A NEW APPROACH TO DEVELOP AN ENGLISH TO BANGLA MACHINE TRANSLATION SYSTEM

Md. Golam Rabiul Alam, Md. Monirul Islam and Nowrin Islam

Department of Computer Science and Engineering
International Islamic University Chittagong, Bangladesh.

E-mail: gra9710@yahoo.com, monirliton@yahoo.com, nowrin_islam@yahoo.com

Abstract: *Machine translation (MT) is always a challenging job. It is really difficult to build up a complete machine translation system for natural languages. Machine translation includes natural language understanding and generation. The proposed system represents a new solution for building a MT system for English to Bangla translation, by modifying the rule-based transfer approach of MT system. In machine translation the searching of word from the lexicon is a compulsory task, here this searching stage is utilized efficiently by proposing an intelligent integer based lexicon system, consists of a number of separate lexicons and an algorithm is also developed for searching words from the lexicon in order to accomplish the basic steps of machine translation.*

Keywords: *English to Bangla, Intelligent, Lexicon, Machine Translation, Parsing and Semantic.*

1. Introduction

In any MT system lexicons bears a lot of importance. As searching the lexicons is an compulsory task, so by utilizing that compulsory task in an effective way, an efficient MT system is tried to be developed .In this paper, based on the idea of rule-based transfer approach [1], we have modified the architecture and try to develop a new MT system. Only simple affirmative sentences are considered here. But by proper modification same system can be used for other sentence patterns of English language.

2. Literature Review

Lexicon is the list of stems and affixes, together with basic information about them. Every lexicon is of a certain class . Lexicon searching is a one of the important task for some Machine Translation (MT) approaches to extract corresponding meaning of a word in the target language and also the syntactic and semantic information about the word. Grammatical rules are also kept in the lexicons. So in a MT system, lexicon is a central component. The existing MT rule based [2] or structure based [3] systems use the lexicons for only the above-mentioned purposes. We utilize the lexicon searching in an effective way to make the whole translation process an efficient one. Our proposed lexicons have the capability of making decisions by their own of where to search for the next valid word and to provide error checking and error message while searching is taken place and at the same time generation can also be done. Here no need of transfer step because the corresponding Bangla words are mapped to their desired position in the time of searching the words in the lexicon, by extracting all the information about the desired position of that word in the target text. The existing MT systems use the separate parser (bottom up or top down) for checking the syntactic and semantic validity of the sentence. And then by transferring the appropriate Bangla meaning for the English word, from the lexicons, the target sentence is generated. So firstly they accomplish searching in the lexicon for all the words in the sentence for extract corresponding lexical, syntactic and semantic information about the word, and also grammar rules are searched in the lexicon for the purpose of parsing. But our system does not use any separate parser and no grammar rules are kept in the lexicon. Rather the lexicons are organized in such a way that they seems to be the leaf node of a parse tree .So lexicons together can act as a parser .The lexicons store semantic information which used for semantic analysis. This system introduce integer based lexicon, so word matching can be done through simple arithmetic operation like subtraction .The binary search can also be used here in order to search out the beginning index of words which begin with same letter. And this system can solve the translation of some ambiguous simple sentence. Our proposed MT system is a modification of the transfer-based MT system. Basically in this approach translation proceeds in three stages, analyzing

Date of submission : 05.01.2010 Date of acceptance : 10.10.2010

input sentences in to a representation which still retains characteristics of the original source language text. This is then input to transfer component, which produces a representation, which has characteristics of the target language and from which a target sentence can be produced [1]. So, transfer-based MT system follows three steps such as analysis, transfer and generation. And our system has only two steps --analysis and generation. So it is comparatively a faster technique. In a transfer-based machine translation system transfer step involves changing the underlying representation of the source text into an underlying representation of the target text. Synthesis or generation step and final major step involves changing the underlying target text representation into the target text, using target language grammar. For example “I eat rice” –this source text will be” ami khai vat “in the transfer step. Then by applying Bangla grammar rule of sentence structure subject-object-verb we get “ami vat khai. “, in the generation step. But in our proposed system the intermediate transfer step can be omitted. As the valid position of a word in the target sentence is given in the lexicon, so the target sentence can be generated using the information about the proper position of the word in the target sentence, according to the grammatical rule of the target sentence. There is no need to involve an underlying representation of the target text. As the proposed system has two steps, it reduces translation time.

3. Proposed Machine Translation System

The translation process accomplished in two steps: analysis and generation. Analysis step involves lexical analysis and encoding of the token by taking of ASCII code for each character, syntactic and semantic analysis accomplished by searching in the lexicon. The generation step is also done while searching is taken place in the lexicon for a word of the input text. Here the translation is accomplished with help of an intelligent, integer based bilingual lexicon system, which not only stores information but also can make decisions and can check errors in the input text.

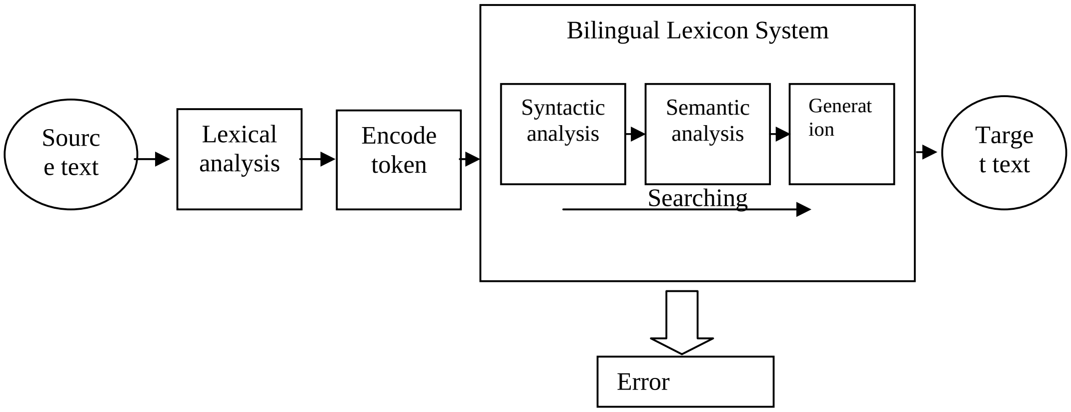

Fig 1: Proposed MT system architecture.

The lexicons will also contain semantic information, which will provide semantic analysis during searching. Not only that, while searching, if the word is found in the valid phase, the lexicon will provide the information where to transfer the corresponding Bangla word validly in order to generate a correct Bangla simple sentence which will be the translated sentence of the given input simple English sentence. That is the lexicons also do generation. The lexicons are numbered to accomplish the translation work properly.

3.1 Intelligent, Integer number based Lexicon System

Our proposed lexicon system is based on two things:

i) Comparative Grammatical structure of English and Bangla

ii) Individual constituent Lexicons

By comparing the grammar rules of both English and Bangla simple sentence, a comparative grammatical structure of both the languages have been followed and basis on this, we arrange our lexicons by following the

English grammar rule. For generation, we follow the Bangla grammar rule. Searching is a challenging factor to make a MT system efficient. Our Intelligent searching technique is done by strategically searching in grammatically arranged individual constituent lexicons and performs the analysis, transfer and synthesis in searching stage by using valid searching phases.

3.2 Lexical Analysis

Lexical analysis involves scanning the input sentence from left to right and tokenize them into individual words. Then each of the words is encoded before the lexicon searching is taken place and the numeric codes are stored in the lexicons. To encode each word we follow the encoding algorithm discussed later.

3.3 Syntactic Analysis

In our system syntactic analysis is done in two ways

1. Constituent structural Analysis: The representation of a parse tree is called constituent structure. [1]. It will analysis if the all compulsory constituents in the sentence are present in valid sequence.
2. Morphological analysis: Which will analysis if the morph or word is of right pattern as we know that in English grammar, the verb changes its structure according to person, number and aspect of tense.

In syntactic analysis, constituent structure analysis and morphological analysis is involved. Searching does constituent structure analysis. If searching in any lexicon contains compulsory constituents, fails, or any violation of valid sequence of finding the word in the lexicons occurs, it will generate an error. For morphological analysis, we extract the person and number information about the subject and auxiliary verb and the person, number and aspect of tense information about the auxiliary verb and the suffix of root verb. Because the auxiliary verb and suffix of verb change their structure according to the aspect of tense of verb, and number, person of subject. We, in our system used bottom-up parsing approach. And context-sensitive grammar and phrase structure rule is followed here. The following table shows us the grammar rules used for both Bangla and English.

The phrases can be shown by using a parse tree (Fig. 2 & Fig 3) for both English and Bangla simple sentence. The phrase structure for English and Bangla simple sentence is shown with the help of following parse tree where simple, affirmative English sentence “The good student is reading the tough essay from the new book” is used as an example. The two parse trees can be expressed in the form of the following table where each phrase along with their constituents are shown according to the valid sequence of their occurrence in both English and Bangla simple, affirmative sentence. The phase number is used here to identify the sequence of valid occurrence of each constituent in both English and Bangla sentence. Here from the above grammar rules, parse tree and structural analysis of both Bangla and English sentence, we can come to the point that generally all the sentence having a subject, an object, a finite verb and a prepositional phrase and also the subset of this sentence structure, are come one after another sequentially in the above way, if we scan the sentence from left to right. On the basis of that analysis, we organized the separate constituent lexicon one after another sequentially and searching of the tokens are taken place in the lexicon in the sequential order as each leaf node or terminal in the parse tree come one after another sequentially**.** Each lexicon behaves here as a leaf node of the parse tree as shown in the figures [Fig 2 and Fig 3]. Moreover, as the lexicons are strategically organized in a way by maintaining the valid phase of occurrence of the constituents of a sentence, an error can easily be detected if any compulsory constituent is absent in the input text. We keep information in the lexicon about the next valid constituent of the sentence. And as we numbered the lexicons, we keep the information in a lexicon about the next valid lexicon where to search next to get the next valid constituent and that give the lexicons decision-making capability. It makes the searching efficient and faster because each phase of searching tells where to search next to find the next valid constituent of the sentence .By detecting an invalid occurrence of any constituents, error checking can be done instantly other than checking for all the constituents of the sentence in the lexicon and after all these, finding that the sentence is wrong. This intelligent lexicon system help us to get rid from searching in one lexicon in unorganized way which is time consuming and inefficient. From the Fig 4, we see how the lexicons are organized one after another in a strategically way, here the upper arrows are used to indicate if a word is found in the lexicon, in which lexicon to go next. If no word is found in a lexicon, the next lexicon has to be searched.

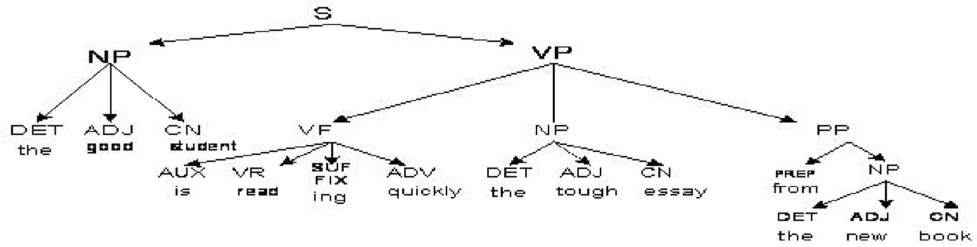

Fig 2: Bottom-up parse tree for simple affirmative English sentence

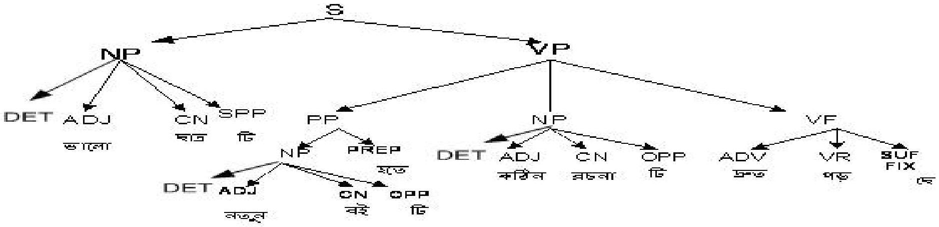

Fig 3: Bottom-up parse tree for translated simple affirmative Bangla sentence

Table 1: The English and Bangla grammar rules of proposed system

English sentence Rules	**Bangla sentence rules**
S= NP+ VP	S= NP+VP
NP= PRN/PRO/DET+ADJ+CN	NP= PRN/PRO/DET+ADJ+CN+SPP/OPP
VP=VF+NP+PP	VP= PP+NP+VF
PP=PREP+NP	PP= NP+PREP
VF=AUX+VR+SUFFIX+ADV	VF=ADV+VR+SUFFIX

Table 2: English simple sentence structure according to bottom-up parse tree.

Sentence															
Subject					**Verb**				**Object**						
NP					VP										
					VF				NP			PP			
												Prep	NP		
PRN	PRO	DET	ADJ	CN	AUX	VR	SUFFIX	ADV	DET/PRO/OP	ADJ	CN	PREP	DET/PRO/OP	ADJ	CN
1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14

Table 3: Bangla simple sentence structure according to bottom-up parse tree

Sentence																	
Subject						**Object**									**Verb**		
NP						VP											
						PP					NP				VF		
						NP				Prep							
PRO	PRN	DET	ADJ	CN	SPP	DET/PRO/OP	ADJ	CN	OPP	PREP	DET/PRO/OP	ADJ	CN	OPP	ADV	VR	SUFFIX
1	1	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16

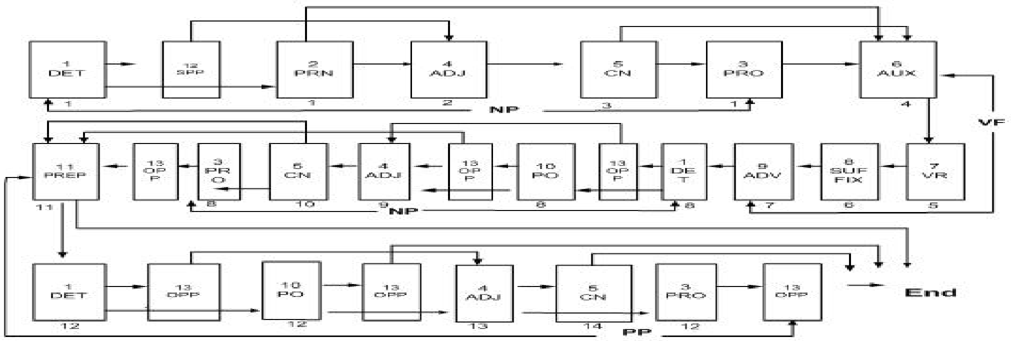

Fig 4 : Syntactic analysis with the help of proposed lexicon system

[Abbreviation: PRN = Pronoun, PRO = Proper Noun, DET = Determiner (a/an/the), Quantifier (one, two, three…), ADJ = Adjective, CN = Common Noun, PO= Pronoun (object form), SPP = Subject Post Position, OPP = Object Post Position, AUX = Auxiliary Verb, SUFFIX = Verb Suffix, ADV = Adverb, PREP = Preposition, VR=Verb root, VP = Verb Phrase, NP = Noun Phrase, PP = Prepositional Phrase, VF = Verb Form]

3.4 Semantic Analysis

Semantic analysis is done to ensure that whether the discrete input constituents fit together meaningfully [2]. We know in Bangla and English language there are several mismatches in semantic meaning though have correct syntactic structure.

3.5 Categorization of subject, object and verb

The validity of the meaning of a sentence depends on the relation among the subject, verb and object in a sentence. Some English sentences can be syntactically valid but not semantically valid. In order to accomplish semantic analysis and to detect error, we categorize the subject, verb and object of a sentence by tagging them using integer numbers. Basically the verb and object of a sentence depends upon the subject because the subject is the performer of a verb and the subject perform a verb with an object or on an object or by an object. So here the subject is categorize according what it can do and by whose help it can perform the task. Here the subject is categorized mainly human and non-human –in these two different types. Human type is again divided into two parts –Honorific and Non-Honorific. Non-Human type is divided into two parts —Material and Non-material. The tagging and mapping of subject with verb and object is stated in table 4&5and examples are given in Table 6. The constituents with the same semantic type can be used in the sentence together validly. Subtracting the tag number of both verb and object from the subject can perform the checking. If the result is zero in both cases, the sentence is semantically valid. If one result of subtraction is non-zero, the sentence will be semantically wrong.

Table 4: Tagging of type of subject, verb and object.

Subject		Verb		Object	
Tag No.	Type	Tag No.	Type	Tag No.	Type
1	Honorific	1.	Honorific	1.	Honorific
		2.	Non-Honorific	2.	Non-Honorific
		3.	0	3.	Material
		4.	Non- Material	4.	Non- Material
2	Non-Honorific	1.	Honorific	1.	Honorific
		2.	Non-Honorific	2.	Non-Honorific
		3.	0	3.	Material
		4.	Non- Material	4.	Non- Material
3	Material	1.	0	1.	0
		2.	0	2.	0
		3.	Material	3.	0
		4.	0	4.	0
4	Non-Material	1.	0	1.	0
		2.	Non-Honorific	2.	Non-Honorific
		3.	0	3.	Material
		4.	Non- Material	4.	Non- Material

Table 5: Mapping of the type of subject with the type of verb and object.

		Subject	Object	Verb
Human Being	1	Honorific	H,NH,M,NM [1,2,3,4]	H,NH,NM[1,2,4]
	2	Non-Honorific	H,NH,M,NM [1,2,3,4]	H,NH,NM [1,2,4]
Not human Being	3	Material	0	M [3]
	4	Non-Material	M, NM [3,4]	NH,NM [2,4]

Table 6: Example of semantic type mapping according to above mapping.

No	Example	Category			
1	The bird writes an essay (4) (1) (1)	**Subject**	**Verb**	**Object**	(4-1)=3 (4-1)=3 Not Correct
		NM	H	H	
2	The student writes an essay (1) (1) (1)	H	H	H	(1-1)=0 (1-1)=0 Correct

3.6 Generation

In machine translation, generation means to generate the input sentence in target language. In our proposed MT system, generation is taken place when a word is found in the valid lexicon. Each lexicon keeps the information about the appropriate position of the corresponding Bangla meaning of the English words that they contain in the resulting text. Then the information about the appropriate position is extracted from the lexicon and the word is placed in the desired position .To find out the appropriate position of a word of English language in Bangla,the comparative analysis of the grammar of both language have to be done. A generalized grammatical structure for source and target language is made at first. From the relative comparison between the generalized form of English and Bangla affirmative, simple sentence with one subject, one finite verb, one object and a prepositional phrase, we can generate the following valid phase structure according to the occurrence of constituents for both languages. While searching the lexicon, if an English word is found, just replace the corresponding Bangla meaning of that word in the appropriate Bangla phase, act as the index of a output array, corresponding to the valid phase of that English word to accomplish generation in time of searching. So the comparative grammatical structure of both languages is needed.

Table 7: Comparative analysis of valid phase structure for English simple sentence and corresponding Bangla simple sentence.

English Phase No.	1	2	3	4	5	6	7	8	9	10	8	11	12	13	14	12
Corresponding Bangla Phase No.	1	2	3	4(SPP)	15	16	14	10	11	12	13(OPP)	9	5	6	7	8(OPP)

Table 8: An English and Bangla sentence translation according to comparative grammatical structure.

PRO	PRN	DET	ADJ	CN	AUX	VR	SUFFIX	ADV	DET	ADJ	CN	PREP	DET	ADJ	CN
		The	good	boy	is	Read	ing	Quickly	the	tough	Essay	from	The	new	book

PRO	PRN	ADJ	CN	SPP	DET	ADJ	CN	OPP	PREP	DET	ADJ	CN	OPP	ADV	VR	SUFFIX
		ভাল	ছেলে	টি		নতুন	বই	টি	হতে		কঠিন	রচনা	টি	তাড়াতাড়ি	পড়	ছে

4. Implementation

The proposed system has been implemented with Visual Basic 6.0 and Microsoft Access. Where Visual Basic 6.0 is used for front end or interface design and MS Access is used for Lexicon design.

5. Time Complexity

In our proposed MT system, we have designed a lexicon of English Grammar which is used for semantic analysis and to generate target language words. As each phases of the lexicon is designed by using integer number so for searching of words from dictionary requires average complexity of $\Theta(\log n)$ as words in the dictionary follows the topological order. The words are grouped into CN,DET, VR, ADJ etc. in the proposed lexicon so, I is not required the search the whole dictionary to find a particular group word. As our proposed system doesn't requires intermediate representation so it reduces the conversion complexity of the existing systems. So if we can define English grammar in our proposed lexicon structure then we can parse and translate it by polynomial time order. So whereas in Normal existing architecture requires one exponential search and one pre-order search to find Bengali Parse tree [6], our architecture requires only one polynomial order search, which proves that our proposed architecture is better.

6. Conclusion

Our system has some limitations. The proposed system does not capable of translating all type of English sentence. It does not include any non-finite verbs like Gerund, participle and more than one subject and object. There is no complete MT system yet of Bangla to English translation, so it will requires proactive research on this field.. This proposed system can be helpful for those who are doing researches on the development of MT system. And hopefully it will create a new direction in the field of English to Bangla machine translation.

References

[1] Md. Abdullah-Al- Mamun, Muhammad Iftekhar Ahmed, Mohammed Alauddin Bhuiyan, Mohammad Reza Selim and Muhammed Zafar Iqbal "*An Implementation of Machine Tanslation Between Bengla and English* " in proceedings of International Conference on Computer & Information Technology (ICCIT), 2000, NSU, Dhaka, Bangladesh.

[2] Lenin mehedy, S.M Niaz Arifin and M Kaykobad "*Bangla Syntax Analysis: A Comprehensive Approach*" in proceedings of International Conference on Computer & Information Technology (ICCIT), 2003, Jahangirnagar university, Dhaka, Bangladesh.

[3] Stuart Russell and Peter Norvig, Artificial Intelligence-A Modern Approach, Prentice Hall ,2003, pp-818-821.

[4] Khandokar Jahirul Alam, Md. Arafatur Rahman, A. K. M. Meherul Haque, Touhidul Islam, Mohammad Tanvir Irfan, "Structure Based Bangla To English Machine Translation", International Conference on Computer & Information Technology (ICCIT), 2005, IUT, Gazipur, Bangladesh.

[5] Shehab Raihan, Muhammad Masroor Ali, "Bangla to English Translation by Rule-Based Approach", International Conference on Computer & Information Technology (ICCIT), 2004, Brac University, Dhaka, Bangladesh.

[6] Sajib Dasgupta Abu Wasif Sharmin Azam, "An Optimal Way of Machine Translation from English to Bengali", International Conference on Computer & Information Technology (ICCIT), 2004, Brac University, Dhaka, Bangladesh.