18. Implicit Actor → Null; 19. Expanded Subject → Sub-expander + Subject; 20. Sub-expander → Adjective | Adjective + Infinite verb | Adjective clause | Relative part (সম্বন্ধ পদ/পদসমষ্টি) | Relative part + Adjective | Adverbial clause; 21. Relative part → Noun + এর (er) | Pronoun + এর | Adjective + এর , 22. Simple predicate → Verb clause | Implicit verb; 23. Implicit verb → Null; 24. Expanded predicate → Pre-expander + Verb clause; 25. Pre-expander → Adverb | Adverb + Adverb | Adverb + Object (কর্মপদ) | Adjective + Object | Adjective expander (বিশেষণের বিশেষণ) + Adjective + Object | Object | Adverbial clause; 26. Object → Noun | Pronoun | Relative part + Noun | Relative part + Pronoun | Null; 27. Verb clause → Infinite verb + Finite verb | Finite verb | Implicit verb | Infinite verb + Finite verb + Indeclinable | Finite verb + Indeclinable(অব্যয় পদ); 28. Indeclinable → না (na) | Other;

* The '→' sign means the phrase "can have the form of", the '|' sign indicates an alternative rule for the left-side term and the '+' sign means join of two terms of a sentence.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

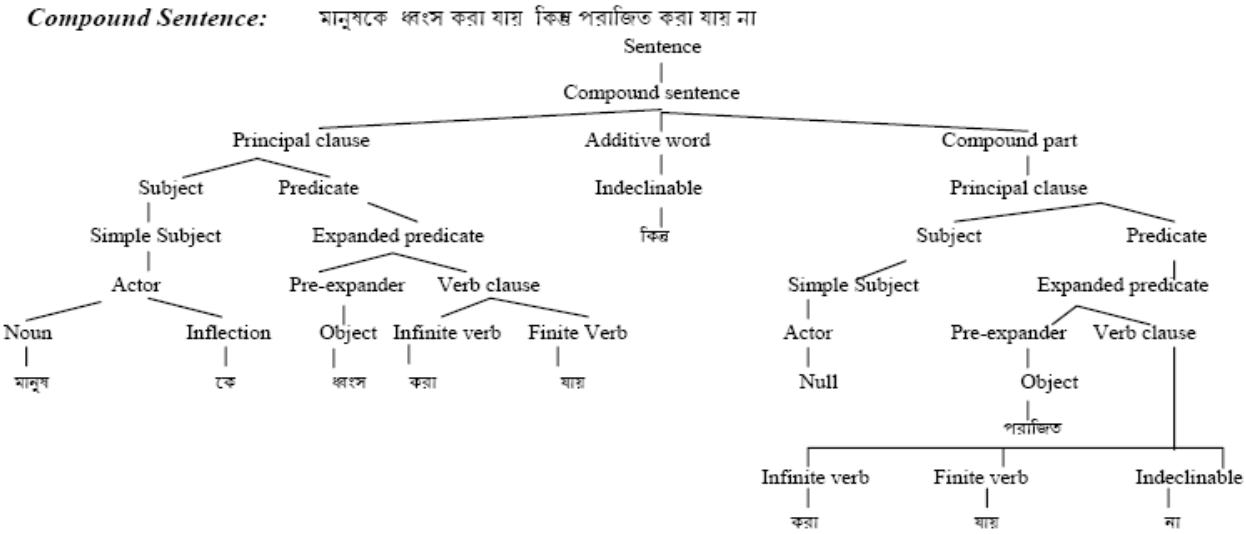

Fig. 4 Parse tree for a Compound Bangla sentence

5. IMPLEMENTATION OF PROPOSED MODEL

5.1 Generation of Structural Representation

Flow-chart of structural representation (SR) generation by means of parsing approach are given bellow:

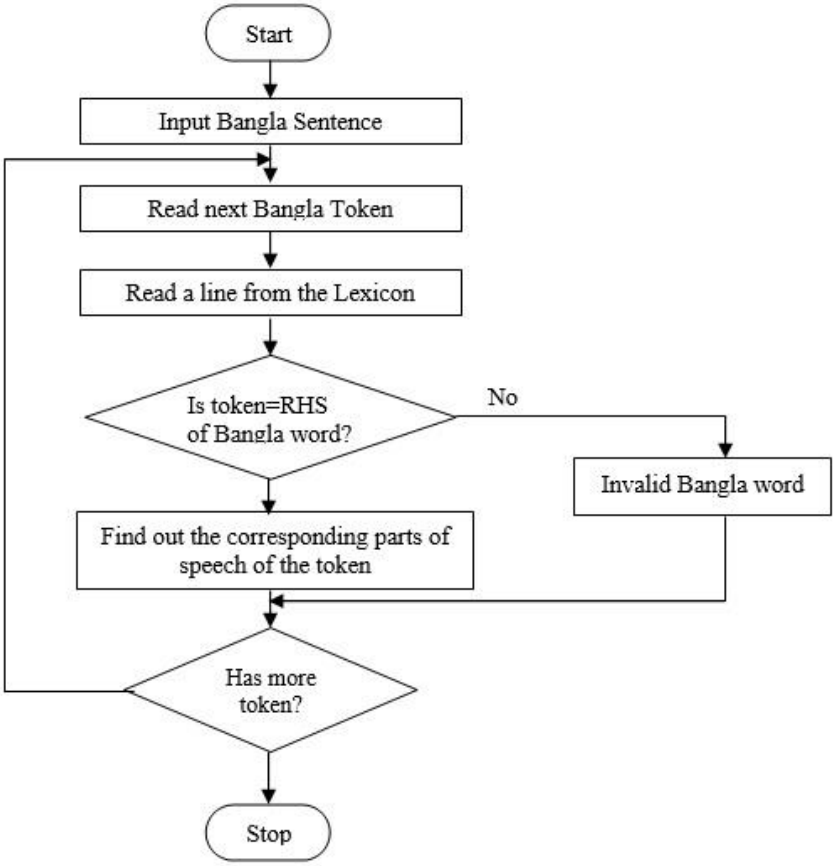

Fig. 5 Flow-chart for finding the respective parts of speech of Bangla token

5.2 Implementation of Language Modeling

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

To implement the language model, a bilingual corpus of large amount of aligned sentence pairs is used as training corpora. This corpus contains an English sentence and a Bangla sentence for each aligned pair. The translation model uses both the Bangla and English sentences to estimate the translation probability which is considered as fuzzy membership value of each Bangla word. Two steps for this purpose are defined as,

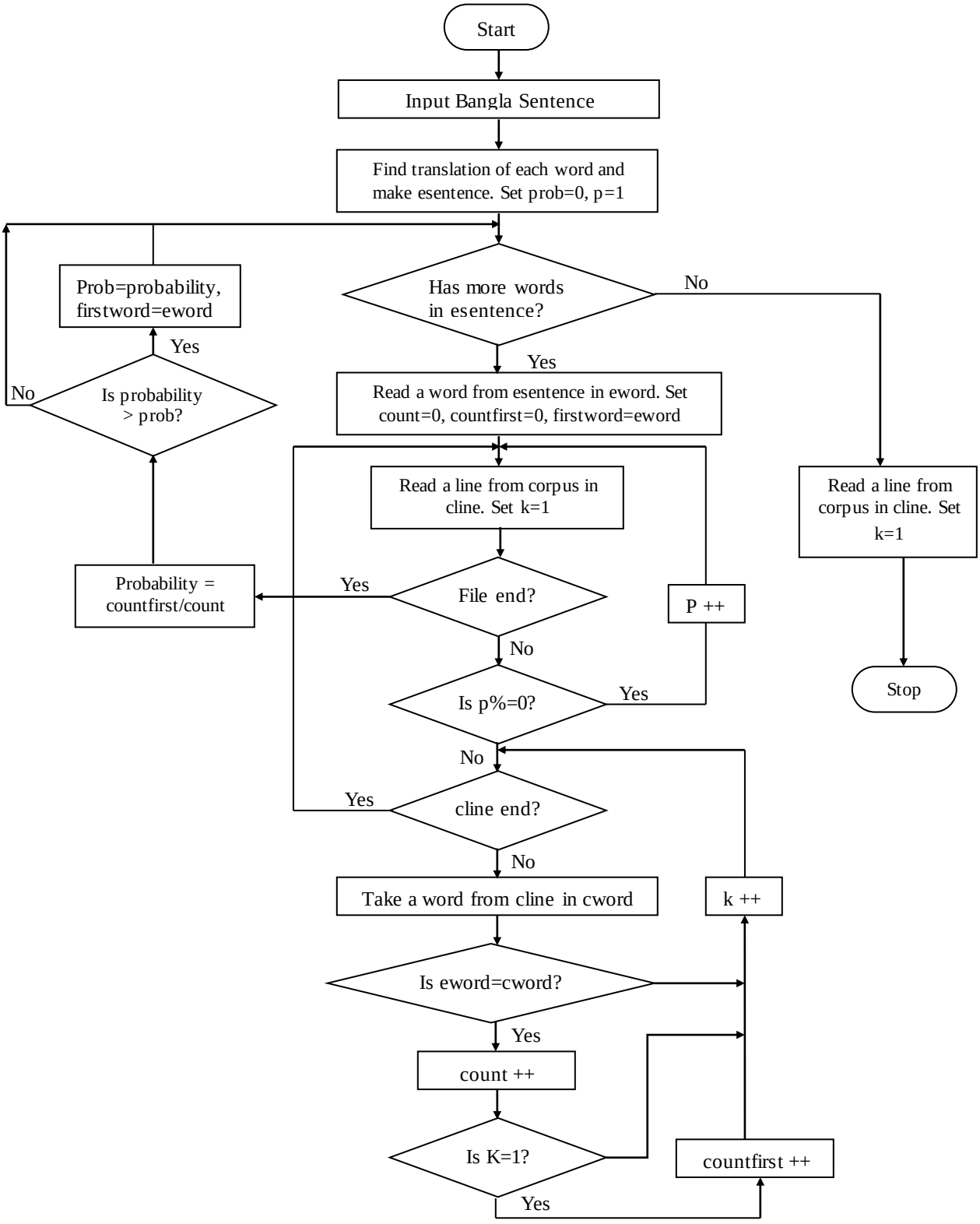

Fig. 6 Flow-chart for calculating the First word of a sentence

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

Calculating the fuzzy membership value of each word to come first of the English sentence

To calculate this probability the number of occurrence of each word is calculated as "count" and the number of occurrence of that word in the first position of an English sentence is also calculated as "countfirst" from the bilingual corpus. Then "countfirst" is simply divided by "count" to get the particular probability of that word. And we assign this probability as fuzzy membership value for that word. Flow-chart for calculating the First word of a sentence is shown in **Fig. 6**.

Calculating the fuzzy membership value of each other words to come next of the English sentence
For calculating the probabilities of each other words to come next following the current word, a combination is formed with each other words. Then the number of occurrence of this combination is calculated as "countcombination" and the number of occurrence of the current word is also calculated as "countindividual". Now "countcombination" is simply divided by "countindividual" to get the probability of every word. And this probability is assigned as fuzzy membership value for the combination. If no words is found the current word is retained.

6. EXPERIMENTAL RESULTS AND DISCUSSIONS

In order to justify the effectiveness of this method, several experiments were conducted. **Fig. 7** illustrates the snapshot of the implemented method. Success rate for different types of sentences is shown in **Fig. 8**.

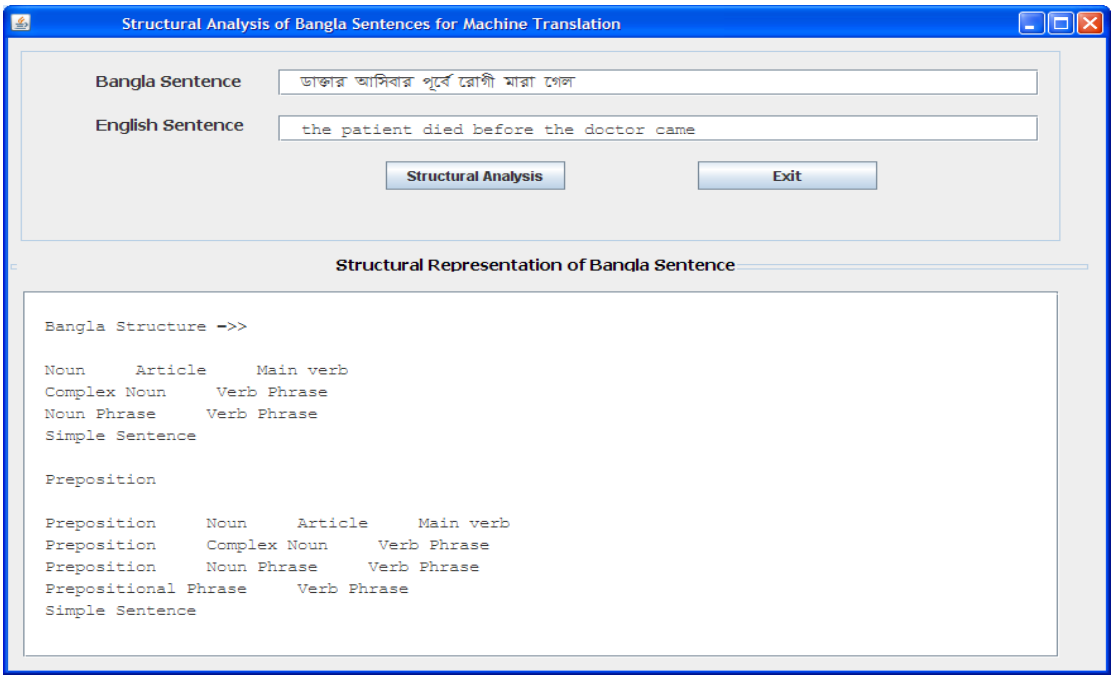

Fig. 7 Sample output of the program for the complex sentence

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

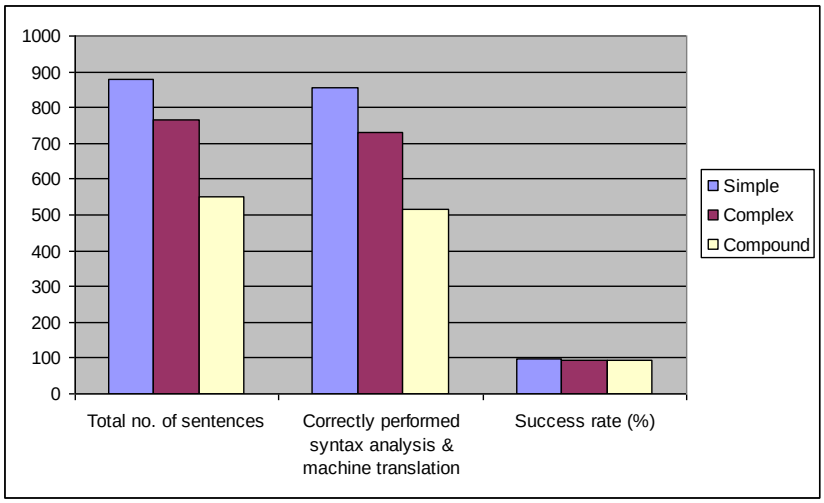

Fig. 8 Success rate for different types of sentences

7. CONCLUSION

This paper mainly focuses on the syntax analysis phase, it sets aside the job of extracting additional information about the other parts of speech such as the noun, the pronoun, the adjective, the adverb and the indeclinable which will be of great use in the semantic analysis phase. The concepts of the change of voice, narration and other special concepts of Bangla grammar such as the composition of words (সমাস) and inflection of the noun or pronoun (নাম-বিভক্তি) should be further analyzed. As the major emphasis was given to parse the finite verb of the sentence, other types of clauses like the adverbial or adjective clauses were not parsed further.

In this paper, we have discussed how to identify the principal simple sentence and the subordinate simple sentence in a Bangla complex sentence and to separate the primitive complex sentence. We have also shown how to generate the translated English complex sentences.

We know that, Bangla grammar has an inherent property in forming the verbs, that is, unlike the English grammar, various necessary information of a sentence such as the tense, the person, the mode of verb (ক্রিয়ার ভাব) etc. can be extracted from a finite verb. Many previous works did so by decomposing the verb phrase. The inflection of verb (ক্রিয়া-বিভক্তি) plays a very important role in this regard; further investigation can be done in decomposing the verb and then extracting the information. Few earlier works have proposed parsing method for different forms of Bangla present tense. We can extend those set in future by proposing methods for all other types. The inflection of Bangla verb (ক্রিয়া-বিভক্তি) can have different forms depending on the tense, the person and the class of subject of the verb.

REFERENCES

[1] M. M. Hoque and M. M. Ali, "A Parsing Methodology for Bangla Natural Language Sentences", Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 277-282 (2003).
[2] K. D. Islam, M. Billah, 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), Dhaka, Bangladesh, pp. 321-326 (2003).
[3] S. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 2nd Edition, Pearson Education publisher, New York, 2003.
[4] 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), Dhaka, Bangladesh, pp. 342-346 (2003).
[5] L. Mehedy, N. Arifin and M. Kaykobad, "Bangla Syntax Analysis: A Comprehensive Approach", Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 287-293 (2003).
[6] D. Rao, P. Bhattacharya and R. Mamidi, "Natural Language Generation for English to Hindi Human-Aided Machine Translation", Proceedings of International Conference on Knowledge Based Computer Systems, (Mumbai, India), pp. 171-189 (1998).
[7] M. G. Uddin, M. Murshed, M. A. Hasan, "A parametric approach to Bangla to English Statistical Machine Translation for complex Bangla sentences -Step 1", Proceedings of International Conference on Computer and Information Technology (ICCIT), Dhaka, Bangladesh, pp. 529-534 (2005).
[8] M. M. Anwar, M. Z. Anwar, M. A. Bhuiyan, "Syntax Analysis and Machine Translation of Bangla Sentences", IJCSNS International Journal of Computer Science and Network Security, VOL.9 No.8, August 2009, pp. 317-326.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500

[9] S. S. Ashrafi, M. H. Kabir, M. M. Anwar, A. K. M. Noman, "English to Bangla Machine Translation System Using Context-Free Grammars", IJCSI International Journal of Computer Science Issues, Vol. 10, Issue 3, No 2, May 2013, pp. 144-153.

https://sites.google.com/site/ijcsis/
ISSN 1947-5500