A Phrasal EBMT System for Translating English to Bengali

Sudip Kumar Naskar
Computer Science and Engineering Department,
Jadavpur University
Kolkata, India, 700032
sudip_naskar@hotmail.com

Sivaji Bandyopadhyay
Computer Science and Engineering Department,
Jadavpur University
Kolkata, India, 700032
sivaji_cse_ju@yahoo.com

Abstract

The present work describes a Phrasal Example Based Machine Translation system from English to Bengali that identifies the phrases in the input through a shallow analysis, retrieves the target phrases using a Phrasal Example base and finally combines the target language phrases employing some heuristics based on the phrase ordering rules for Bengali. The paper focuses on the structure of the noun, verb and prepositional phrases in English and how these phrases are realized in Bengali. This study has an effect on the design of the phrasal Example Base and recombination rules for the target language phrases.

1 Introduction

Bengali is in the list of world's top 10 languages and is estimated to be spoken by more than 200 million people in Bangladesh, India and Singapore. Bengali is also referred to as Bangla by its native speakers. In order to translate from English to Bengali, the first thing we do is lexical analysis of the English sentence, to gather the lexical features of the morphemes. During morphological analysis, the root words / terms (including idioms and named entities), along with associated grammatical information and semantic categories are extracted. A shallow parser identifies the constituent phrases of the source language sentence and tags them to encode all relevant information that might be needed to translate this phrase and perhaps resolve ambiguities in other phrases. A phrasal Example Base is used to retrieve the target language phrase structure corresponding to each input phrase. Then these phrases are translated individually to the target language (Bengali) using Bengali synthesis rules. Finally, those target language phrases are arranged using some heuristics, based on the phrase ordering rules of Bengali, to form the target language representation of the source language sentence. The noun phrases (NP) and the prepositional phrases (PP) are translated using Example bases of syntactic transfer rules. These PPs can take the role of nouns and adverbs in certain cases. Verb phrase (VP) translation scheme is *rule based* and uses Morphological Paradigm Suffix Tables. In order to translate properly we also need to identify and translate gerunds, infinitives and participles.

The structure of noun phrase is somewhat similar in both English and Bengali. But the verb phrase and prepositional phrase (PP) constructions differ markedly in English and Bengali. First of all, in Bengali, there is no concept of preposition. English prepositions are handled in Bengali using inflexions to the reference objects, i.e., the noun that follows a preposition in a PP, and / or post-positional words after them. Moreover, inflexions in Bengali get attached to the reference objects and relate it with the main verb of the sentence in *case* or *karaka* relations. An inflexion has no existence of its own in Bengali, and it does not have any meaning as well, but in English prepositions have their own existence, i.e., they are separate words. Verb phrases in both English and Bengali depend on the person, number information of the subject and tense and aspect information of the verb. But for any particular root verb, there are only a few verb forms in English, whereas in Bengali it shows a lot of variations.

The structural divergence between English and Bengali suggests that a Phrasal EBMT system would suit best for translating English to Bengali. In the present system, grammar examples expressed as context-sensitive rewrite rules (using semantic features) are stored in a phrasal example base. Some references on Example Based Machine Translation (EBMT) strategy can be found in (Nagao, 1984; Furuse and Iida, 1992; Somers, 2000; Carl and Andy, 2003).

A hybrid approach based on fixed rules (rule-based) and stochastic methods (corpus-based) which extracts word order transfer rules between two languages has been presented in (Dien et. al., 2003). They have used this approach for translating from English to Vietnamese. They have used the Transfer-based Learning (TBL) machine learning approach for learning transfer rules from English-

Vietnamese bilingual corpus in order to transfer word orders from English into Vietnamese. An account of the structural and categorical divergence problem for English and Hindi have been given in (Gupta and Chatterjee, 2003). They have proposed an algorithm for identification of this divergence for an English-Hindi EBMT system. This identification helps in partitioning the example database into divergence / non-divergence categories, which in turn should facilitate efficient retrieval and adaptation in an EBMT system. The development of an English to Hindi Machine-Aided Translation System named AnglaHindi based on the ANGLABHARTI translation methodology has been described in (Sinha and Jain, 2003). AnglaBharti is a pattern directed rule based system with context free grammar like structure for analysis of English as source language. It analyses English sentences and creates an intermediate structure that has the word and word-group order as per the structure of the group of target languages. The intermediate structure is then converted to final translation in the target language through a process of text-generation.

The next three sections 2, 3 and 4 describe the English NPs, VPs and PPs and how these phrases are translated in Bengali. Section 5 discusses about the phrase ordering rules of English and Bengali based on which some heuristics have been devised that direct how translated Bengali phrases are rearranged to generate the final Bengali output. Section 6 gives a brief account of the system. The evaluation technique is outlined in section 7. The conclusion is drawn in section 8.

2 English NPs and their Bengali Translation

The basic structure of an NP in English is: [specifier/article] [adv] [adj] [noun] [plural marker] [case marker], and in Bengali: [specifier] [adv] [adj] [noun] [plural marker] [case marker]. One or many of the above components may be excluded except a noun. By "noun" we include: noun (dog, money, love etc.), proper noun (Bangkok, Mary etc.), pronoun (you, him, us etc.), noun group (my first job etc.) and gerund (swimming etc.). When discussing Noun Morphology we have to consider many aspects of noun: - number, case, classifiers, person etc.

The Nominative suffix and the Accusative-Dative (Objective) suffix are Φ (null) in English (e.g., my friend). The Nominative suffix is Φ (null) in Bengali as well (e.g., *bondhu* (friend)), but the Objective suffix is *–k*e, as in *bondhu-ke* for animate nouns, and Φ for nonanimate nouns. The *-'*s, and *-'* suffixes mark the genitive forms in English (e.g., boy's, boys'). Bengali has three genitive markers: *–e*r, *–r* and *–ye*r. The genitive suffix is *–er* when the Bengali noun ends in a consonant as in *bon-er* 'sister's', it is *–r* when the noun ends in a matra as in *bondhu-r* 'frined's', and it becomes *–yer* if the noun ends in a vowel as in *boi-yer* 'book's'. The locative suffix is *–e* when the noun ends in a consonant as in *math-e* 'in the ground', it is *–te* when the noun ends in a matra as in *bari-te* 'in the house'. These suffixes are added after the enclitic counting expression (or pluralizer *-gul*o), if any; e.g., *tak-gulo-te* 'on the shelves'.

There are two pluralizers in English: *-s* and *–e*s. Some words in English have two distinct forms in singular and plural, and the plural forms can not be derived from the singular forms (e.g., datum vs. data). In Bengali, *-ra* and *-der* make the nominative and objective/genitive human plural marker respectively, e.g., *bondhu-ra* 'friends', *chele-der* 'boys''. Similarly, *-gulo/guli* makes up a nonhuman plural marker, e.g., *boi-guli* 'books' in nominative and objective case, and *boi-guli-r* 'books'' in genitive case.

An article in English appears at the start of an NP, and there are only three of them (a, an, the). They are handled in Bengali by means of classifiers, e.g., *-jon* is a human classifier. For example, *ek-jon lok* 'a man', *du-jon mohila* 'two women', an*ek-jon lok* 'many people'. The other two important classifiers are the general classifier *ti/ta* and the piece-classifier *khani/khana* which signals non-animate single objects. E.g., *chele-ti* 'the boy', *boi-khani* 'the book'. When the numeral-classifier appears before the noun, even if the noun phrase may refer to a plural entity, it does not have a plural ending, e.g., *anek-jon sadasso* 'many members', *du-ti boi* 'two books'.

An NP in English also includes pronouns and adjectives and in the next two subsections (2.1 and 2.2) we discuss how such words are handled in English and Bengali. Subsection 2.3 includes a snapshot of the phrasal example base for NPs that encode the NP transfer rules.

2.1 Pronouns

Personal pronouns in English vary depending on the number, case and person information (and gender information for third person singular pronoun). Personal pronouns in Bengali also have many forms depending on the number, case, person and formality information, but it does not depend on the gender information. Each personal pronoun takes two distinct forms in singular and plural. Each of these pronouns has three different forms for three cases. Again 2nd person pronouns have three (formal, neutral and intimate) and 3rd person pronouns have 2 honorific variants (formal and neutral). So, translating pronouns from English to Bengali involves anaphora resolution (i.e., what

a personal pronoun refers to and which honorific variant is applicable to it). The different forms of the 2nd person pronouns in Bengali are given in Table 1.

2nd person	Nom (Sg/Pl)	Acc/Dat (Sg/Pl)	Gen (Sg/Pl)
Formal	*apni* / *apnara*	*apnake* / *apnader(ke)*	*qpnar* / *qpnader*
Neutral	*tumi* / *tomra*	*tomake* / *tomader(ke)*	*tomar* / *tomader*
Intimate	*tui* / *tora*	*toke* / *toder(ke)*	*tor* / *toder*

Table 1: 2nd person pronouns in Bengali

2.2 Adjectives

In English, comparative and superlative forms of an adjective are formed in two ways: either suffixing *–er*, *-est* to the adjective if it is a monosyllabic word (e.g., rich, richer, richest), or adding a comparative expression before the adjective (e.g., *more* effective, *most* effective, *less* effective, *least* effective,). Some adjectives exhibit exceptions to these principles (good-better-best; bad-worse-worst). Maintaining a list of such adjectives is the way to identify them.

The three adjectives forms (positive, comparative and superlative) are handled in Bengali in a similar way. Comparative and superlative adjectives are formed by suffixing *–tara* and *–tama* respectively to the positive adjective, e.g., *priyo* 'dear', *priyotara* 'dearer' and *priyotama* 'dearest'. Sometimes, the adjective is left as it is and a comparative expression is placed before the adjective (as in English *more*, *most*, *less*, *least* are placed before the adjective) to make it a comparative or superlative one, e.g., *bhalo* 'good', *beshi/aro bhalo* 'better' (= more good), *sabcheye bhalo* 'best' (= most good).

2.3 A Snapshot of the NP Phrasal Example Base

Some examples of transfer rules are given below for NPs: (n represents a noun and n′ its Bengali counterpart).

<art & a > <n & sng, human, nom> ←→ <*ekjon*> <n′ >

<art & a > <adj> <n & sng, inanimate)> ←→ <*ekti*> <adj′> < n′>

<prn & gen> <n & plr, human, nom> ←→ <prn′> <n′ > <- *ra*>

Using these transfer rules, we can translate the following NPs.

<art & a > <n & boy (sng, human, nom)> ←→ <*ekjon*> <*chele*>

<art & a > <n & book (sng, inanimate, acc)> ←→ <*ekti*> <*boi*>

<prn & my (gen)> <n & friends (plr, human, nom)> ←→ <*amar*> <*bondhura*>

<n & Ram's (sng, gen)> <n & friends (plr, human, dat)> ←→ <*ramer*> <*bondhuderke*>

3 English VPs and their Bengali Translation

English verbs have to agree with the subject in person and number information. But in contrast, Bengali verbs have to agree with the subject in person and formality. Thus 'you went' has singular */apni gelen, tumi gele, tui geli*/, plural */apnara gelen, tomra gele, tora geli*/. Similarly, 'he came' has meaning /*tini elen, se elo*/. English verb phrases are formed by placing one or more auxiliary verbs in front of the root verb, then either suffixing the root verb or taking any of the three forms (present, past, past participle). But in Bengali, verb phrases are formed by appending appropriate suffixes to the root verb. Again some verbs in English are translated in Bengali using a combination of a semantically "light" verb and another meaning unit (a noun, generally) to convey the appropriate meaning. In English to Bengali context such happenings are very common, e.g., to swim – *santar* (swimming) *kata* (cut), to try – *chesta* (try) *kara* (do).

Bengali verbs are morphologically very rich. A single verb has many morphological variants. The Bengali representation of the 'be' verb is formed by suffixing to the present root *ach*, past root *chil* and the future root *thakb* for appropriate tense and person information. The negative form of the 'be' verb in present tense is *nei* for any person information. And in past and future tense, it is formed by simply adding the word *na* postpositionally after their corresponding assertive form.

Root verbs in Bengali can be classified into different groups according to the spelling pattern. Some examples of spelling patterns are *Ca, Caa, Ci, Cu, CaaC, CuC, CaCCaa/CaaCCaa/CeCCaa, CiCCaa* etc. The symbol *C* represents a consonant, symbol *CC* represents a conjunct and the letters in lower case represent *matra* (vowel). '*ja*'('to go'), '*kha*'('to eat') have the spelling pattern *Caa*; '*di*' ('to give'), '*ni*' ('to take') have the spelling pattern *Ci*; whereas '*kar*' ('to do'), '*has*' ('to laugh'),

'*khel*' ('to play') fall into the group that has the spelling pattern *CaC/CaaC/CeC*. All the verbs belonging to the same spelling pattern category, take the same suffix for same person, tense, aspect information. These suffixes also change from the classical to colloquial form of Bengali. There are separate morphological paradigm suffix tables for the verb stems that have the same spelling pattern. there are some exceptions to these rules. Further details can be seen in (Naskar and Bandyopadhyay, 2004).

The negative forms for the perfect aspect are handled by adding Simple Present suffixes to verb stem and then adding *ni* postpositionally to it. For all other cases, *na* is added postpositionally to the suffixed stem for the appropriate tense and aspect.

The other verb forms like gerund-participle, dependent gerund, conjunctive participle, infinitive-participle, conditional participle are also handled in the same way by adding appropriate suffixes from a suffix table.

When subjects differ in number, or person, or both are connected by *and,* the subject of the verb is treated as a first person, if one of the subjects is of the first person; as a second person, if one of the subjects is of that person, and none of the first; and else as a third person. The Bengali verb is generated accordingly.

Bengali does not have an inflected passive. The analytical or periphrastic passive is usually formed by the verbal noun representation of the verb stem in Bengali and the appropriate form of the auxiliary verb *ja* 'to go'. The *ja*-passive forms an impersonal construction, e.g., with the verb 'to see': *dekha jay* 'it is seen'. If a promoted object is used, it is in the Dative: *ama-ke dekha jay* 'I am seen' (= to me is seen). A few cases of idiomatic passive formation take place with the verb root *pad* 'to fall', *cal* 'to go', *hoy* 'to happen' etc instead of *ja*: *mara pode* 'gets killed'; *mara gelo* 'got killed', *khawa chole* 'can be eaten' *amake dekha hoy* 'I am seen' (=They see me).

4 English PPs and their Bengali Translations

A preposition is a word placed ("pre-positioned") before a "noun" to show in what relation the noun stands in regard to something else. A PP is a group of words starting with a *preposition*, and containing an *object of the preposition*, and any *modifiers of the object*. A PP can take the role of a NP or an adverbial phrase also. In Bengali, there is no concept of preposition. English prepositions are handled in Bengali adding inflexions to the reference objects and post-positional words after them. Inflexions get attached to the reference objects and relate it with the main verb of the sentence in *case* or *karaka* relations. An inflexion has no existence of its own in the language, and it does not have any meaning as well. There are only a few inflexions in Bengali: Φ (null), -e, -y, -*y*e, -*t*e, -*et*e, -*k*e, -*r*e, -*er*e, -*r* and -*e*r. On the other hand, post-positional words are independent words. They have meanings of their own and are used independently like other words. A post-positional word is positioned after an inflected noun (the reference object) and relates it with the main verb of the sentence in *Case* or *Karaka* relations. Some examples of the post-positional words in (colloquial) Bengali are *diye* (by)*, theke* (from)*, jonno* (for)*, kachhe*(near)*, samne* (in front of) etc.

As far as the selection of the appropriate target language representation of a preposition is concerned the reference object plays a major role in determining the correct preposition sense. Deciding whether the preposition is used in a spatial sense, as opposed to a temporal or other senses, is determined by the semantics of the head noun of the reference object. A noun phrase (NP) denoting a place gives rise to a spatial PP. Similarly, an object referring to a time entity produces a temporal expression.

When an English PP is translated into Bengali, the following transformation takes place: (preposition) (reference object) (reference object) [(inflexion)] [(postpositional-word)]. The correspondence between English prepositions and corresponding Bengali inflexions and/or postpositions is not direct, they have a many-to-many relationship. Any of the three Bengali inflexions '*-e*', '*-te*' and '*-y*' can be placed after the reference object for any of the spatial and temporal prepositions *in*, *on* and *at*. The rule is: if the last letter of the Bengali representation of the reference object is a consonant, '*-e*' is added to it, else if the last letter of the Bengali word is a consonant with a *matra* (vowel marker) and if the *matra* is '*-a*', any of '*-te*', or '*-y*' can be added to the Bengali reference word, otherwise '*-te*' is added to it. When translating the temporal expressions, some additional rules are applied, e.g., if '*on*' is followed by a day (like Sunday, Monday etc.) or a date in English, no postpositional marker is added.

To translate prepositional phrases with prepositions other than *in*, *on*, *at*, we take the help of an example base, which contains the necessary transfer rules. Here are some examples from the example base (TLR – target language representation of the reference object).

from (place/physical object/time) ←→ (TLR) - *theke*

before (place/physical object) ←→ (TLR) - (*r* / *er* / *yer*) *samne*

before (time) ←→ (TLR) - (*r* / *er* / *yer*) *age*

near (place/physical object) ←→ (TLR) - (*r* / *er* / *yer*) *kachhe*

round (place/physical object) ←→ (TLR) - (*r* / *er* / *yer*) *chaardike*

towards (place/physical object) ←→ (TLR) - (*r* / *er* / *yer*) *dike*

without (object) ←→ (TLR) - *chharha*

after (time) ←→ (TLR) - (*r* / *er* / *yer*) *pare*

since (place/physical object/time) ←→ (TLR) *theke*

till (time/place) ←→ (TLR) *porjonto*

within (time) ←→ (TLR) - (*r* / *er* / *yer*) *modhdhe*

with (instrument) ←→ (TLR) *diye*

behind (place/physical object) ←→ (TLR) - (*r* / *er* / *yer*) *pichhone*

The semantic information about the reference objects are accumulated from the Wordnet. In some cases, translation of PPs does not require knowledge of semantic features of the reference objects. For example, *since* (as a preposition) is always translated as *theke* in Bengali, irrespective of the reference object. Again in some cases, these semantic information about the reference object do not suffice to translate the preposition properly. Given below are some PPs that are translated using the transfer rules:

<prep & with> <prn & his (gen)> <n & friends (plr, human, nom)> ←→ <*tar*> <*bondhuder*> <*sathe*>
(corresponding transfer rule: <prep & with> <prn & gen> <n & plr, human> ←→ <prn´> <n´> <- *der*> <*sathe*>)

<prep & with> <prn & him (dat)> ←→ <*tar*> <*sathe*>
(corresponding transfer rule: <prep & with> <prn & dat> ←→ <prn´ & gen> <*sathe*>)

<prep & in> <n & school (sng, inanimate, loc)> ←→ <*bidyalaye*>
(corresponding transfer rule: <prep $ in / at / on> <n & sng, inanimate, loc)> ←→ <n´> - < *e* / *te* / *y* >)

<prep & inspite_of> <art & the> <n & incident (sng, inanimate, event) > ←→ <*ghatanati*> <*swatteo*>
(corresponding transfer rule: <prep & inspite_of> <art & the> <n & sng > ←→ <n´> < - *ti*> <*swatteo*>)

5 Phrase ordering in English and Bengali

There are basic differences in phrase ordering in English and Bengali. English is a fixed phrase order language with SVO pattern whereas Bengali is a relatively free phrase order language. The most acceptable pattern in Bengali is SOV. Some salient points in phrase ordering rules in English and Bengali are identified below.

(i) English is a SVO language and an English sentence has the [NP (Subj)] [VP] {[NP (ind. Obj.)]} {[NP (dir. Obj.)]} construction. Whereas, Bengali is a SOV language (like other Indian languages) and a Bengali sentence has the [NP (Subj)] {[NP (ind. Obj.)]} {[NP (dir. Obj.)]} [VP] construction (the braces indicate that the component is optional). But, Bengali is a relatively free phrase-order language. So, not obeying to the usual pattern does not necessarily lead to an incorrect translation.

(ii) In English, if there are any complement(s) of the subject (or object or verb), it usually sits after the subject (or object or verb). But, in Bengali, the complements appear after it. E.g., '<The Prime Minister>[1] <of Pakistan>[2]' ←→ <*Pakistan-er*>[2] *pradhanmontri*[1] .

(iii) If the verb has two complements – a temporal and a spatial, the temporal complement usually appears before the spatial complement in Bengali. E.g., '<He>[1] <came>[2] <to our house>[3] <yesterday>[4]' ←→ *se*[1] *gatakal*[4] <*amader badi-te*>[3] *esechilo*[2].

(iv) The subordinate clause usually sits before the principal clause in Bengali.

(v) Relative Clauses (RC) are formed in English by relative pronouns and adverbs (wh-words) and they sit after the referent. In Bengali, RCs are formed with relative pronouns *(j*e, *jini*), adverbs *(jekhane*), interjections *(jodi*) and their co-relatives *(se*, *tini*, *sekhane*, *tahale*) both at the initial and the final position. When translating a complex English sentence having a RC, the relative pronoun is followed by the referent in the relative clause, and the co-relative is placed in the principal clause. E.g., '<The boy>[1] <who>[2] <was playing>[3] <football>[4] <is>[5] <Ram's brother>[6]' ←→ '*je*[2]

<chele-ta>[1] *phutbal*[4] *khelchilo*[3] *se*[2] *<ram-er bhai>*[6]'.

(vi) If the prepositional phrase in English has more than one noun, suffix is added only to the last noun in Bengali, e.g., '<The people>[1] <of India and Pakistan>[2]' ←→ '*<bharat ebong pakistan-er>*[2] *lokra*[1]'.

The heuristic rules used for recombination of the translated Bengali phrases have been developed based on the phrase ordering rules as described.

6 System Description

The translation methodology incorporated in the system, is to identify the constituent phrases of the source language input sentence, translate the phrases individually to the target language and finally arranging the translated phrases, using some heuristics, to form the target language equivalent of the source language sentence.

A shallow parser identifies the constituent phrases of the source language sentence and tags the phrases to encode all relevant information. A phrasal Example Base is used to retrieve the target language phrase structure corresponding to each input phrase. Then these phrases are translated individually to the target language (Bengali) using Bengali synthesis rules. Finally, those target language phrases are arranged using some heuristics, based on the word ordering rules of Bengali, to form the target language representation of the source language sentence. The noun phrases (NP) and the prepositional phrases (PP) are translated using syntactic example bases that contain translation examples. Thus the system takes somewhat hybrid approach. These PPs can take the role of nouns and adverbs in certain cases. Verb phrase (VP) translation scheme is *rule based* and uses Morphological Paradigm Suffix Tables. The next three subsections 6.1 to 6.3 discusses about the different modules of the system. An example English sentence is translated to Bengali in subsection 6.4 following the approach in the system. Sense disambiguation in the system is described in subsection 6.5.

6.1 Morphological Analysis

In order to translate from English to Bengali, the first thing we do is lexical analysis of the English sentence to gather the lexical features of the morphemes. During morphological analysis, the root words / terms (including idioms, named entities, phrasal adjectives, phrasal prepositions), along with associated grammatical information and semantic categories are extracted. Wordnet 2.0 (Fellbaum, 1998) is the main resource that we have used for performing the Morphological analysis of the Source Language (SL – English) input sentence. Wordnet contains four types of lexical terms – Noun (79689), Verb (13508), Adjective (18563) and Adverb (3664). The other types of lexical terms (article, pronoun, preposition, conjunctions) are stored in different files. At first, the tokens (words/abbreviations/acronyms/proper nouns) are identified in the input sentence. The tokens are then searched in the Wordnet and in other lexicons for their syntactic categories. The tokens are morphologically analyzed for suffixes. We have used the suffixes provided by the Wordnet. The other irregular suffixed words are considered exceptions and are kept in *exception files*. Each word is searched for all its syntactic categories and is assigned all the possible categories.

Multiword expressions or terms are identified in this module and are treated as a single token. These include phrase prepositions (e.g. "in spite of", "by dint of"), phrase adjectives (e.g. "a lot of", 'a flock of'), idioms ("bag and baggage") etc. Sequences of digits and certain types of numerical expressions, such as dates (June 23) and times (9 am), money expressions ($3 to $4 million), and percents (31.5%) are also treated as a single token. They can also appear in different forms (e.g., 'November 18, 1989', 'Nov. 18, 1989', '18th November, and '18/11/1989') as any number of variations. *Wordnet* contains a lot of multiword expressions including noun – noun compounds (e.g., federal bureau of investigation, graphical user interface etc.), and phrasal verbs. In addition to that, we maintain a list of idiomatic expressions. We also have a list of proper names (more than 10,000 entries) that contains names of person, location and organization.

6.2 Parsing

Currently we have a rule-based shallow parser, which takes the words / terms together with all relevant information and depending on rules assigns a word its correct part of speech (POS) tag. Then it identifies the constituent phrases of the sentence depending on a grammar. The phrases are tagged during extraction. The tags include all the necessary information that might be needed to translate this phrase and perhaps resolve ambiguities in other phrases.

To improve performance of the parser, we are adding a statistical component to the Tagging module using the Hidden Markov Model (HMM). For this purpose, we have computed the word frequencies and the n-gram tag-probabilities from the POS-Tagged Brown Corpus. From these two data, we can assign a word its most probable tag.

6.3 Example Base

Examples are stored in the Example base in three ways: (a) literal examples; (b) pattern examples with variables instead of words; and (c) grammar examples expressed as context-sensitive rewrite rules, using semantic features. There is a hybrid nature in this approach, where the type (a) examples are pure strings, type(c) are effectively transfer rules of the traditional kind, with type (b) half-way between the two (Furuse and Iida, 1992). In the present work, the system makes use of all these three types of example bases.

The tables that contain the proper nouns, acronyms, abbreviations and figure of speech in English and the corresponding Bengali translation are example bases of type (a). The phrasal templates (translation examples) for the noun and the prepositional phrases store the part of speech of the constituent words along with necessary semantic information. The source and the target phrasal templates are stored in example bases of type (c). These translation examples are effectively transfer rules that are stored in the example base instead of explicit coding. This adds flexibility – new rules can be added and existing rules can be modified easily without changing or adding a single line of code. The idiom table contains pattern examples with variables instead of words (e.g., "make up <possessive pronoun> mind").

6.4 A Sample Translation

Let us consider the candidate English sentence: "*A man had given the boy a good book in the school*." The shallow parser identifies the phrases as: "<np1> *A man* </np1> <vp1> *had given* </vp1> <np2> *the boy* </np2> <np3> *a good book* </np3> <pp1> *in the school* </pp1>." The phrase translator module translates the phrases as: 'A man' ←→ '*ekjon lok*', '*had given*' ←→ '*diyechhilo*', '*the boy*' ←→ '*chheletike*', '*a good book*' ←→ '*ekti bhalo boi*' and '*in the school*' ←→ '*bidyalaye*'. Then the translated Bengali phrases are reordered using heuristics that are based on phrase ordering rules of Bengali. The system generates the translated Bengali sentence as: "*<ekjon lok>*[1] *<chheletike>*[3] *<bidyalaye>*[5] *<ekti bhalo boi>*[4] *< diyechhilo >*[2]".

6.5 Sense Disambiguation

Sense disambiguation has been attempted for nouns and verbs. To identify the proper sense, we have utilized selectional preference information of the verb on the head nouns of the subject and object. Reports on an earlier work on sense disambiguation in the system using selectional restriction information can be found in (Naskar and Bandyopadhyay, 2004). The classifications (Synsets) and Frame information in Wordnet are used intelligently to disambiguate the noun and verb senses. We have computed Selectional Preference information for each verb synset from the sense tagged Brown Corpus. For a particular verb sense, if the head nouns of the subject and object (if any) of the sentence in hand do not match with the selectional preference information for that particular verb sense, we discard the possibility of that verb sense. For each verb synset, in addition to the frame structure (as in wordnet), we have assigned to it information (synset) about what it prefers as a subject or object (if it takes object at all). This selectional preference information together with synset information of the participating noun(s) and verb are used to disambiguate both polysemous nouns and verbs.

For disambiguating verb sense, the frame information of the root verb of the verb phrase is used first. Sometimes, this frame information alone suffices to disambiguate the verb sense. If the frame information alone cannot disambiguate the verb sense, we employ the selectional preference knowledge of the verb. Although the frame information alone may not disambiguate the verb sense, it often reduces the degree of ambiguity in polysemous verbs. Say, the verb (say v) has m (m>1) senses, and among these m senses, frame information of only n (n>0) senses match with that of the sentence. If n = 1, the verb is disambiguated. Now if n>1, we proceed further. In this case we apply the selectional preference information of the verb. Say, the head noun (say n1) of the subject has p senses and that (say n2) of the object has q senses. So, there are (p*n*q) unique possibilities. Among these (p*n*q) possibilities, we choose i^{th} sense of n1, j^{th} sense of v and k^{th} sense of n2 that maximizes the probability of occurrence of n1 at the subject position, v at the verb position and n2 at the object position.

7 Proposed Evaluation Scheme

The most widely used MT evaluation schemes like BLEU (Papineni et. al., 2002) are based on the viewpoint that *"The closer a machine translation is to a professional human translation, the better it i*s". These evaluation schemes basically work at the blackbox level, i.e., performances of the component modules of the MT system are not evaluated separately. We are currently in the process of developing a glassbox evaluation scheme for our MT system. The MT system has been conceptualized as consisting of the following three component modules: the shallow parser that identifies the NP or VP or PP, the retrieval of the target language phrase structure using the phrasal example base and finally the phrase rearrangement

module that rearranges the synthesized target language output. Any error at an earlier stage would propagate to the succeeding stages with much higher weight and the system would have to pay a big penalty in the overall performance for a small error at an earlier stage. We are working on a glassbox evaluation scheme that will evaluate the performance of each individual module. These weighted evaluation metrics will be combined linearly to generate the evaluation metric for the whole system. This would pinpoint where the system falls back, so that we can improve the low scoring modules and thus evolve the system for better performance.

8 Conclusion

Divergence is a key aspect of machine translation between two languages. Divergence assumes special significance in the domain of Example-Based Machine Translation (EBMT). In the present system, we have considered the structural divergence between English and Bengali. We have observed that categorial divergence also exists between English and Bengali and its handling would have improved the quality of the output. The sense disambiguation module that we have incorporated in the system can disambiguate senses of nouns in subject and object position, but can not tackle nouns in PPs.

9 Acknowledgements

Our thanks go to Council of Scientific & Industrial Research, Human Resource Development Group, New Delhi, India for supporting Sudip Kumar Naskar under Senior Research Fellowship Award (9/96(402)2003-EMR-I).

References

S. Naskar and S. Bandyopadhyay. 2004. *Translation of Verb Phrases from English to Bengali*, In "Proceedings of the CODIS 2004", pages 582-585, Kolkata, India.

S. Naskar and S. Bandyopadhyay. 2004. *Sense Disambiguation Scheme Using Selectional Restrictions in Anuvaad – An English-Bangla MT System.* In "Proceedings of iSTRANS-2004", pages 214-216, New Delhi, India.

Michael Carl and Andy Way. 2003. *Recent Advances in Example-Based Machine Translation*. Kluwer Academic Publishers. Dordrecht.

Dien Dinh, Nguyen Luu Thuy Ngan, Do Xuan Quang and Van Chi Nam. 2003. *A Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translatio*n. In "Proceedings of MT Summit IX", New Orleans, Louisiana, USA.

R.M.K. Sinha and A. Jain. 2003. *AnglaHindi: An* English to Hindi Machine-Aided Translation *Syste*m. In "Proceedings of MT Summit IX", New Orleans, Louisiana, USA.

Deepa Gupta and Niladri Chatterjee. 2003. *Identification of Divergence for English to Hindi EBM*T. In "Proceedings of MT Summit IX", New Orleans, Louisiana, USA.

Kishore Papineni, Salim Roulos, Todd Ward and Wei-Jing Zhu. 2002. *BLEU: a Method for Automatic Evaluation of Machine Translatio*n, in the *proceedings of the 40th Annual meeting of the Association for Computational Linguistics (ACL-2002*). pages 311-318, Philadelphia.

Harold Somers. 2000. *Example Based Machine Translation*. In "Handbook of Natural Language Processing", Robert Dale, Hermann Moisl, Harold Somers, ed., pages 611-627, Marcel Dekker, Inc., New York.

C. Fellbaum. ed. 1998. *WordNet – An Electronic Lexical Database,* MIT Press, Cambridge, MA.

Furuse O, H Iida. 1992. *An example-based method for transfer-driven machine translation*. In "Isabelle P, ed. Proceedings of TMI-92", pages 139-150, Monteral, Canada.

Nagao M. 1984. *A framework of a mechanical translation between Japanese and English by analogy principle*. In "Artificial and Human Intelligence", A Elithorn, R Banerji, ed., pages 173-180, Amsterdam, North-Holland.