Thesis Report

Example Based English to Bengali Machine Translation

Supervisor’s Name: Dr. Mumit Khan

Student Name: Khan Md. Anwarus Salam

Student’s ID: 07141002

Email: kmanwar@gmail.com

August 2008

DECLARATION

I hereby declare that this thesis is based on the results found by myself. Materials of work found by other researcher are mentioned by reference. This thesis, neither in whole nor in part, has been previously submitted for any degree.

Signature of
Supervisor

Signature of
Author

ACKNOWLEDGMENTS

I have to thank specially Dr. Mumit Khan for over viewing my thesis work to completion. I also thank all team members of Center for Research on Bangla Language Processing for assisting me in every step.

1. ABSTRACT

In this thesis we propose a new architecture for example based English to Bengali machine translation. The proposed Example Based Machine Translation (EBMT) system has five steps: 1) Tagging 2) Parsing 3) Prepare the chunks of the sentence using sub-sentential EBMT 4) Using an efficient adapting scheme match the sentence rule 5) Translate from English to Bengali in the chunk and generate output with morphological analysis. We prepared our tag set for tagging the English sentence. Here we proposed an optimal adapting scheme for choosing sentence rule from the knowledge base of the EBMT system. Our current system can translate simple sentences. We also defined a way to translate a complex sentence using sub-sentential EBMT. As this system can add more rules in the knowledge base, eventually it can be use for general purpose English to Bengali machine translation.

Keywords: Machine Translation, English to Bengali, EBMT, Example Based, Adaptation

TABLE OF CONTENTS

Page

Chapter 2: Introduction

The title of my research is "Example Based English to Bengali Machine Translation". The goal of this research topic is to develop efficient machine translation system for English to Bengali language. To develop an efficient machine translation system is very important but it is really expensive as it requires a huge amount of time and resources. In all languages there are many words that may have multiple meanings and also some sentence may have multiple grammar structure to express the same meaning, it is a great challenge to do the right semantic analysis. But it is very important to have a machine translation system which can compute all possible outputs in reasonable time and able to choose the best option.

In present there are many ways of machine translation system. Many researchers came up with different approaches. But still it is not possible to get the finest possible result. I want to use the example based machine translation system, to get all possible outputs. For achieving this I have to plan to prepare a dictionary with morphological analysis and a Parallel Corpus. Then from semantic analysis it may possible to choose the best desired output.

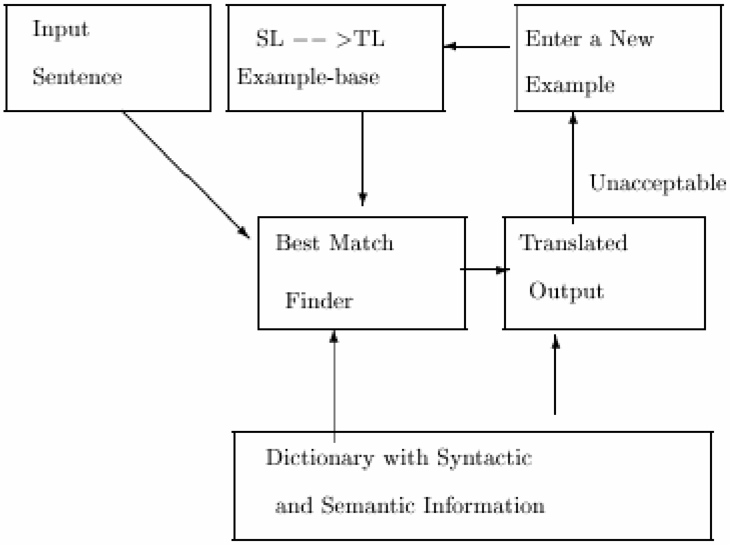


Figure 1: Role of Examples in Translation

Chapter 3: Background

3.1 What is Machine Translation?

Machine Translation is the process of translating text units of source language into a target language by using computers. The term Machine Translation can be defined as “translation from one natural language (source language (SL)) to another language (target language (TL)) using computerized systems, with or without human assistance” (Hutchins and Somers 1992, pg. 3).

3.2 Generations and Types of Machine Translation

Machine translation systems can be divided in two generations direct systems and indirect systems. First generation systems are known as direct systems. In such systems, translation is done word by word or phrase by phrase. In such systems very minimal linguistic analysis of input text is conducted (Hutchins and Somers 1992). This architecture is still being extensively used in commercial MT systems. The main idea behind direct systems is to analyze the input text to the extent that some transformational rules can be applied. This analysis could be parts of speech of words or some phrasal level information. Then using a bilingual dictionary, source language words are replaced with target language words and some rearrangement rules are used to modify the word order according to the target language (Arnold et al. 1993).

This architecture is very robust because it does not fail on any erroneous or ungrammatical input. Since the analysis level is very shallow and the system contains very limited grammatical information, it hardly considers anything ungrammatical. In the worst case if the rule does not apply to the input, the input is passed on without any alteration as output. This kind of system is hard to extend because all the rules are written in one direction and are language specific. To make another language pair work, all the rules have to be re-written. Since the system does not perform very deep analysis, its time complexity is low. These systems work very well for closely related languages but are not suitable for modeling languages with diverse syntactic nature. Since the system does not explicitly contain the grammatical rules of the target language, there is a chance that the output will not be grammatical but it will be similar to the target language (Arnold et al. 1993)

Owing to the fact that linguistic information helps an MT system to produce better quality target language translation, with the advance of computing technology, MT researchers started to develop methods to capture and process the linguistics of sentences. This was when the era of second generation MT systems started. Second generation machine translation systems are called indirect systems. In such systems the source language structure is analyzed and text is transformed into a logical form. The target language translation is then generated from the logical form of the text (Hutchins and Somers 1992). The transition from direct systems to indirect systems is illustrated in Figure 2.1, taken from (Hutchins and Somers 1992, pg. 107).

SYSTRAN is one of the most well-known direct systems. It is described in Hutchins and Somers (1992) and Wilks (1992). Indirect systems can be further divided into Interlingua and Transfer based systems.

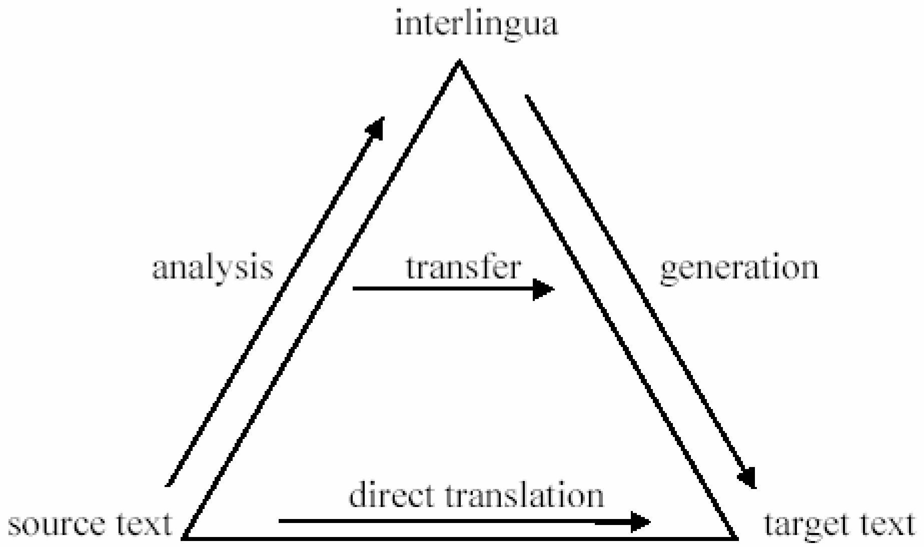


Figure 2.1: Transfer and interlingua 'pyramid' diagram

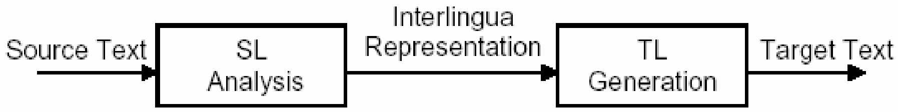


Figure 2.2a: Interlingua Based System

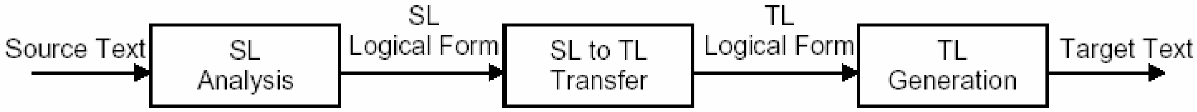


Figure 2.2b: Transfer Based System

In the transfer method, the source language is analyzed to an abstract level. Then, through a transfer module, this abstract form is converted to the corresponding abstract form in the target language through which the target translation text is generated. The module '*SL Analysis*' captures the required linguistic information about the source language sentences to aid the translation. '*SL to TL Transfer*' module transfers the representation generated by '*SL Analysis*' to a target language representation. The module '*TL Generation*' generates the translation text using this logical representation. Such a system requires independent grammars for the source and target languages. Moreover it requires a comparative grammar or transfer roles to relate source structures to target structures. Since the system assumes full grammatical knowledge it does not allow ungrammatical sentences to be parsed, thus reducing the output of the system. This kind of system is easy to extend because to add a new language, grammar and transfer rules for the new language need to be written but the grammar of the other language is reusable. Such systems are theoretically reversible. The same grammars can be used in the reversed system. Practically there are problems in reversing the system because some transfer rules which are correct in one direction may not be correct in the other direction. The system has the explicit grammar of the target language, which ensures grammatical output (Arnold et al. 1993). Examples of transfer systems include ARIANE (Vauquois and Boitet 1985), SUSY (Maas 1987), MU (the Japanese National Project) (Nagao et al. 1986), METAL (Slocum et al. 1987; Bennett and Slocum 1988), TAUM-AVIATION (Isabelle 1987), ETAP-2 (Apresian et al. 1992), LMT (McCord 1989), EUROTRA (Arnold 1986; Arnold and des Tombe 1987;

Copeland et al. 1991a,b), CAT-2 (Sharp 1988), MIMO (Arnold and Sadler 1990), MIMO-2 (van Noord et al. 1990) and ELU (Estival et al. 1990).

The Interlingua approach involves the use of an intermediate language (i.e. an Interlingua) for the transfer, with the source language text translated to the Interlingua and the Interlingua translated to the target language text. As suggested by Hutchins and Somers (1992), an Interlingua is an intermediate 'meaning' representation and this representation:

"*includes all information necessary for the generation of the target text without 'looking back' to the original text. The representation is thus a projection from the source text and at the same time acts as the basis for the generation of the target text; it is an abstract representation of the target text as well as a representation of the source text.*" (Hutchins and Somers 1992, p. 73)

Interlingua appears to be an attractive approach for machine translation due to several reasons. Firstly, from a theoretical point of view it is very interesting to establish a representation which is independent of language. Secondly, Interlingua systems are more easily extendable because only analysis and generation modules are required to add a new language and no language specific transfer information is needed. But it is difficult to define such a language independent representation even for closely related languages (Arnold et al. 1993).

An attempt to define an Interlingua to represent the language in the form of a semantic relation is The Universal Networking Language (UNL) project. This project was initiated by the University of United Nations based in Tokyo in 1996. An utterance is represented as a hyper-graph in UNL. Normal nodes in the graph bear Universal Words (UWs) with semantic attributes and arcs bear semantic relations (deep cases, such as agt, obj, goal, etc.). UNL representation is being built in many languages including Arabic, Chinese, French, German, Hindi, Indonesian, Italian, Japanese, Mongolian, Portuguese, Russian, and Spanish. Some other Interlingua systems are Rosetta (Landsbergen 1987b,a), KBMT (Goodman 1989; Goodman and Nirenburg 1991). (Arnold et al. 1993).

There are new emerging approaches to MT known as the empirical approaches. They apply statistical or pattern matching techniques for MT. These techniques are called empirical since the knowledge for translation is derived empirically by examining text instead of linguistic rules. There are two such approaches, the ‘example’ or ‘analogy’ based approach, and the ‘statistical’ approach (Arnold et al. 1993).

In the ‘example-based’ approach, translation is done by matching the given text with stored example translations. The basic idea is to collect a bilingual corpus of translation pairs and then use a best match algorithm to find the closest example to the source phrase to be translated. This gives a translation template, which can then be filled in by a word for word translation. A limitation of this technique is that it requires a large bilingual aligned corpus. But these examples can also be built incrementally, increasing the quality of translation. Such systems are efficient because they need not to go through complex grammars to analyze the text, but if many examples match the input text then finding the best match can be a complex task. A pure example based system will include no linguistic knowledge but addition of some linguistic knowledge can improve the system by increasing its capability of dealing with more patterns concisely as one can specify categories instead of raw words (Arnold et al. 1993).

The second approach, the ‘statistical approach’, uses probabilistic analysis in MT as the name suggests. This term sometimes refers to the use of probability based techniques in parts of the MT task like word sense disambiguation or structural disambiguation. The other use of this term refers to a pure statistical machine translation system which uses probabilistic models to determine the correct translation of input text. In this approach, two statistical models, namely a ‘language model’ and a ‘translation model’ are built. This technique has been successfully used in speech recognition. A language model provides probabilities of occurrence of the sentence in the language, P(S) and a translation model provides probability of a target sentence given source sentence, P(T/S). An N-gram model is used to build the language model. Language models for both source and target languages are built. The translation model is computed using a word-level aligned bilingual corpus. For details of the modeling process, refer to Brown et al. (1990). Using language model probabilities and conditional probabilities of the translation model, P(S/T) is computed using the following formula:

$$P(S/T) = \frac{P(S)P(T/S)}{P(T)}$$

This approach does not require explicit encoding of linguistic information. On the other hand, it is heavily dependent on the availability of good quality bilingual data in very large proportions, which is currently not available for most languages (Arnold et al. 1993).

3.3 Why Example-based Machine Translation?

Example-based Machine Translation makes use of past translation examples to generate the translation of a given input. An EBMT system stores in its example base of translation examples between two languages, the source language and the target language. These examples are subsequently used as guidance for future translation tasks. In order to translate a new input sentence in SL, similar SL sentence is retrieved from the example base, along with its translation in TL. This example is then adapted suitably to generate a translation of the given input. It has been found that EBMT has several advantages in comparison with other MT paradigms (Sumita and Iida, 1991):

1. It can be upgraded easily by adding more examples to the example base;
2. It utilizes translators' expertise, and adds a reliability factor to the translation;
3. It can be accelerated easily by indexing and parallel computing;
4. It is robust because of best-match reasoning.

Even other researchers (e.g. (Somers, 1999), (Kit et. al., 2002)) have considered EBMT to be one major and effective approach among different MT paradigms, primarily because it exploits the linguistic knowledge stored in an aligned text in a more efficient way. We apprehend from the above observation that for development of MT systems from English to Bengali, EBMT should be one of the preferred approaches. This is because a significant volume of parallel corpus is available between English and Bengali in the form of Newsletters, Bi-lingual websites, government notices, translation books, advertisement material etc. Although all data is generally not available in electronic form yet, converting them into machine readable form is much easier than formulating explicit translation rules as required by an EBMT system.

3.4 Difficulties of Example-based Machine Translation

- Can not use in general translation
- But improvable by increasing Knowledge Base
- Match sentence rule is very difficult
- No tools available

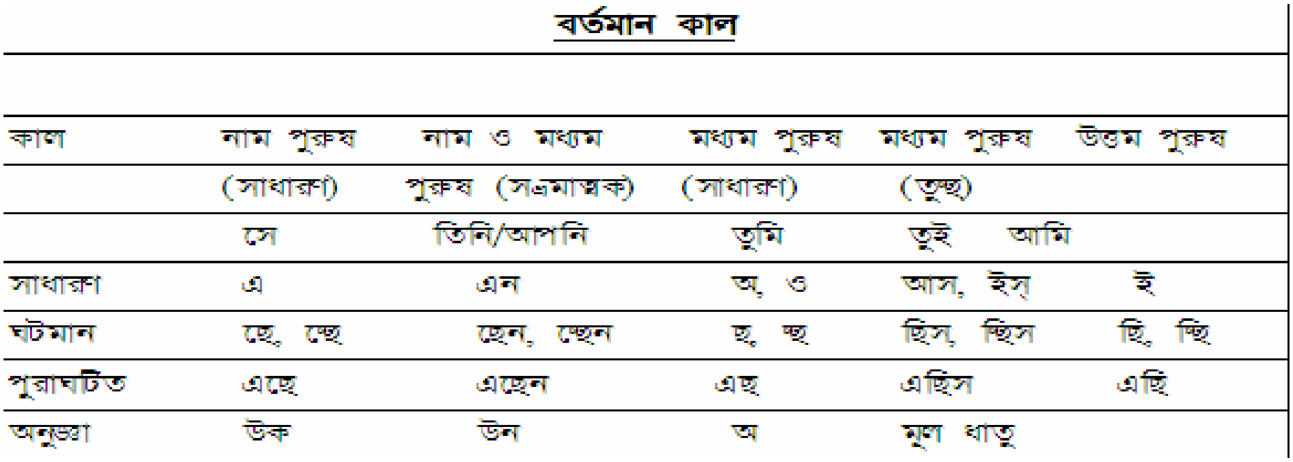
3.5 Initial Requirement for EBMT

- Prepare Language Model
- Generate Sentence Rule
- Morphological Analysis
- English to Bengali Dictionary with Analysis

Morphological Analysis

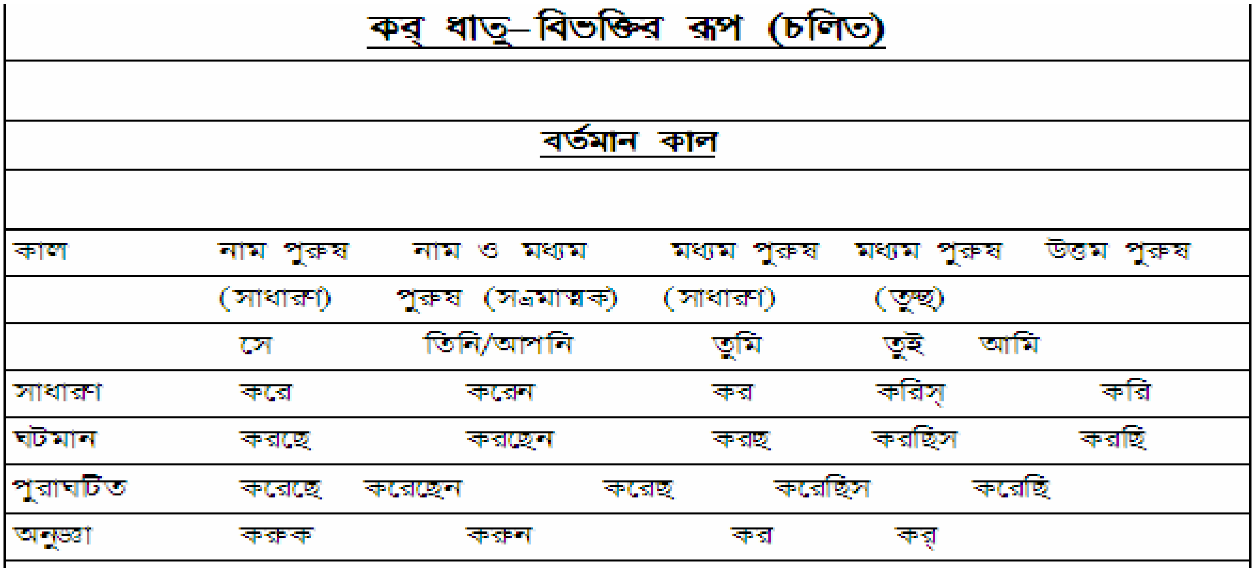
For English normally we can have 4 forms of a word. Eg. Do, Did, Done, Does

But in Bengali we may have nearly 20 forms of the same word meaning “koro”. Below we give the figure which explains the word forms.



বর্তমান কাল

কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	এ	এন	অ, ও	আস, ইস্	ই
ঘটমান	ছে, চ্ছে	ছেন, চ্ছেন	ছ, চ্ছ	ছিস, চ্ছিস	ছি, চ্ছি
পুরাঘটিত	এছে	এছেন	এছ	এছিস	এছি
অনুজ্ঞা	উক	উন	অ	মূল ধাতু	

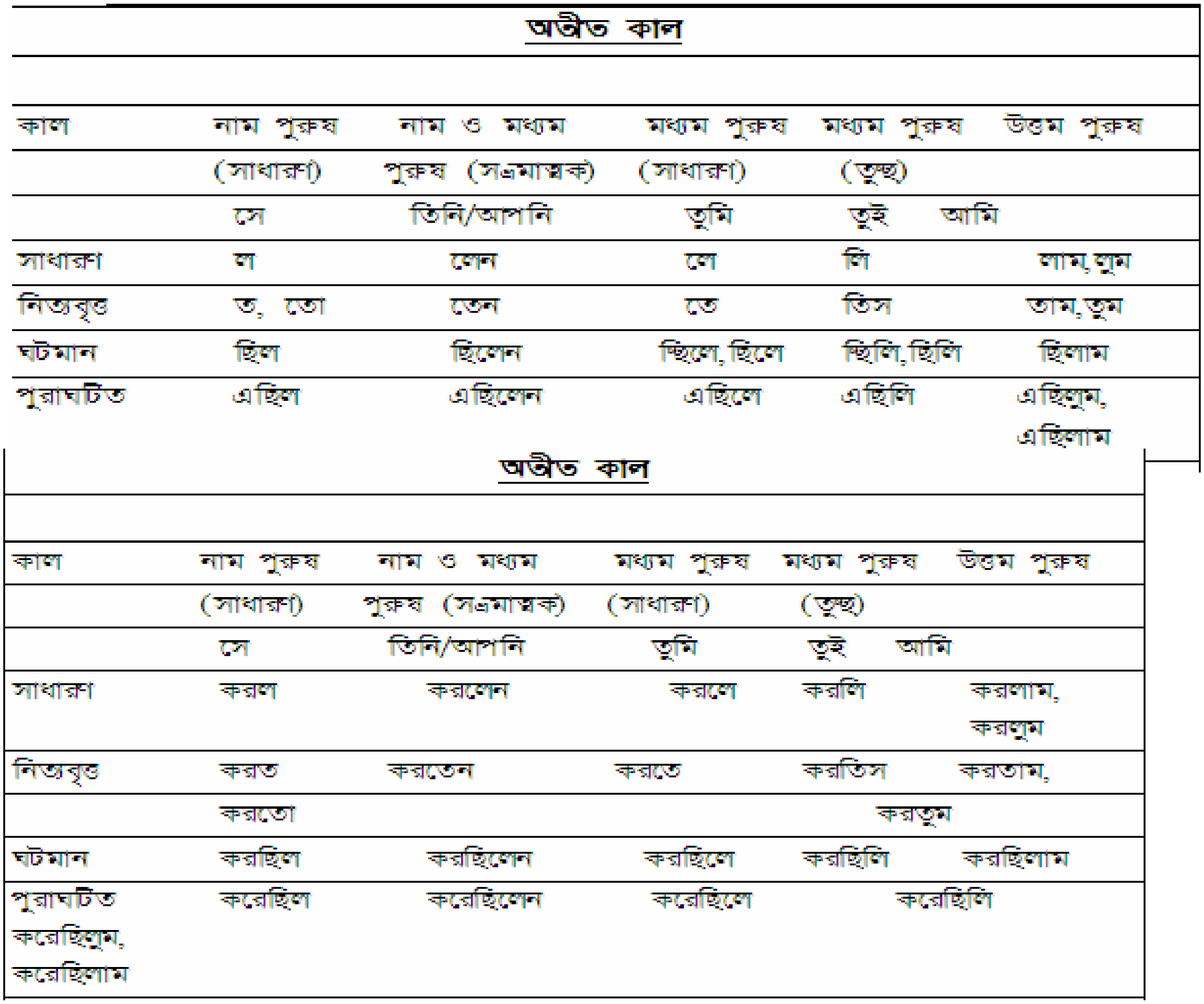


কর্ ধাতু–বিভক্তির রূপ (চলিত)

বর্তমান কাল

কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	করে	করেন	কর	করিস্	করি
ঘটমান	করছে	করছেন	করছ	করছিস	করছি
পুরাঘটিত	করেছে	করেছেন	করেছ	করেছিস	করেছি
অনুজ্ঞা	করুক	করুন	কর	কর্	

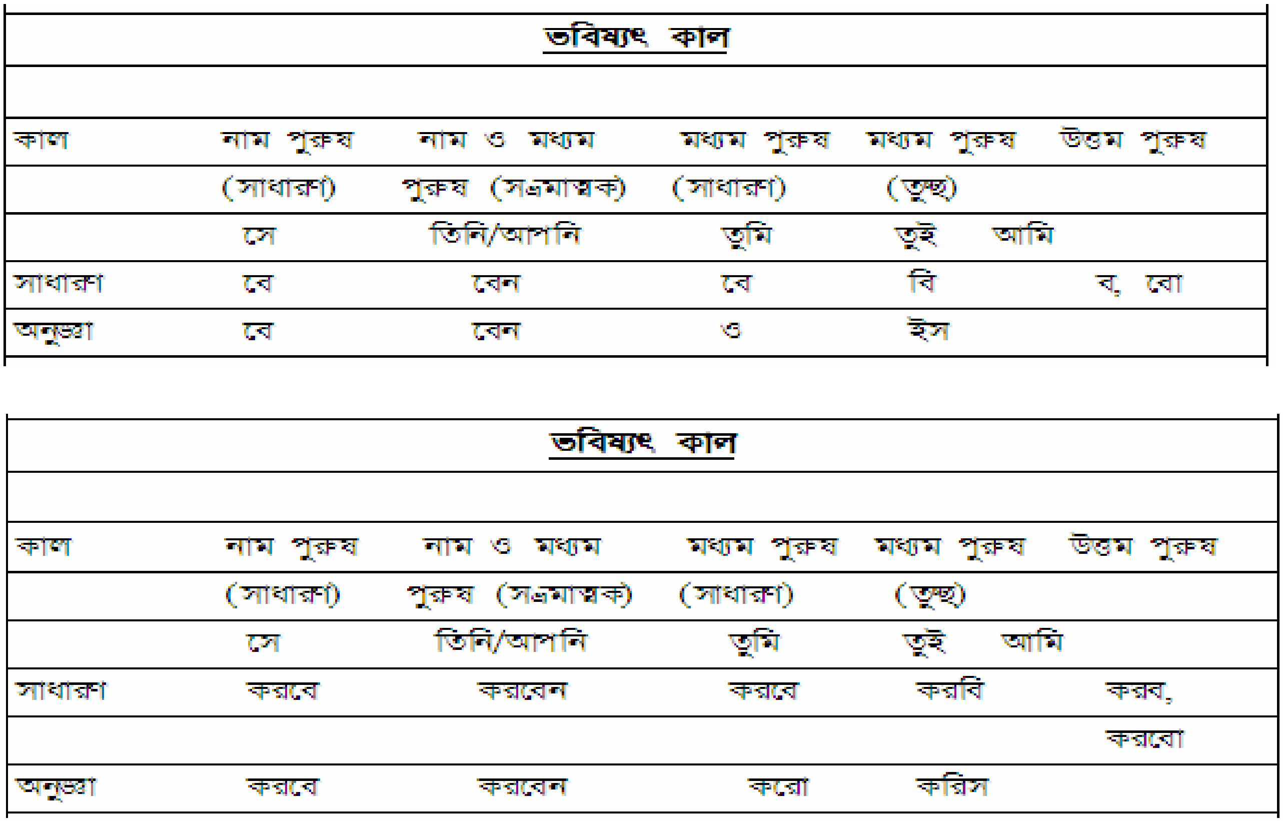
Figure: For present Tense



অতীত কাল					
কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	ল	লেন	লে	লি	লাম, লুম
নিত্যবৃত্ত	ত, তো	তেন	তে	তিস	তাম, তুম
ঘটমান	ছিল	ছিলেন	চ্ছিলে, ছিলে	চ্ছিলি, ছিলি	ছিলাম
পুরাঘটিত	এছিল	এছিলেন	এছিলে	এছিলি	এছিলুম, এছিলাম

অতীত কাল					
কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	করল	করলেন	করলে	করলি	করলাম, করলুম
নিত্যবৃত্ত	করত	করতেন	করতে	করতিস	করতাম,
	করতো				করতুম
ঘটমান	করছিল	করছিলেন	করছিলে	করছিলি	করছিলাম
পুরাঘটিত করেছিলুম, করেছিলাম	করেছিল	করেছিলেন	করেছিলে	করেছিলি	

Figure: For PastTense



ভবিষ্যৎ কাল					
কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	বে	বেন	বে	বি	ব, বো
অনুজ্ঞা	বে	বেন	ও	ইস	

ভবিষ্যৎ কাল					
কাল	নাম পুরুষ	নাম ও মধ্যম	মধ্যম পুরুষ	মধ্যম পুরুষ	উত্তম পুরুষ
	(সাধারণ)	পুরুষ (সম্ভ্রমাত্মক)	(সাধারণ)	(তুচ্ছ)	
	সে	তিনি/আপনি	তুমি	তুই	আমি
সাধারণ	করবে	করবেন	করবে	করবি	করব,
					করবো
অনুজ্ঞা	করবে	করবেন	করো	করিস	

Figure: For Future Tense

Table1: Sample knowledge base of the English to Bengali EBMT System

English	English Chunk	Transfer to Bengali Chunk	Bengali
He reads a book	[NP He/PRP] [VP reads/VBZ] [NP a/DT book/NN]	[NP সে/PRP] [VP পড়ে /VBZ] [NP একটি/DTবই/NN]	সে একটি বই পড়ে
The sun Rises in the east	[NP The/DT sun/NN Rises/NNS] [PP in/IN] [NP the/DT east/JJ]	[NP /DT সূর্য /NN উদিত /NNS] [PP হয়/IN] [NP /DT পূর্বে /JJ]	সূর্য পূর্বে উদিত হয়
He is reading a book	[NP He/PRP] [VP is/VBZ reading/VBG] [NP a/DT book/NN]	[NP সে /PRP] [VP /VBZ পড়ছে /VBG] [NP এক টি /DT বই /NN]	সে এক টি বই পড়ছে
I am reading a book	[NP I/PRP] [VP am/VBP reading/VBG] [NP a/DT book/NN]	[NP আমি /PRP] [VP /VBP পড়ছি /VBG] [NP এক টি /DT বঈ /NN]	আমি এক টি বঈ পড়ছি
They are reading a book	[NP They/PRP] [VP are/VBP reading/VBG] [NP a/DT book/NN]	[NP তারা /PRP] [VP /VBP পড়ছে /VBG] [NP এক টি /DT বই /NN]	তারা এক টি বই পড়ছে
I have done the work	[NP I/PRP] [VP have/VBP done/VBN] [NP the/DT work/NN]	[NP আমি /PRP] [VP /VBP কাজটি /VBN] [NP টি/DT কাজ/NN]	আমি কাজটি করেছি
He has gone to Dhaka	[NP He/PRP] [VP has/VBZ gone/VBN] [PP to/TO] [NP Dhaka/NNP]	[NP সে /PRP] [VP /VBZ গিয়াছে /VBN] [PP /TO] [NP ঢাকা /NNP]	সে ঢাকা গিয়াছে
They have livedat this house five	[NP They/PRP] [VP have/VBP] [PP livedat/IN] [NP this/DT	[NP তারা /PRP] [VP করছে /VBP] [PP বাস /IN] [NP এই /DT বাড়িতে /NN	তারা এই বাড়িতে পাচ বছর যাবত বাস করছে

years.	house/NN] [NP five/CD years/NNS] ./.	] [NP পাচ /CD বছর /NNS] ./.	
He has been reading the book for two hours	[NP He/PRP] [VP has/VBZ been/VBN reading/VBG] [NP the/DT book/NN] [PP for/IN] [NP two/CD hours/NNS] ./.	[NP সে /PRP] [VP /VBZ যাবত /VBN পড়ছে /VBG] [NP টি /DT বই /NN] [PP for/IN] [NP দুই /CD ঘন্টা /NNS] ./.	সে দুই ঘন্টা যাবত বই টি পড়ছে
I did the work	[NP I/PRP] [VP did/VBD] [NP the/DT work/NN] ./.	[NP আমি /PRP] [VP করছিলাম /VBD] [NP টি/DT কাজ/NN] ./.	আমি কাজটি করছিলাম
He went home yesterday	[NP He/PRP] [VP went/VBD] [NP home/NN] [NP yesterday/NN] ./.	[NP সে /PRP] [VP গেল /VBD] [NP বাড়ি /NN] [NP গতকাল /NN] ./.	সে গতকাল বাড়ি গেল
He wrote the letter	[NP He/PRP] [VP wrote/VBD] [NP the/DT letter/NN] ./.	[NP সে /PRP] [VP লিখলো /VBD] [NP এক /DT চিঠি /NN] ./.	সে এক টি চিঠি লিখলো
The boys were playing	[NP The/DT boys/NNS] [VP were/VBD playing/VBG] ./.	[NP /DT বালকগুলো /NNS] [VP /VBD খেলতে ছিল /VBG] ./.	বালকগুলো খেলতে ছিল
You will do the sum	You/PRP] [VP will/MD do/VB] [NP the/DT sum/NN]	[NP tumi/PRP] [VP /MD korobe/VB] [NP Ti/DT Angk/NN]	তুমি অংক টি করবে
He will be doing the work	[NP He/PRP] [VP will/MD be/VB doing/VBG] [NP the/DT work/NN]	[NP se/PRP] [VP thak/MD be/VB korIte/VBG] [NP Ti/DT kaj/NN]	সে কাজ টি করতে থাকবে
You will be reding The book	[NP You/PRP] [VP will/MD be/VB reding/VBG] [NP The/DT book/NN]	[NP tumi/PRP] [VP thak/MD be/VB pRite/VBG] [NP Ti/DT bo`i/NN]	তুমি বই টি পড়িতে থাকবে

Goals of the thesis

The primary goal of this thesis is to study various aspects of designing an EBMT system for translation from English to Bengali. It may be observed that in today's world a lot of information is being generated around the world in various fields. However, since most of this information is in English, it remains out of reach of people at large for which English is not the language of communication. As a consequence, an increasing demand for developing machine translation systems from English to Bengali is being felt very strongly. However, development of MT systems typically demands availability of a large volume of computational resources, which is currently not available for Bengali. Moreover, generating such a large volume of computational resources (which may comprise an extensive rule base, a large volume of parallel corpora etc.) is not an easy task. EBMT scheme, on the other hand, is less demanding on computational resources making it more feasible to implement in respect of these languages.

I will try to find suitable solutions for the following aspects:

a) Development of efficient retrieval and adaptation scheme: Efficient adaptation of past examples is a major aspect of an EBMT system. There are many adaptation schemes available for an EBMT system. Even an efficient similarity measurement scheme and a quite large example base cannot guarantee an exact match for a given input sentence. As a consequence, there is a need for an efficient and systematic adaptation scheme for modifying a retrieved example, and thereby generating the required translation.

b) Study of divergence for English to Bengali translation, and how translation divergence can be effectively handled within an EBMT framework.

c) How to handle complex sentences - which are in general considered to be difficult to deal with in an MT system.

Chapter 4: Proposed Architecture of Example Based English to Bengali Machine Translation

The proposed EBMT system has five steps

1. Tagging the English sentence
2. Parsing the English sentence
3. Using sub-sentential EBMT prepare the chunks of the sentence
4. Using an efficient adapting scheme match the sentence rule.
5. Translate from English to Bengali in the chunk and generate with morphological analysis

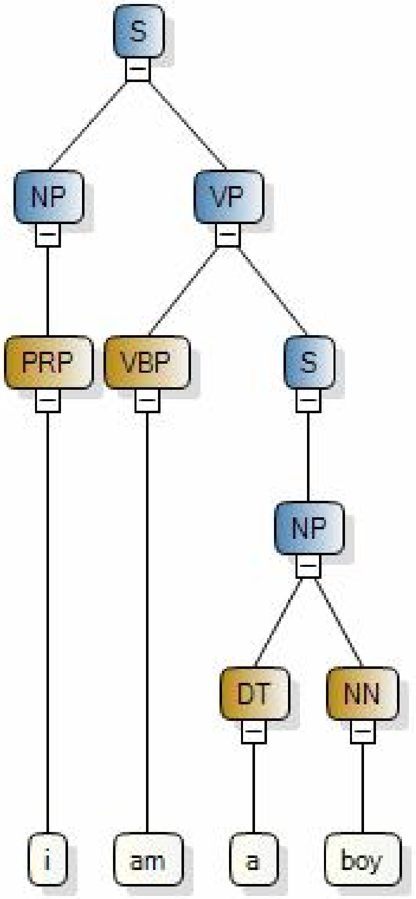
4.1 Tagging and Parsing

Tagging, is the process of marking up the words in a text as corresponding to a particular part of speech, based on both its definition, as well as its context—i.e., relationship with adjacent and related words in a phrase, sentence, or paragraph. Eg. I do-> I/PRP do/VBP

Parsing, is the process of analyzing a sequence of tokens to determine grammatical structure with respect to a given formal grammar. We used the tag set of Table2 for tagging the English sentence. Eg. Eg. I am a boy-> (S (NP (PRP I)) (VP (VBP am) (NP (DT a) (NN boy))))

Table2: Tag set used for English to Bengali EBMT

Level 1	Level 2	Tag
Noun	Common	NN
	Proper	NNP
	Compound Common Noun	NNC
	Compound Proper Noun	NNPC
	Verb Root	NNV
	Temporal	NNT
	Question Temporal	QNT
	Locative	NNL
	Question Locative	QNL
Pronoun	Personal Pronoun	PRP
	Question Pronoun	QPR
Adjective	Simple	JJ
	Verb Root	JJV
	Question Adjective	QJJ
Vocatives	Vocatives	VOC
Verb	Main Finite Verb	VB
	Nonfinite Nominal	VBM
	Nonfinite Conditional	VBC
	Nonfinite Perfective	VBT
	Nonfinite	VBF
	Past tense	VBD



	Gerund/present participle	VBG
	Past participle	VBN
	Non-3rd ps. sing. Present	VBP
	3rd ps. sing. Present	VBZ
	Existential	VBE
Adverb	Adverb	RB
	Question Adverb	QRB
Conjunction	Co-ordinating	CC
	Compound Co-ordinating	CCC
	Suspicion	CN
	Eternal Joining	CET
	Subordinating	CS
	Compound Subordinating	CSC
Numbers	Cardinal Numbers	CD
Adposition	Adposition	ON
Interjection	Interjection	UH
Particle	Particle	RP
	Question Particle	QRP
Determiner	Common	DT
	Singular	DTS
	Question Determiner	QDT
Quantifier	Quantifier	QF
Foreign Word	Foreign Word	FW
Symbol	Symbol	SYM
List Item Marker	List Item Marker	LS
Suffixes	Adpositional	SFON
	Accusative	SFAC
	Possessive	SF$
Punctuation Marks & Others	Sentence Final Punctuation	.
	Comma	,
	Colon, Semi-colon	:
	Dash, Double-Dash	-
	Left Parenthesis	(
	Right Parenthesis	)
	Opening Left Quote	LQ
	Closing Right Quote	RQ
	Preposition/subordinate conjunction	IN

	Adjective, superlative	JJS
	Adjective, comparative	JJR
	Modal	MD
	Proper noun, plural	NNPS
	Noun, plural	NNS
	Predeterminer	PDT
	Possessive ending	POS
	Possessive pronoun	PRP$
	Adverb, comparative	RBR
	Adverb, superlative	RBS
	to	TO
	wh-determiner	WDT
	wh-pronoun	WP
	Possessive wh-pronoun	WP$
	wh-adverb	WRB
	Left open double quote	``
	Comma	,
	Right close double quote	'
	Sentence-final punctuation	.
	Colon, semi-colon	:
	Dollar sign	$
	Pound sign	#
	Left parenthesis (	-LRB-
	Right parenthesis	-RRB-

4.2 Handle Complex Sentence Using Sub-Sentential EBMT:

Handling complex sentence in general considered to be difficult to deal with in an MT system. Since exact sentence matches only occur in special domains, we want to extend this to sub-sentence matches. For this we need to:

- Find the most similar example (involves segmenting by preparing chunks)
- Alter source side to match current input.

Similarity requires a “distance metric” in the source language (English).

This can be closeness:

- of the lexical items in a hierarchy of terms/ concepts from ontology
- of the sequence of syntactic categories and function words,
- of the two syntactic structures,
- or combinations of these.

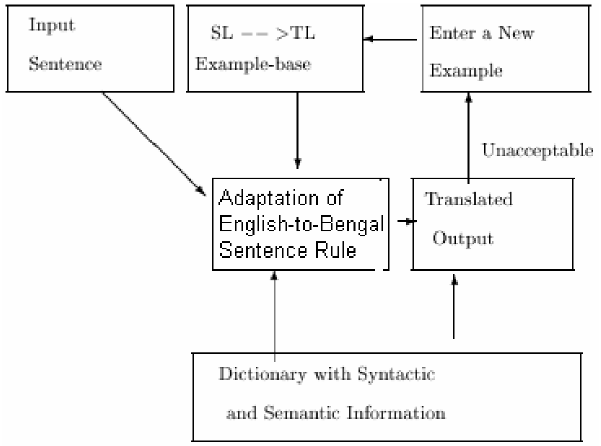

Input Sentence
SL -- >TL Example-base
Enter a New Example
Unacceptable
Adaptation of English-to-Bengal Sentence Rule
Translated Output
Dictionary with Syntactic and Semantic Information

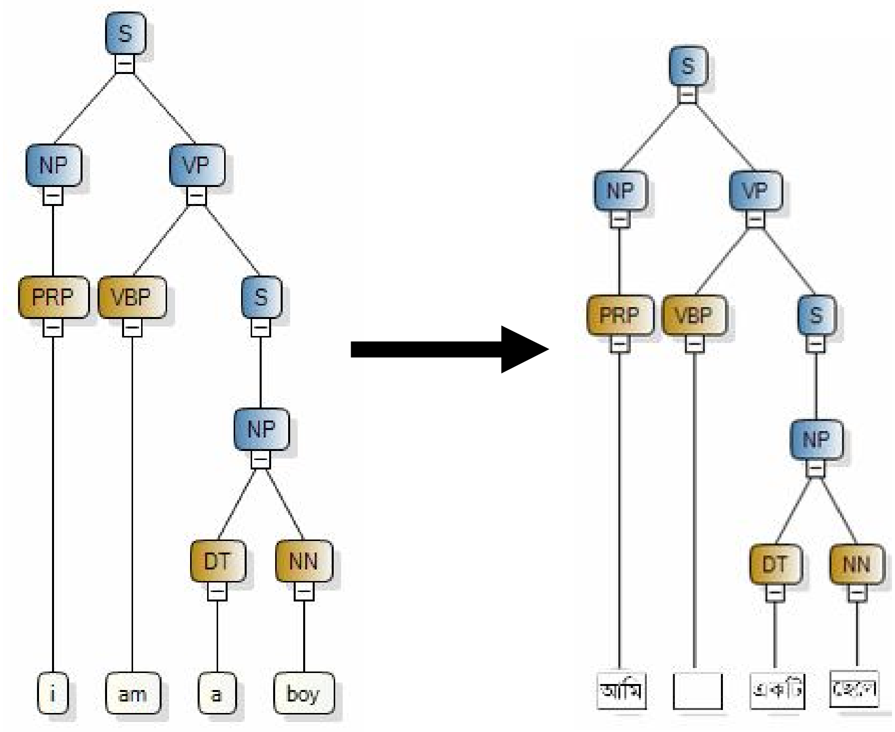
Figure 1: Role of Examples in Translation

4.3 Adapting Scheme to Match Sentence Rule

Efficient adaptation of past examples is a major aspect of an EBMT system. There are many adaptation schemes available for an EBMT system. Even an efficient similarity measurement scheme and a quite large example base cannot guarantee an exact match for a given input sentence. As a consequence, there is a need for an efficient and systematic adaptation scheme for modifying a retrieved example, and thereby generating the required translation. In section 5 we discuss details about our proposed adaptation scheme. In Table1 we gave a sample knowledge base of the English to Bengali EBMT System. During translation our adapting scheme chooses the best rule for the source sentence.

4.4 Match the sentence rule from the Knowledge Base

Figure: Tree Conversion for English to Bengali



4.5 Translate from English to Bengali

Study of divergence for English to Bengali translation is also required. Translation divergence can be effectively handled within an EBMT framework. As in earlier step we have the sample rule and the parsed sentence. Now we can easily translate the sentence by matching the rule.

Chapter 5. Adaptation in English to Bengali Translation

A successful EBMT system requires a good adaptation scheme. The need for an efficient and systematic adaptation scheme arises for modifying a retrieved example, and thereby generating the required translation. Researcher came with various approaches to deal with adaptation aspect of an EBMT system. Overall the adaptation procedures employed in different EBMT systems primarily consist of four operations:

- **Copy**, where the same chunk of the retrieved translation example is used in the generated translation;
- **Add**, where a new chunk is added in the retrieved translation example;
- **Delete**, when some chunk of the retrieved example is deleted; and
- **Replace**, where some chunk of the retrieved example is replaced with a new one to meet the requirements of the current input.

5.1 Adaptability/Mappability for a chunk has 4 discrete values:

Depending on the Level 3: Source Language (SL): Target Language (TL) mapping is one-to-one for all words
Level 2: Syntactic Functions map, but not some POS tags
Level 1: Functions differ, but lexical correspondence still holds
Level 0: Cannot establish correspondence

5.2 Description of the Adaptation Operations

1. Constituent Word Replacement (WR): One may get the translation of the input sentence by replacing some words in the retrieved translation example. Suppose the input sentence is: \The bird was eating apples.", and the most similar example retrieved by the system (along with its Bengali translation) is: \The elephant was eating fruits."\haathii phol khacchilo". The desired translation may be generated by replacing \haathii" with the Bengali of \birds", i.e. \pakhi" and replacing \phal" with the Bengali of \apples", i.e. \aapel". These are examples of the operation of constituent word replacement.

2. Constituent Word Deletion (WD): In some cases one may have to delete some words from the translation example to generate the required translation. For example, suppose the input sentence is: \Animals were dying of thirst". If the retrieved translation example is : \Birds and Animals were dying of thirst." \pakhi ebong pashu trishnay mara jacche", then the desired translation can be obtained by deleting \pakhi ebong" (i.e the Bengali of \birds and") from the retrieved translation. Thus the adaptation here requires two constituent word deletions.

3. Constituent Word Addition (WA): This operation is the opposite of constituent word deletion. Here addition of some additional words in the retrieved translation example is required for generating the translation. For illustration, one may consider the example given above with the roles of input and retrieved sentences being reversed.

4. Morpho-word Replacement (MR): In this case one morpho-word is replaced by another morpho-word in the retrieved translation example. For illustration, if the input sentence is \He eats rice.", and the retrieved example is: \He is reading a book." _ \se akte boi porChe", then to obtain the desired translation[4] first the morpho-word \Che" is to be replaced by “\e”

5. Morpho-word Deletion (MD): Here some morpho-word(s) are deleted from the retrieved translation example.

6. Morpho-word Addition (MA): This is the opposite case of morpho-word deletion. Here some morpho-words need to be added in the retrieved example in order to generate the required translation.

7. Suffix Replacement (SR): Here the suffix attached to some constituent word of the retrieved sentence is replaced with a different suffix to meet the current translation requirements. This may happen with respect to noun, adjective verb, or case ending.

8. Suffix Deletion (SD): By this operation the suffix attached to some constituent word may be removed, and thereby the root word may be obtained.

9. Suffix Addition (SA): Here a suffix is added to some constituent word in the retrieved example.

10. Copy (CP): When some word (with or without suffix) of the retrieved example is retained in to in the required translation then it is called a copy operation.

5.3 Study of Adaptation Procedure for Morphological Variation of Active Verbs

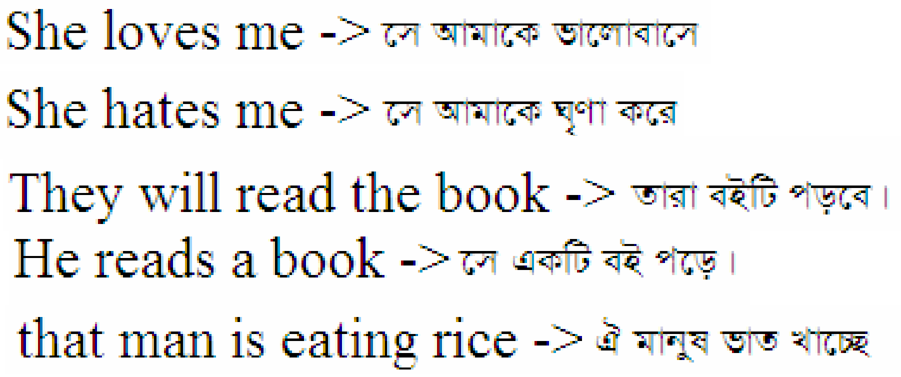
Verb morphology variations are divided into four groups. These are:

1. Same tense same verb form
2. Different tenses same verb form
3. Same tense different verb forms
4. Different tenses different verb forms

Chapter6. Implementation of the proposed EBMT system

Currently our system is working for English to Bengali Translation. Our current system can translate simple sentences which are given in the knowledge base. We also defined a way to translate a complex sentence using sub-sentential EBMT. So while separating the chunk in the machine translation process we can use our knowledge base for translating the separated small chunks of the sentence. As this system can add more rules in the knowledge base, eventually it can be used for general purpose English to Bengali machine translation.

Sample Outputs
Our system can now do following types of English to Bengali translation:



She loves me -> সে আমাকে ভালোবাসে

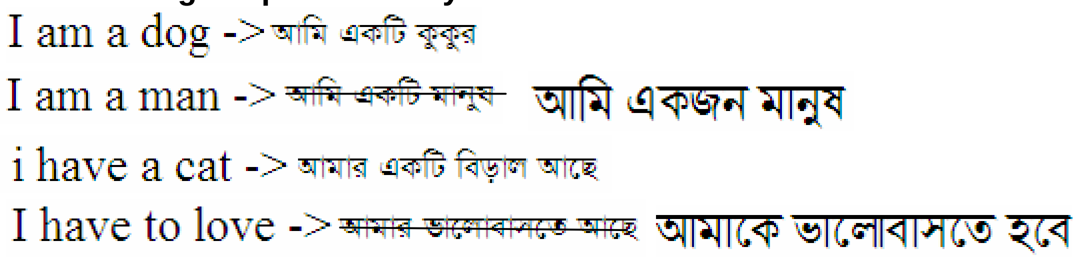
She hates me -> সে আমাকে ঘৃণা করে

They will read the book -> তারা বইটি পড়বে।

He reads a book -> সে একটি বই পড়ে।

that man is eating rice -> ঐ মানুষ ভাত খাচ্ছে

Some Wrong Output! But Why?



I am a dog -> আমি একটি কুকুর

I am a man -> ~~আমি একটি মানুষ~~ আমি একজন মানুষ

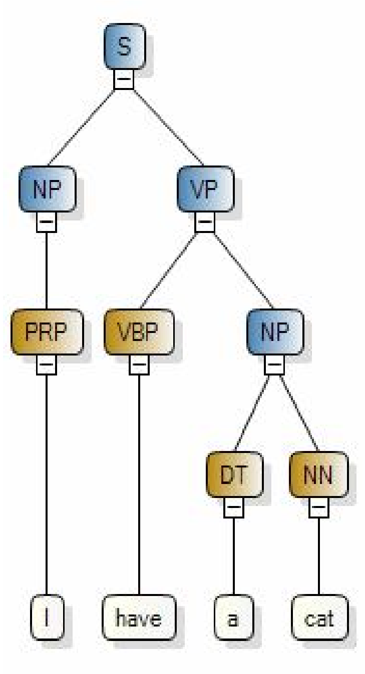
i have a cat -> আমার একটি বিড়াল আছে

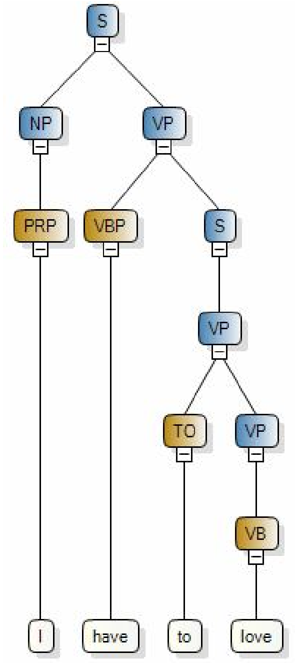
I have to love -> ~~আমার ভালোবাসতে আছে~~ আমাকে ভালোবাসতে হবে

Now we can see that,

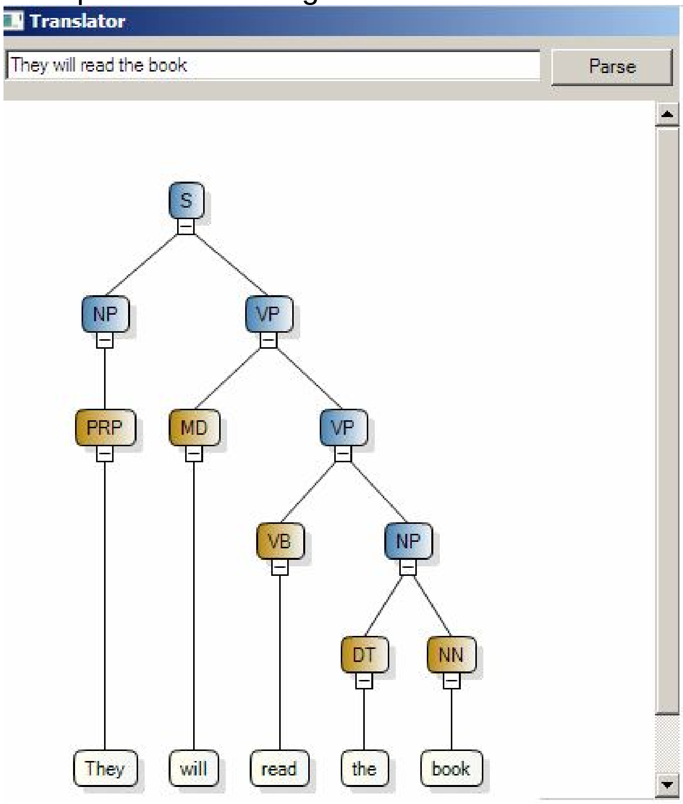
I have a cat = PRP + VBP + NP
I have to love = PRP + VBP + S

If we add a new rule for “PRP + VBP + S” in our knowledge base then we can also translate “I have to love” properly. In this way we can update the system.

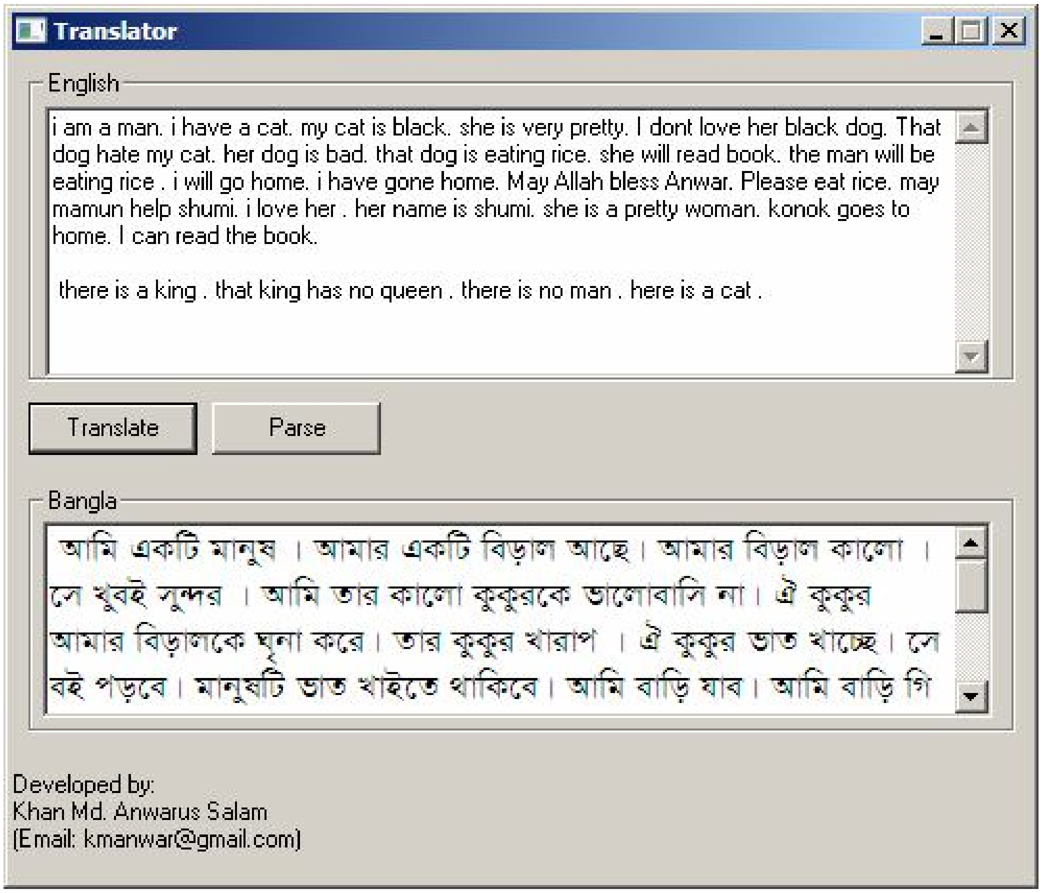




The screen shots of our implementation is given below:



They will read the book



i am a man. i have a cat. my cat is black. she is very pretty. I dont love her black dog. That dog hate my cat. her dog is bad. that dog is eating rice. she will read book. the man will be eating rice . i will go home. i have gone home. May Allah bless Anwar. Please eat rice. may mamun help shumi. i love her . her name is shumi. she is a pretty woman. konok goes to home. I can read the book.

there is a king . that king has no queen . there is no man . here is a cat .

আমি একটি মানুষ । আমার একটি বিড়াল আছে। আমার বিড়াল কালো । সে খুবই সুন্দর । আমি তার কালো কুকুরকে ভালোবাসি না। ঐ কুকুর আমার বিড়ালকে ঘৃনা করে। তার কুকুর খারাপ । ঐ কুকুর ভাত খাচ্ছে। সে বই পড়বে। মানুষটি ভাত খাইতে থাকিবে। আমি বাড়ি যাব। আমি বাড়ি গি

Developed by:
Khan Md. Anwarus Salam
(Email: kmanwar@gmail.com)

Figure2: Screen shot of the implementation

Chapter 7. Limitations of our System and Future Work

Limitations:

- It can not differentiate different meaning of same word
  - Eg. Let the light light light (3 meaning)
- It can not identify the names unless it is given in the dictionary.
- It can not translate if the given sentence do not match in the knowledge base
- It can not determine different meaning of ‘a’ for pen/elephant. Different translation of animate/non animate objects etc (Need Bangla Wordnet)
- Can not do voice, narration or thematic translation

Future Work:

- Connect with wordnet. (But wordnet does not have all sufficient information yet. But we can prepare the bangla wordnet)
- Make a machine learning system so that user can train it (HAMT, Suggest a translation). We can then improve its efficiency for general purpose use
- Transliterate nouns. But due to limited dictionary we also miss many actual words which can create problems

Chapter 8. Conclusion

We have presented a new approach for English to Bengali machine translation. Right now our system can translate English to Bengali sentence. But it has limited knowledge base. By increasing the knowledge base we can improve its efficiency for general purpose use. We can extend this research for doing EBMT based Bengali to English translation as well.

Chapter 9: Reference

[1]. Book: Machine Translation: An Introductory Guide , *By Doug Arnold, Lorna Balkan, Siety Meijer, R.Lee Humphreys, Louisa Sadler; Colchester, August 1993*

[2]. Contributions To English To Hindi Machine Translation Using Example-Based Approach, Phd Theses in January 2005, Deepa Gupta, IIT Delhi.

[3]. D. Gupta and N. Chatterje., Study of Divergence for Example Based English-Hindi Machine Translation. STRANS-2001, IIT Kanpur, 2001 pp. 43-51.

[4]. Balanced Bengali Language Corpus: A Proposal, *By Khan Md. Anwarus Salam, S M Murtoza Habib and Dr. Mumit Khan*, Research work in BRAC University in 2008*.*

[5]. H.A. Guvenir and I. Cicekli., Learning Translation Templates from Examples. Elsevier Science Ltd., 1998

[6]. R. Jain , R.M.K Sinha and A. Jain., ANUBHATRI: Using Hybrid Example-Based Approach for Machine Translation.. STRANS-2001, IIT Kanpur, 2001 pp. 20-32.

[7]. Verb Transfer For English To Urdu Machine Translation, Thesis by Nayyara Karamat, FAST-Lahore, 2006

[8]. An Optimal Way Towards Machine Translation from English to Bengali, By Sajib Dasgupta, Abu Wasif and Sharmin Azam. In the Proceedings of the 7th International Conference on Computer and Information Technology (ICCIT), Bangladesh, 2004.

[9]. Shah 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), Bangladesh, 2003.