Int. Conf. on Computer and Information Technology (ICCIT2000)
NSU, Dhaka, 25-26 January, 2001

An Implementation of Machine Translation Between Bangla and English

Md. Abdullah-Al-Mumin, Mohammad Iftekhar Ahmed, Mohammed Alauddin Bhuiyan,
Mohammad Reza Selim and Muhammed Zafar Iqbal
Dept. of Electronics and Computer Science,
Shahjalal University of Science and Technology, Sylhet, Bangladesh.
E-mail: selim@sust.edu; ranju-ecs@sust.edu

Abstract: This paper describes an implementation process of a Machine Translation (MT) system between Bangla and English. Transfer based Linguistic Knowledge (LK) architecture is preferred to implement this architecture. In this system we have used Lexical Functional Grammar (LFG) frameworks to define the Bangla and English grammar. A functional specification is specified to extract grammatical relations from the constituent structure of a Bangla sentence. We have implemented a number of dictionaries in this system and their structure is also discussed here. This system presently is capable to translate the affirmative sentences only.

1. Introduction

Machine Translation (MT) is the transfer of meaning from one natural language to another with the aid of a computer [1]. Machine Translation potentially can melt away the language barriers, It could even unite the world both intellectually and culturally. Although it is far from complete realization, practical translation systems are already available for grammatically constrained languages in restricted domains. Modern computational linguistic theories have been developed for many languages to aid their translation process. In this context Bangla as a natural language is still neglected as a research item in current MT field. In this paper we describe our work to implement a Bangla translation system. Since we communicate with the international knowledge world mainly through English, we've started with a Bangla to English translation system.

Grammar has an important role in the implementation of any MT system. There exists a well-defined grammatical model of English to use in computer applications, but the situation is quite different in Bangla. Fortunately, an appreciable preliminary work on Bangla grammatical model for computer use was done by Mohammad Reza Selim and Dr. Muhammed Zafar Iqbal [2] and we used their grammatical model to parse the Bangla sentence in our translation system.

2. Different Approaches for Machine Translation

Machine Translation (MT) system can be classified by their architecture i.e. the overall processing organization. Traditionally, MT has been based on **Transformer** architecture system. The newer one is **Linguistic Knowledge** (LK) architecture, which involves more extensive and sophisticated kinds of linguistic knowledge. The LK architecture includes two approaches. The first is so-called **interlingual** approach. The interlingual approach would seem preferable for the multilingual translation system, but since our system is designed to interact between two languages, Bangla and English, we have chosen the second approach called **transfer** approach. In this approach translation proceeds in three stages, analyzing input sentences into 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 [3].

3. Linguistic Functional Grammar

In general, syntax of a sentence for both Bangla and English can be represented with two different structures. The first is constituent or phrase structure (c-structure), which is used to parse a sentence using phrase structure rules. The second is functional structure (f-structure), which works with grammatical relations such as subject, object and verb. Each language can be defined with the grammatical functions but they have the different phrase structure rules. For this reason f-structure is used to transfer the information between the languages. Lexical Functional Grammar (LFG) [4] is a strong computational formalism that addresses how to extract grammatical relations from a sentence in a positional language such as Bangla and English. LFG formalism has two major components, a context free grammar and a functional specification. The c-structure is produced using phrase structure rule defined by the context free grammar. The f-structure is produced by using functional specification together with the c-structure. The functional specification specified in this system discussed in the parsing section later.

4. The System Architecture

Here we present the design concepts of the translation system we have developed. The system has been developed on transfer based LK architecture (Fig 1), which has three parts: analysis, transfer and synthesis. Analysis step involves using the parser and the Bangla grammar to analyze the Bangla input sentences. Transfer step involves changing the underlying representation of the Bangla sentence into an underlying representation of an English sentence. Synthesis step and final major step involves changing the underlying English representation into an English sentence, using a generator and the English grammar.

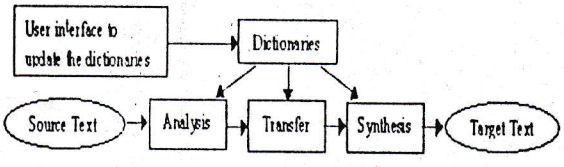

Fig 1: Transfer based translation system architecture.

ISBN: 984-584-005-1

4.1 Analysis

Analysis process takes an input sentence, gather lexical and morphological information and then construct a c-structure of the sentence. Finally, it produces f-structure from the c-structure to provide as input of the transfer system. It is performed through two steps: (a) Lexical analysis or scanning and (b) Syntactic analysis or parsing. Fig 2. shows the overall architecture of an analysis process.

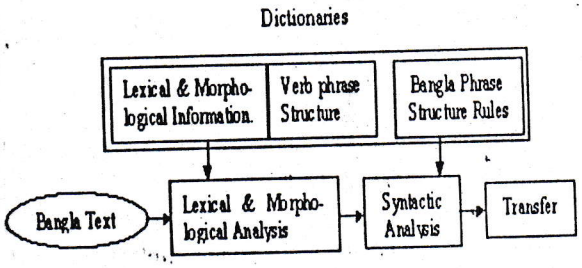

Fig 2: Architecture of an analysis process.

4.1.1 Lexical Analysis

Lexical analysis involves reading and extracting words from the input sentence. After being extracted, the word is looked up into the dictionary database called lexicon. The lexicon contains the necessary information corresponding to a word. The information may include type of word (e.g. বই(boi) = NP) and morphology of word. Morphology is concerned with the internal structure of words (e.g.পড়ছিল (porchilo) = পড় (por)+ছিল(chilo)) and how words can be formed (e.g. ভাল(bhalo) করে (kore) = ভাল করে (bhalo kore)).

Word grouping

Looking up into the dictionary, some words, if necessary, are made combined into groups. Because two or more words may be combined to represent a single word type. For example, in the sentence "*bhalo kore lekhapora koro* (ভাল করে লেখাপড়া কর)" the word combination "ভাল করে (bhalo kore)" represents an adjective. Similarly the word combination "লেখাপড়া কর (lekhapora koro)" represents a single Verb Form.

Morphological analysis

After grouping has been done, if possible, words are decomposed into components, There are a number of inflectional suffixes denoting number (singular or plural) of the nouns and pronouns in a sentence. For example, the suffix রা(ra) of the word ছেলেরা(chhelera) indicates the plural number of the noun ছেলে(chhele). There are well-defined suffixes to represent cases (কারক) of the nouns. These suffixes perform the job of English prepositions. Bangla words has a very strong and structured inflectional morphology for its verb forms for different tenses and persons (e.g. করেন(koren)= কর (kor)+এন (en)) . Using these suffixation, tense and person of a sentence can be detected almost unambiguously. We can get the tense, aspect and person of a sentence from the suffix of the verb form using Table 1(Appendix A) [2]. For example, the suffix ছিল(chhilo) of the word পড়ছিল (porchhilo) retrieves information from Table 1. that it is Past-Continuous-TPNH. If the scanner fails to find a word in the Lexicon database, then it will generate an error message. For convenience a separate Lexicon handler is also needed that allows addition of new words into the dictionary, correction of any previous incorrect entry and showing dictionary items according to Lexical word class. For each token, decomposed or grouped together, the scanner retrieves their type from the lexicon. Thus the scanner provides a surface structure of the input sentence for the parser. For example, the surface structure of the sentence " মেয়েটি একটি বই পড়ছিল (mayeti ekti boi porchhilo)" will have the form shown as the constituent structure:

TPNH	PP	Qntfr	PP	TPNH	VR	Continuous	Past	Con-TPNH
মেয়ে	টি	এক	টি	বই	পড়	ϕ	ϕ	ছিল
(maye)	(ti)	(ek)	(ti)	(boi)	(por)			(chhilo)

[TPNH: Third Person Non-Honorific; PP: Post Position; Qntfr: Quantifier; VR: Verb Root]

4.1.2 Syntactic Analysis

The task of an automatic parser is to take a formal grammar and the surface structure provided by the scanner and apply the grammar to the sentence in order to (a) check that it is indeed grammatical and (b) given that it is grammatical, represent the sentence in a structure which can be used to transfer into the target language structure.

Verifying the sentence

Now taking the surface structure of the sentence the parser verifies whether the sentence is grammatical or not by producing a parse tree using the grammar. The procedure used here to parse a sentence is 'bottom-up' parsing, which starts working from the leaves up to the root. If the sentence can be parsed up to the root using the grammar rules we can conclude that the sentence is grammatical. The Bangla grammar rules used by this system are stored in a dictionary that is used during producing the parse tree. The representation of a parse tree also called constituent structure (c-structure). The c-structure of the sentence " মেয়েটি একটি বই পড়ছিল (mayeti ekti boi porchhilo)"is shown in Fig.3.

Functional specifications of LFG Formalism

The functional specifications of LFG may consists some rules which is used to produce the f-structure (Fig. 5(a)) from the c-structure (Fig. 3). The functional specifications used in this system are as follows:

1. Traverse down the c-structure tree. Beginning at the root node, the left sub-tree, rooted at NP, contains the information of the subject of the sentence.

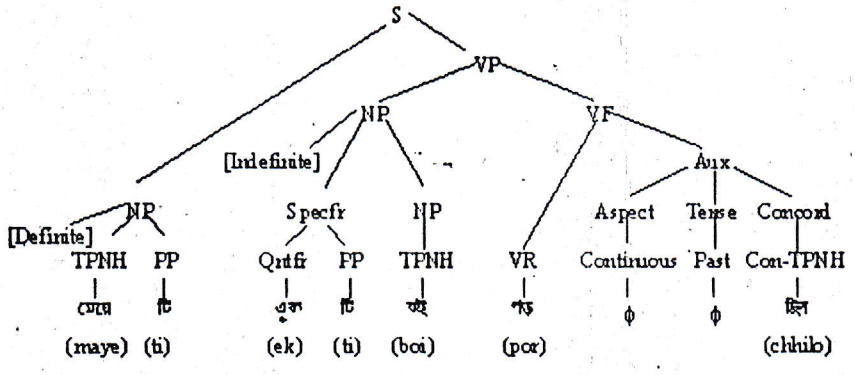

Fig. 3: The constituent structure of the sentence "মেয়েটি একটি বই পড়ছিল (mayeti ekti boi porchhilo)".

[NP: Noun Phrase; Specfr: Specifier; Aux: Auxiliary; VF: Verb Form; VP: Verb Phrase]

2. Now pointed at the right child of the root (VP), the left sub-tree of the VP, rooted at NP, contains the information of the object of the sentence.
3. Now pointed at the left child of the VP (VF), the left child of the VF, VR, is the verb root of the sentence. The right sub-tree of the VF, rooted at Aux, contains the Aspect, Tense and Person of the sentence.

4.2 Comparative Grammar and Transfer

Now we focus on the transfer component, which embodies the comparative grammar that links the analysis and synthesis components together – the module in the center of a translation system. We can investigate at how the comparative grammar relates a representation for Bangla sentence to the corresponding representations for English sentence. The comparative grammar has bilingual dictionary rules. In the simplest case, these may just relate source lexical items to target lexical items (Fig. 4):

মেয়ে - girl এক - one বই - book পড় - read
ছেলে - boy দুই - two ভাত - rice খেল - play

Fig. 4: Bilingual dictionary

These dictionary rules can be seen as relating leaves on the source language tree to leaves on the target language tree. The comparative grammar also contains some structural rules which relate other parts and nodes of the two functional structure to each other. One such structural rule might be read as: "The translation of the whole sentence is normally made up of the translation of the subject + the translation of the object + the translation of the verb".

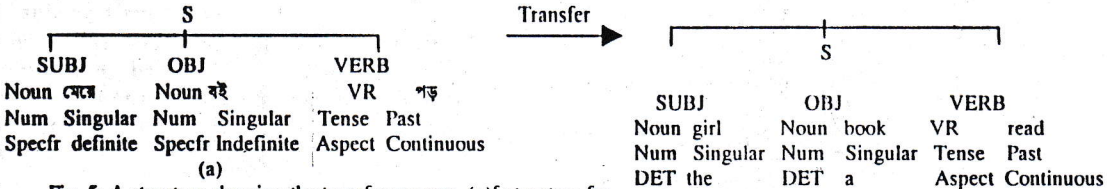

Fig. 5: A structure showing the transfer process. (a)f-structure for

In this context 'translation' has the restricted sense of translation into the corresponding target language representation – this representation has to be input to synthesis before a 'full' translation is reached. In our translation system the auxiliary information are simply carried over from source structure to target structure. Applying these rules to the Bangla representation will result into the construction of the corresponding English representation. Fig. 5 shows this transfer process. This representation serves as input for the English synthesis module, which applies the rules of the English grammar to produce an English sentence.

4.3 Synthesis

The synthesis process works in reverse way as the parser in the works of analysis process works. The synthesis process takes the f-structure provided by the transfer process. This f-structure has the grammatical information – subject, object and verb – of the target language, English, as shown in Fig. 5(b). Now the task of the synthesis process involves mapping the f-structure into the c-structure and generating the sentence from the c-structure, using the English grammar (Fig. 6).

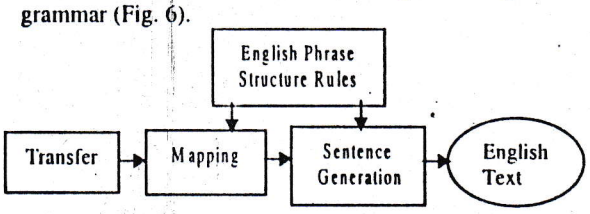

Fig. 6: Architecture of a synthesis process.

4.3.1 Mapping

We have implemented here the approach to map the grammatical information from the f-structure into some partial c-structures using the English grammar rules. These partial c-structures are then used to

construct a complete c-structure using the rules. These grammar rules are also stored in a dictionary as was done for the Bangla grammar rules. The English sentence thus produced must be grammatical.

In this mapping process the grammar rules are applied until a c-structure is produced that match each part of the f-structure – subject, object and verb. If the structure produced by a particular rule matches the input f-structure, then a partial c-structure (Fig. 7) is

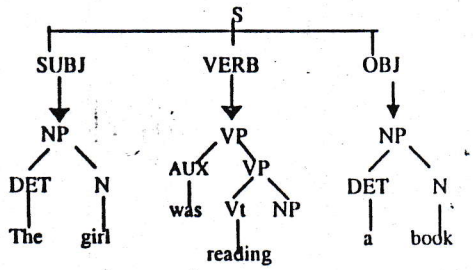

Fig. 7: Mapping of the c-structure from English f-structure of Fig. 5(b).

built with this rule. For mapping the verb information into c-structure it requires the *auxiliary form* and *correct form* of the verb. Here we have designed two dictionaries, one of which contains the auxiliary value and the form the way the verb root will be changed, other contains various root verbs and their modified values corresponding to a different form. Table 2(Appendix A) and Table 3(Appendix A) shows the format of the two dictionaries. Now we can obtain the auxiliary value and correct form of the verb root using the 'Tense' and 'Aspect' information and looking up into the dictionaries mentioned above.

4.3.2 Sentence Generation

Now all that is necessary is attaching the partial c-structures that have been constructed, in the appropriate places using the grammar rules. Thus the complete constituent structure we obtain is shown in Fig 8.

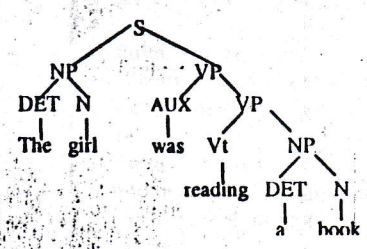

Fig 8: Complete c-structure by attaching the partial c-structure using English grammar rules.

Thus the leaf nodes of the complete c-structure represents the word order of the translated target sentence. So the translation system we have developed gives the output " The girl was reading a book "for an input Bangla sentence "মেয়েটি একটি বই পড়ছিল (maeyeti ekti boi porchhilo)".

5. Dictionaries

In an MT system dictionaries are probably the central component. The size and quality of the dictionary limits the scope and coverage of a system, and the quality of translation. So it is very important to implement an efficient dictionary for the system. To analyze a sentence in an MT system different types of word information is required. It may include word-type, morphology of the word (internal structure of words and how words can be formed) etc. The system also requires grammar rules for both the source and the target languages, which must be provided in the dictionaries. The important dictionary we need at the center of an MT system is the bilingual dictionary, which is used to translate information between source and target languages. To make a system really useful the system should allow the end user to make some additions to the system dictionaries.

6. Conclusion

The MT system we have designed can verify the grammatical validity of the input Bangla sentence because the sentence is analyzed at the front end with the Bangla grammar. Since a grammar for the English language is also provided at the synthesis portion, the translated output will tend to be grammatical. Presently our system works with affirmative sentences only, but it can be implemented for other types of sentences. The LK architecture, the approach we have implemented, is a reversible process [3]. So our implemented "Bangla to English" translation system can also be made "English to Bangla" by adding some relevant modules and dictionaries at both ends of the system.

7. References

1. *Computer Translation of Natural Language*, W.Goshawke, I.D.K. Kelly, J.D.Wigg
2. *Syntax analysis of Phrases and Different Types of Sentences in Bangla*, Md. Reza Selim and Md. Zafar Iqbal, Proceedings, International Conference on Computer and Information Technology, Sylhet, 1999.
3. *Machine Translation: An Introductory Guide*, Doug Arnold, Lorna Balkan, Siety Meijer
4. *Natural Language Processing: A Paninian Perspective*, Akshar Bharati, Vineet Chaitanya, Rajeev Sangal.
5. *Language Files* (Fifth Edition), Crabtree Monica, Powers Joyce, 19991, Ohio State University Press, Columbus.
6. *Compilers: Principles, Techniques and Tools*, Aho A.V., Shethi R., Ulman J.D., Bell Telephone Laboratories.

Appendix A

Table 1: Showing the verb auxiliary depends on Aspect, Tense and Person + Class of the subject.

Person & Class	Aspect	Verb Auxiliary			
		Present	Past	Future	Habituated Past
First Person	Indefinite	i (ই)	lam (লাম)	bo (ব)	tam (তাম)
	Continuous	chhi (ছি)	chhilam (ছিলাম)	Te-thakbo (তে থাকব)	te-thaktam (তে থাকতাম)
	Perfect	echhi (এছি)	echhilam (এছিলাম)	e-thakbo (এ থাকব)	e-thaktam (এ থাকতাম)
Second & Third Person Honorific	Indefinite	en (এন)	len (লেন)	*ben (বেন)*	ten (তেন)
	Continuous	chhen (ছেন)	chhilen (ছিলেন)	Te-thakben (তে থাকবেন)	te-thakten (তে থাকতেন)
	Perfect	echhen (এছেন)	echhilen (এছিলেন)	e-thakben (এ থাকবেন)	e-thakten (এ থাকতেন)
	Imperative	un (উন)	*not applicable*	*ben (বেন)*	*not applicable*
Second Person Non-Honorific	Indefinite	*a (অ)*	le (লে)	*be (বে)*	te (তে)
	Continuous	chho (ছ)	chhile (ছিলে)	*te-thakbe (তে থাকবে)*	te-thakte (তে থাকতে)
	Perfect	echho (এছ)	echhile (এছিলে)	*e-thakbe (এ থাকবে)*	e-thakte (এ থাকতে)
	Imperative	*a (অ)*	*not applicable*	o (ও)	*not applicable*
Second Person Pejorative	Indefinite	*is (ইস)*	li (লি)	bi (বি)	ti (তি)
	Continuous	chhis (ছিস)	chhili (ছিলি)	Te-thakbi (তে থাকবি)	te-thakti (তে থাকতি)
	Perfect	echhis (এছিস)	echhili (এছিলি)	e-thakbi (এ থাকবি)	e-thakti (এ থাকতি)
	Imperative	(No Auxliary)	*not applicable*	*is (ইস)*	*not applicable*
Third Person Non-Honorific	Indefinite	e (এ)	lo (ল)	*be (বে)*	to (ত)
	Continuous	chhe (ছে)	chhilo (ছিল)	*te-thakbe (তে থাকবে)*	te-thakto (তে থাকত)
	Perfect	echhe (এছে)	echhilo (এছিলি)	*e-thakbe (এ থাকবে)*	e-thakto (এ থাকত)
	Imperative	uk (উক)	*not applicable*	*be (বে)*	*not applicable*

Table 2: A dictionary format containing the Tense information.

Present			
Indefinite	VR	VR	TP_singular
Continuous	am + ing_form	are + ing_form	is + ing_form
Perfect	have + past_participle	have + past_participle	has + past_participle
Past			
Indefinite	past_form	past_form	past_form
Continuous	was + L_ing_form	were + ing_form	was + ing_form
Perfect	had + L_past_participle	had + past_participle	had + past_participle
Future			
Indefinite	shall + VR	will + VR	will + VR
Continuous	shall be + ing_form	will be + ing_form	is + ing_form
Perfect	shall have + past_participle	will have + past_participle	has + past_participle

Table 3: A dictionary format containing the different forms of verb.

verb_name	TP_singular	ing_form	past_form	past_participle
live	lives	living	lived	lived
read	reads	reading	read	read