A Semantics-based English-Bengali EBMT System

for translating News Headlines

Diganta Saha
Computer Science and Engineering
Department
Jadavpur University
Kolkata, India, 700032
neruda0101@yahoo.com

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

Abstract

The paper reports an Example based Machine Translation System for translating News Headlines from English to Bengali. The input headline is initially searched in the Direct Example Base. If it cannot be found, the input headline is tagged and the tagged headline is searched in the Generalized Tagged Example Base. If a match is obtained, the tagged headline in Bengali is retrieved from the example base, the output Bengali headline is generated after retrieving the Bengali equivalents of the English words from appropriate dictionaries and then applying relevant synthesis rules for generating the Bengali surface level words. If some named entities and acronyms are not present in the dictionary, transliteration scheme is applied for obtaining the Bengali equivalent. If a match is not found, the tagged input headline is analysed to identify the constituent phrase(s). The target translation is generated using English-Bengali phrasal example base, appropriate dictionaries and a set of heuristics for Bengali phrase reordering. If the headline still cannot be translated using example base strategy, a heuristic translation strategy will be applied. Any new input tagged headline along with its translation by the user will be inserted in the tagged Example base after generalization.

1   Introduction

The present work aims to develop a methodology for a semantics-based Example Based Machine Translation (EBMT) system for translating news headlines from English to Indian languages. The methodology is being deployed to implement a machine translation system for translating news headlines from English to Bengali, a major Indian language and the fifth language in the world in terms of the number of native speakers. It is the official language of Bangladesh. The reason for choosing English as the source language is that most news are generated in English, even in India, and the vernacular dailies carry out a translation before publishing them. The semantic and syntactic classification schemes developed for English news headlines may be useful for building news headline machine translation systems from English to other languages.

Most of the International and National news wire service agencies send news items in English. Manual translation of these news items into any other language is slow and tedious. The inflow of news items is not evenly distributed, therefore there is burst of translation required just before the newspaper is to go out. The domain of news items has attracted the attention of Machine Translation (MT) researchers all over the world. The internet editions of newspapers in English and regional languages are now a reality.

Translation of news headlines plays a crucial role in the translation of a news item. The headline is an important component in a news item. The headline must be informative, i.e., it should indicate sufficiently about the content of the news item. At the same time it must attract the attention of the reader, i.e., it must have its own style. The informative property of the news headlines must be retained as far as possible while translating into the target language. Each language has its own style of writing headlines. The style of the source language news headline can be preserved by assigning semantic tags to the words in the news headline in addition to the syntactic tags. The style of the news headline in the target language can be maintained by developing a parallel example base of news headlines in source and target languages, assigning semantic as well as the syntactic tags to both sides for generalizing the paralle example base, aligning the entries and then following an example based machine translation strategy. A direct parallel example base of news headlines may be necessary for those headline pairs which are unique in nature

and thus cannot be generalized. The system described in the present work follows this strategy.

News headlines are generally not grammatical sentences in nature. They can be or can consist of root word(s), surface level word(s), named entity(ies) (person names, location names, organization names, and miscellaneous e.g., temporal expressions, monetary expressions, cinema names, book names, hotel names, train names), acronym(s), noun phrase(s), sentence without an auxiliary verb, quotation or a grammatical sentence. The syntactic structure of the news headlines suggests that while they cannot always be defined by the sentence level grammar formalisms, news headlines follow a sublanguage of its own. Thus, the Rule based machine translation strategy that uses sentence level grammar formalisms is not suitable for the translation of news headlines. If the input news headline cannot be translated using either the direct or the generalized example base, the tagged input headline may be analysed to identify the constituent phrase(s). The target translation is then generated using the parallel phrasal example base, appropriate dictionaries and a set of heuristics for target language phrase reordering. This rule based machine translation strategy has been followed in the present work.

In India, most English newspapers have their vernacular publication but the layout of news and their headlines are not parallel, i.e., not exact translation of each other. Thus corresponding news headlines in English and vernacular editions cannot be directly used to create a large parallel example base of English-Vernacular news headlines. The two machine translation systems for translating English news headlines to Hindi (Sinha, 2002; Rao et. al., 2000) do not have a large parallel example base of English-Hindi news headlines. Thus Statistical machine translation (SMT) system is also not suitable for machine translation of English news headlines to Indian languages. In this work, we are creating the tagged parallel example base of news headlines with the help of English and Bengali newspapers of the same date. The present system generalizes the tagged English news headlines. The corresponding set of tagged Bengali news headlines may not be identical. The system displays the possible generalized tagged Bengali news headlines and the developer chooses one of them. The chosen target language news headline may be edited to maintain the informative nature and the style. This collection of parallel example base is not large enough to attempt SMT. In view of these, it has been considered that EBMT strategy is most suitable for translation of news headlines. The EBMT strategy also allows the

system to integrate different resources, namely, Direct example base, Generalized tagged example base and the Phrasal example base which are discussed later.

Related works on machine translation of news headlines in India as well as elsewhere in the world are discussed in section 2. Semantic and syntactic classification of news headlines have been outlined in section 3 and 4 respectively. Tag set definition and tagging of English and Bengali news headlines are discussed in section 5. Creation of generalized tagged example base of English and Bengali news headlines are described in sections 6 and 7 respectively. Section 8 describes the different example bases in the system, specifically the Phrasal example base. The dictionary design is outlined in section 9. MT system development methodology is described in section 10 and the conclusion is drawn in section 11.

2   Related Works

In India, a heuristic approach for translating news headings from English to Hindi is found in (Sinha, 2002). A human-aided MT system for translating English news texts to Hindi is being developed at the Centre for Development of Advanced Computing, Mumbai (Rao et al, 2000). The system is now being enhanced and adopted for web translation service to the news agencies. A hybrid system for translating news items from English to Bengali (Naskar & Bandyopadhyay, 2005; Bandyopadhyay & Saha, 2002; Bandyopadhyay, 2000a, 2000b) is being developed at the Jadavpur University, India.

The NHK System of Japan which translates English newspaper articles to Japanese is described in (Hutchins, 1999). The improvement of translation quality of English newspaper headlines by automatic pre-editing in the English to Japanese machine translation system being developed at the Sharp Corporation of Japan is discussed in (Yoshimi, 2001). The work focuses on the absence of the verb *be* and formulates a set of rewriting rules for putting the verb properly into headlines, based on information obtained by morpholexical and rough syntactic analysis. The improvement of translation style and the target words of English news headlines by identifying and resolving the coreference of acronyms, abbreviations and proper names in the English to Japanese machine translation system being developed at the Toshiba Corporation of Japan is discussed in (Ono, 2003).

3   Semantic Classification of News Headlines

News items in a news paper generally follow a classification on the basis of geographical

hierarchy (Metro -> State -> Country -> World) as well as a separate topic based one (Sports, Business etc.). Though there does not exist any standard classification of news items in the journalistic world, we conducted a study on six English news papers that are published from Kolkata. It has been observed that all of them follow the same geographic classification as well as a topic based one, though the names of the classes are not always identical. Named entities and acronyms occur in very large number in news items as well as in the associated news headlines and these named entities and acronyms tend to cluster around each class in the classification scheme. The classification of the news items as well as the associated headline, on the basis of the content of the news has been termed as the **Semantic Classification**. Having separate bilingual named entity and acronym dictionaries under each semantic class help in the disambiguation of these words also. In the present work, English news headlines have been semantically classified as follows: *Front Page, World, India, Bengal, Kolkata, Business and Sports*. The news in the *Front Page* include the important news events for the day that may belong to any category. There are further classification like *Editorial, Perspective, Cinema, Entertainment or Campus* whose contents are mainly feature based. Headlines for these items have not been considered in the present work. The Bengali news papers published from Kolkata carry more news from the state and hence they follow a more detailed classification on *Bengal*. In the present work, we have followed identical classification schemes for both English and Bengali news headlines.

News items can be further classified into the following two categories on the basis of the number of paragraphs in the news item: (i) Short Single Paragraph News Items and (ii) Long Multi-paragraph News Items as they follow distinct styles. Long multi-paragraph news items are more informative in nature. Headlines for both these types of news items also differ in their style and informative nature. Headlines for short single paragraph news items are generally one-, two- or three words long; may be a named entity, compound noun or noun phrase and occassionally may be sentences. Long multi-paragraph news items may include two separate headlines. Sometimes, within the bodies of these news items short news along with a separate headline are found, either originating from the same place as the main news or dealing with a related topic. Apart from these syntactic differences which are discussed in the next section, headlines from these

two categories of news are also different on their information content. For example, the headlines *Tea Strike* and *Garden workers go on an indefinite strike for pay hike / Trouble brews in tea estates* correspond to the short and long versions of the same news event. It may be noted that there are two headlines for the long news. On the basis of these observations, the following semantic classes *Front Page, World, India, Bengal, Kolkata, Business and Sports* have been further divided into *short* and *long* classifications. The example news headlines for the various semantic classes are shown in Table 1:

Semantic Class	Example News Headline
Front Page	Snaps say error camps exist: Natwar
Front Page – short	FB threat
World	Rice no-show invites criticism
World-short	Van Gogh trial
India	PM assures left on eve of US trip
India-short	Ex-servicemen
Bengal	Bandh to protest against blasts
Bengal-short	SFI clash
Kolkata-	More courses at Presidency
Kolkata-short	Train services hit
Business	Assam Tea workers want basic pay revised Productivity-linked wages rejected by ACMS
Business-short	Microsoft
Sports	ICC says 2004-05 was corruption-free
Sports-short	Selections

Table 1: Semantic Classification of News Headlines

4   Syntactic Classification of News Headlines

News headlines for short single paragraph news and those for the long multi-paragraph news show different syntactic structures. Headlines for short single paragraph news items can be classified at the top level on the basis of the number of words they contain, viz., one-, two-, three- or more than

three words. Similarly, headlines for long multi-paragraph news items can be classified at the top level on the basis of the number of words they contain, viz., three- or more than three words. Such headlines are generally grammatical sentences in nature. A headline may consist of two sentences, also. Sometimes, within the body of these news items short news are found, either originating from the same place as the main news or dealing with a related topic. It has been observed that the structure of these news headlines follow the same for the short single paragraph news headlines.

The *single word headlines* can be a root word, named entity, acronyms or a surface level word. *Two word headlines* can be a compound noun, sentence without an auxiliary verb, grammatical sentence or a collocation. *Three word headlines* can be a noun phrase, compound noun, sentence without an auxiliary verb or a grammatical sentence. *Headlines with more than three words* can be a noun phrase, sentence without an auxiliary verb, grammatical sentence or quotation.

On the basis of above observations, the news headlines have been syntactically classified at the top level on the basis of the number of words, viz., one-, two-, three- and more than three words. Since, three words or more than three words headlines appear for both short and long news items, each of the 7 semantic classes have been synactically classified at the top level further into 4 classes as above. In the present work, a total of 28 parallel example bases have been designed. Further syntactic classification (i.e., root word, named entity, acronym, surface word, compound noun, collocation, noun phrase, sentence without an auxiliary verb, grammatical sentence, quotation) is included as an attribute of the example news headline. This organization of the example bases makes it easier to identify the appropriate example base for an input news headline, where it is most likely to be present, given its semantic class and the number of words present in it. The syntactic classification provides appropriate information for alignment of the tagged source and target language news headlines. The syntactic class of the input news headline helps in the application of the appropriate rule based translation strategy when it cannot be translated using the example based translation methods. The example news headlines for the various syntactic classes are shown in Table 2.

Some headlines are elliptical in nature. An example of ellipsis is in the headline *Train kills 1* where the number *1* is not explicitly qualified but the implicit qualification is *person*. The elliptical resolution in this case is not necessary for translating it to Bengali as the ellipsis is retained in

Bengali. Another example of an elliptical news headline is *Bhajji claims a couple*. In this case, ellipsis resolution is necessary for translation. Since, the news headline is for a sports news in which *claiming a couple* means *claiming a couple of wickets*, the headline will be extended as *Bhajji claims a couple of wickets* and then translated.

Syntactic Class	Example News Headline
One word	• Accident (root word) • Kirloskar (named entity) • HDFC (acronym) • Selections (surface word)
Two words	• Flight problem (compound noun) • Buddha's gesture (compound noun) • RBI report (compound noun) • Kanika critical (sentence without an auxiliary verb) • India wins (grammatical sentence) • Pulse Polio (collocation)
Three words	• Woods on top (noun phrase) • Shastri Bhavan fire (compound noun) • Tour de France (compound noun) • Sania No. 70 (sentence without • an auxiliary verb) • Train services hit (sentence without an auxiliary verb) • Australia wins again (grammatical sentence)
More than three words	• Breakthrough in diagnosing HIV (noun phrase) • Rail contracts under cloud (sentence without an auxiliary verb) • Family health drive enters fifth round (grammatical sentence) • Intelligence couldn't have prevented attack: Blair (quotation)

Table 2: Syntactic Classification of News Headlines

5 Tag set definition and Tagging of News Headlines

The present system is being developed for translating English news headlines to Bengali. Since, parallel example base of English and Bengali headlines is not available, we started with the collection of English news headlines under different semantic and syntactic classes. The headlines in English and Bengali are tagged with a set of syntactic and semantic tags.

5.1 Tag Sets

Noun and verb words are tagged with the corresponding Wordnet Lexicographer file names.

The following are some example tags used for noun words:

noun.act: noun denoting acts or actions,

noun.animal: nouns denoting animals,

noun.artifact: noun denoting man-made objects,

noun.body: noun denoting body parts,

noun.event: noun denoting natural events,

noun.food: noun denoting foods and drinks,

noun.group: noun denoting grouping of people or objects,

noun.location: noun denoting spatial position,

noun.person: noun denoting people,

noun.time: noun denoting time and temporal relations.

It may be noted that when the noun words in the last four types identify a specific object they denote a named entity and are appropriately tagged.

The following are some example tags used for verb words:

verb.change: verbs of change of size, temperature, intensity, etc.,

verb.cognition: verbs of thinking, judging, analyzing, doubting, etc.,

verb.communication: verbs of telling, asking, ordering, singing, etc.,

verb.competition: verbs of fighting, athletic activities, etc.,

verb.consumption: verbs of eating and drinking,

verb.contact: verbs of touching, hitting, tying, digging, etc.,

verb.creation: verbs of sewing, baking, painting, performing, etc.,

verb.motion: verbs of walking, flying, swimming, etc.,

verb.possession: verbs of buying, selling, owning and transfer,

verb.social: verbs of political and social activities and events,

verb.weather: verbs of raining, snowing, thawing, thundering, etc..

The tagging of the verb words in the headlines helps to identify the source and the target language verb patterns (Kim et. al., 2002). Each verb can have several meanings and each meaning of a verb is represented by a verb pattern. A verb pattern consists of a source language pattern part for the analysis and the corresponding target language pattern part for the generation. The meaning of a verb can be identified using the associated noun and the adjective words. For example, the verb *kill* is tagged as *verb.contact*. The associated noun words for one meaning of the verb are *accident, attack* etc. and the adjective word *dead* is associated with the same meaning of the verb.

Named entities are further tagged as *Person Name*, *Location Name*, *Organization Name* and *Miscellaneous* e.g. *temporal expressions, monetary expressions, cinema names, book names, hotel names, train names* etc.. Strictly speaking, further tagging of named entities are not necessary for headline translation except in tagging of *person names* and *organization names* and that too, when the headline includes a verb word. The verb form in Bengali depends on the associated named entity. Words of other parts of speech (*adjective, adverb, preposition, article, conjunction*) are tagged by their part of speech category only. Further tag sets are Anaphora Classes (*personal pronoun, demonstrative pronoun, abbreviation, special symbol*) and *Numbers*. Personal and demonstrative pronouns generally occur when the headline is a quotation. Special symbols like *$* have been considered as a separate anaphora class as in many

target language headlines the transliteration of the the full form of the symbol, i.e., *dollar*, is used. The *abbreviation* class includes both abbreviated words and acronyms. *Abbreviations* have been considered as a special class of anaphora as they are incomplete in nature and have to be resolved by either looking into the dictionary or in the first paragraph of the associated news item. The first paragraph in a news items is likely to include the content words of the associated headline. The abbreviated words can be resolved while the news headlines are collected from the corpus of English news items. Numbers can appear either in the form of digits or words in the source and the target language headlines and hence these are tagged separately.

5.2 Tagger / Recognizer / Classifier

The words / terms in the input English headlines in English are identified with the help of a tokenizer, a morphological analyzer and a Named entity recognizer(NER) and classifier. The system uses a lexicon of English words developed from the Wordnet 2.0 which includes the lexicographer file level tags associated with each word and term. The lexicon is being developed at the bilingual level and the Bengali meaning of the words are being entered in phases. A separate bilingual list is maintained for words that are pronouns, prepositions, articles and conjunctions. The words / terms are initially tagged at the part of speech (POS) level and then further tagged by a semantic tagger. The semantic tagger uses separate bilingual tables for abbreviations, acronyms, special symbols and various types of named entities. Identification of acronyms in long multi-paragraph news items causes problem as all words in the headline are sometimes written in all capital. Acronyms in short news headlines can be identified by looking for words which are all capital or may include a vowel in small case (e.g., HoD) or a special symbol (e.g., J & K). The system uses a Named Entity Recognizer and Classifier System for English developed in-house as part of a separate research activity. The NER system uses a *frequent starter's list* containing words that appear at the beginning of headlines but are not named entities themselves. This list has been prepared by looking into the English headlines collected in the example base. The NE classifier system is basically table driven and uses a limited set of features. A shallow parser for English (Naskar & Bandyopadhyay, 2005) is being used for identifying the compound nouns, noun phrases and the verb phrases in the input headline. The shallow parser can also detect whether the input headline is a quotation or a grammatical sentence. Thus, the shallow parser is identifying the syntactic category of the English news headline. The system also maintains a bilingual collection of collocations from English to Bengali. The Bengali portion of the parallel news headline is tagged by searching each word of the English headline in the appropriate dictionary or list and finding the Bengali equivalent. A match for the Bengali word is searched in the headline using a Bengali morphological analyzer. If a named entity or an acronym cannot be found in the bilingual dictionary, it is transliterated into Bengali and then searched in the Bengali headline. The Bengali headline may include additional words (nouns or adjectives) which are associated with the verb in the English headline. The system maintains a list of noun and adjective words associated with each meaning of a verb. These additional words are tagged separately using the Bengali lexicon which associates each Bengali noun and verb word with tags similar to those for English.

Let us consider the following examples of parallel English-Bengali headlines. The English gloss of the Bengali words are shown in brackets.

(i)   Train kills two ←→ ট্রেনে কাটা পড়ে মৃত দুই
       [*traine kaataa parhe mrita dui*]

(ii)   Bus kills 1 ←→ বাস দুর্ঘটনায় মৃত এক
       [*bus durghatanaaya mrita ek*]

(iii) Train kills 3 ←→ ট্রেন দুর্ঘটনায় মৃত তিন
       [*train durghatanaaya mrita tin*]

(iv)   Elephant kills three ←→ হাতির আক্রমণে মৃত তিন

       [*haatir aakramane mrita tin*]

(v)   Two killed in train accident ←→ ট্রেন দুর্ঘটনায় মৃত দুই
       [*train durghatanaaya mrita dui*]

The parallel example base of headlines after tagging will look like

(i)   <*train*, noun.artifact> <*kill*, verb.contact> <*two,*number> ←→ < ট্রেন [*train*], noun.artifact> <-ে [*-e*]> < কাটা পড় [*kaataa parh*], noun.event> <-ে [*-e*]> <মৃত্ [*mrita*], adjective>< দুই [*dui*], number>

(ii)   <*bus*, noun.artifact> <*kill*, verb.contact> <*1,*number> ←→ <বাস [*bus*],

130

noun.artifact> <দুর্ঘটনা.[*durghatanaa*],
noun.event> <-য [*-ya*]> <মৃত[*mrita*],
adjective> <এক [*ek*], number>

(iii) <*train*, noun.artifact> <*kill*, verb.contact>
<*3*,number> ←→
<ট্রেন [*train*], noun.artifact> <দুর্ঘটনা.
[*durghatanaa*], noun.event> <-য [*-ya*]>
<মৃত [*mrita*], adjective> <তিন [*tin*],
number>

(iv) <*elephant*, noun.animal> <*kill*,
verb.contact> <*three*,number> ←→ <হাতি
[*haati*], noun.animal> <-র [*-r*]> <আক্রমণ
[*aakraman*], noun.event> <-ে [*-e*]> <মৃত
[*mrita*], adjective> <তিন [*tin*], number>

(v) <*two*,number> <*kill*, verb.contact> <*in*,
preposition> <*train*, noun.artifact>
<*accident*, noun.event> ←→
<ট্রেন [*train*], noun.artifact> <দুর্ঘটনা.
[*durghatanaa*], noun.event> <-য [*-ya*]>
<মৃত [*mrita*], adjective> < দুই [*dui*],
number>

The two tags <noun.artifact> and <number>
can be directly aligned. The tags <-য>, <-র> and
<-ে> are Bengali inflections to be attached to
the preceeding word. The two tags <noun.event>
and <adjective> are associated with the tag
<verb.contact>. The system maintains a list of
verb.contact words and the associated
noun.event word. The *adjective* word is basically
used to qualify the object of the *verb.contact* and
the system maintains a list of such Bengali
adjectives for the verbs.

6 Creation of Generalized Tagged Example Base of English News Headlines

We are creating the tagged parallel example base
of news headlines with the help of English and
Bengali newspapers of the same date. The tagged
English headlines are automatically generalized.
The generalization process of tagged news
headlines is basically identifying the identical
tagged news headlines and then generalizing them.
Two tagged headlines can be considered identical
if they have identical tags at all the corresponding
positions. In the above example, tagged headlines
(i) <*train*, noun.artifact> <*kill*, verb.contact>
<*two*,number> and (ii) <*bus*, noun.artifact> <*kill*,
verb.contact> <*1*,number> and (iii) <*train*,
noun.artifact> <*kill*, verb.contact> <*3*,number> are

identical and they can be generalized to
<noun.artifact> <verb.contact> <number>. Two
tagged headlines can be considered similar if all
the tags present in one headline are present in the
other and the later headline includes noun and
adjective tags which can be derived from the verb
tag. In the above example, tagged headlines
(i)<*train*, noun.artifact> <*kill*, verb.contact>
<*two*,number>, (ii)<*bus*, noun.artifact> <*kill*,
verb.contact> <*1*,number>, (iii) <*train*,
noun.artifact> <*kill*, verb.contact> <*3*,number> and
(v) <*two*,number> <*kill*, verb.contact> <*in*,
preposition> <*train*, noun.artifact> <*accident*,
noun.event> are considered similar since all the
tags present in (i), (ii) and (iii) are present in (v)
and (v) includes the noun.event tag with the word
accident that can be derived from the verb tag
<*kill*, verb.contact>. The system will maintain the
list of such noun and adjective words that can be
derived from the verb word. Headlines which are
similar can be generalized at the next level.

The headlines that do not take part in any
generalization are kept in the Direct Example base
alongwith the tagging. During the development of
the parallel example base further match with
headlines in the Direct example base can occur and
the generalized headline can then be included in
the Generalized Tagged Example Base. It appears
from the above that the headlines (i), (ii) and (iii)
can be generalized on the English side and the
generalized tagged headline will be stored as
<noun.artifact> <verb.contact> <number> in the
Generalized Tagged Example Base. Since the
headlines (iv) and (v) cannot be generalized, the
original headlines will be kept in the Direct
Example base with its tags.

7 Creation of Generalized Tagged Example Base of Bengali News Headlines

Let us consider the five example parallel
English-Bengali news headlines as mentioned in
the section 5.2.. It can be seen that the set of
Bengali news headlines corresponding to the
English news headlines (i), (ii) and (iii), which
have been generalized, are not identical. This is a
general phenomenon and has been observed during
the devlopment of the example base. It also shows
the stylistic variations in news headlines across
languages and within the same language also. This
module identifies the tagged Bengali news
headlines that are identical for a tagged English
news headline and generalizes the Bengali news
headlines under each subset. In this example, the
two generalized tagged Bengali headlines
corresponding to the English news headlines (i),
(ii) and (iii) are identified as <noun.artifact>

<noun.event> <-য়> <adjective> <number> and <noun.artifact> <-েো> <noun.event> <-েো> <adjective><number>. These two generalized Bengali news headlines are shown to the developer for choosing one of them. The chosen Bengali tagged news headline can also be edited by the developer to maintain the informative nature and the style and the edited tagged Bengali news headline is associated with the generalized tagged English news headline and stored in the generalized tagged Example base. The parallel headlines are automatically aligned by their tags.

8 Creation of Example Bases (Direct Example Base, Generalized Tagged Example Base, Phrasal Example Base)

The system consists of three types of Example bases: (i) Direct Example Base, (ii) Generalized Tagged Example Base and (iii) Phrasal Example Base. The Direct Example Base is like the Translation Memory that stores the headlines in the source language and their translation in the target language. The Generalized Tagged Example Base stores the tagged examples with proper alignment. The Phrasal Example Base, stores the various phrase patterns in terms of the part of speech of the constituent words in the source language and their corresponding translation in the target language. The system uses the phrasal example base of English and Bengali phrase patterns used in a phrasal example based machine translation system (Naskar and Bandyopadhyay, 2005). The phrasal EBMT system is being developed for translating English news items to Bengali.

The Phrasal Example base consists of translation examples (phrasal templates) that store the part of speech of the constituent words of the phrases along with necessary syntactic information. Some examples of noun phrasal examples are:

(i)   <art $ a / an> <noun & singular, human, nominative> ←→<একজন [*ekjan*]> <noun>

(ii)   <art $ the> <noun & singular, human, objective> ←→ <noun> <-টিকে [*-tike*]>

(iii)   <art $ a / an> <adj> <noun & singular, inanimate, objective> ←→ <একটি [*ekti*]> <adj> <noun>

An example of a prepositional phrase is

(iv)   <prep $ to / at / in> <art $ the> <noun & singular, place> ←→ <noun> <- েো / েত [*-e/te*]>.

The headline "A hearty walk" may be translated by using the phrasal example base as the headline matches with the Noun phrasal example (iii).

9 Dictionary Design

The system uses a lexicon of English words developed from the Wordnet 2.0 which includes the lexicographer file level tags associated with each word and term. The lexicon is being developed at the bilingual level and the Bengali meaning of the words are being entered in phases. A separate bilingual list is maintained for words that are pronouns, prepositions, articles and conjunctions. There are separate bilingual tables for abbreviations and special symbols. Separate bilingual dictionaries for named entities and acronyms are maintained for each semantic and syntactic class in which the named entities and acronyms are most likely to occur. The named entity recognizer uses a *frequent starter's list* containing words that appear at the beginning of headlines but are not named entities themselves. This list has been prepared by looking into the English headlines collected in the example base. The system also maintains a bilingual collection of collocations from English to Bengali. The Bengali headline may include additional words (nouns or adjectives) which are associated with the verb in the English headline. The system maintains a list of noun and adjective words associated with each meaning of a verb.

10 MT System Development Methodology

During translation, the input headline is initially searched in the Direct example base for an exact match. If a match is obtained, the Bengali headline from the example base is produced as output. If there is no match, the headline is tagged and the tagged headline is searched in the Generalized Tagged Example base. If a match is obtained, the output Bengali headline is to be generated after appropriate synthesis. If a match is not found, the Phrasal example base will be used to generate the target translation. If the headline still cannot be translated, the following heuristic translation strategy will be applied: translation of the individual words or terms in their order of appearance in the input headline will generate the translation of the input headline. Appropriate dictionaries will be consulted to attempt a translation of the news headline.

Let us consider the five example parallel English-Bengali news headlines as mentioned in section 5.2.. If the headline *Elephant kills three* is given for translation, a match will be obtained in the Direct Example Base. The corresponding Bengali translation will be retrieved and then shown as output. If the headline *Tiger kills 1* is to be translated there will be no exact match in the Direct Example Base but the tagged version of the input headline will match with the tagged version of *Elephant kills three.* The two headlines will now be generalized on the source side and the tagged Bengali headline corresponding to *Elephant kills three* will be considered as the generalized tagged Bengali headline. The generalized tagged headlines will be included in the appropriate example base. The headline *Elephant kills three* will be removed from the direct example base. If the input headline is *Tram kills 2*, it will obtain a match in the generalized tagged example base and will be translated accordingly. If the input headline is *A sweet dream,* it will not find any match with either the direct or the generalized tagged example base. The Phrasal example base will then be consulted and the input headline will match with the noun phrase structure. The Bengali translation can be obtained accordingly. If the input headline is *Shastri Bhavan fire,* it will not find any match even in the phrasal example base. The headline will be iedntified as a compound noun as *Shastri Bhavan* is a named entity and *fire* is a noun. The system will produce an output following the heuristic. The named entity will be transliterated and the Bengali equivalent of the word *fire* will be obtained from the dictionary. The sequence of the two Bengali words will be presented as the output. The output will not be accepted by the user in this case as an inflection is necessary after the transliterated named entity and the heuristics could not produce that.

A preliminary version of the machine translation has been developed. The different example bases and the dictionaries are under development. Work is also going on for the development of a Bengali lexicon that includes the tags which are similar to those used in the English Wordnet at the Lexicographer file level.

11 Conclusion

Our news headline corpus is a collection of 2000 news headlines from the Kolkata edition of the News paper 'The Statesman'. A preliminary version of the machine translation has been developed. The different example bases and the dictionaries are under development. Work is also going on for the development of a Bengali lexicon that includes the tags which are similar to those

used in the English Wordnet at the Lexicographer file level. Initial testing of the MT System has started and no formal evaluation of the system has been carried out.

References

R.M.K. Sinha. 2002. *Translating News Headings from English to Hindi*. In "The 6th IASTED International Conference on Artificial Intelligence and Soft Computing (ASC2002) Proceedings", Banff, Canada.

D. Rao, and K. Mohanraj et al. 2000. *A Practical Framework For Syntactic Transfer Of Compound – Complex Sentences For English – Hindi Machine Translation*. In "International Conference KBCS-2000 Proceedings", Mumbai, India : 343-354.

S. Bandyopadhyay & D. Saha. 2002. *Anaphora / Coreference in Machine Translation of News Headlines*. In "Discourse Anaphora and Anaphora Resolution Colloquium (DAARC) 2002 Proceedings", Portugal.

S. Bandyopadhyay. 2000a. *An Example Based MT System in News Items Domain from English to Indian Languages*. In "Machine Translation and Multilingual Applications in the New Millennium Proceedings", Exeter, UK : 10-1 – 10-5.

S. Bandyopadhyay. 2000b. *ANUBAAD – Translating News Items from English to Bengali*. In "International Conference KBCS2000 Proceedings", Mumbai, India : 297-307.

J. Hutchins. 1999. *The Development & Use of Machine Translation Systems and Computer-based translation tools*. "The International Symposium on Machine Translation and Computer Language Information Processing Proceeding", China.

T. Yoshimi. 2001. Improvement of Translation Quality of English Newspaper Headlines by Automatic Pre-editing. *Journal of Machine Translation*, 16(4): 233-250.

C. Kim, M. Hong, Y. Huang, Y. Kim, S. Yang, Y. Seo and S. Choi. 2002. *Korean-Chinese Machine Translation Based on Verb Patterns*. Lecture Notes in Computer Science, Volume 2499: 94-103.

Kenji Ono. 2003. *Translation of News Headline*. In "MT SUMMIT IX Proceedings", New Orleans, Louisiana, USA.

Sudip Naskar and Sivaji Bandyopadhyay. 2005. *A Phrasal EBMT System for translating English to Bengali*. MT SUMMIT X, Phuket, Thailand.