An Example Based MT System in News Items Domain from English to Indian Languages

Dr. Sivaji Bandyopadhyay
Jadavpur University, Calcutta, India.

Abstract

The paper reports an on-going research project on a Knowledge driven Generalized Example based Machine Translation system from English to Indian languages. It is currently translating short single paragraph news items from English to Bengali. Headlines are translated using Knowledge bases and Example structures. The sentences in the news body are translated by analysis and synthesis. Semantic categories are associated with words to identify the inflections to be attached in the target language and to identify the context in the sentence. Context identification is also done by using context templates for each word. The Example base also includes the mapping of grammatical phrases from the source to the target language. The methodologies can be used for developing similar systems for other Indian languages.

1. Machine Translation in the News Items Domain

The domain of news items has attracted the attention of Machine Translation (MT) researchers all over the world. In India, a human-aided MT system for translating English news texts to Hindi is being developed at National Centre for Software Technology, Mumbai (http://www.ncst.ernet.in/kbcs/NLP.html).

The MT Research Group at the Information Sciences Institute (ISI), University of Southern California (USC), is developing programs that translate Japanese, Arabic and Spanish unrestricted newspaper texts into English. The work is reported in Knight et. al. (1994a), Hatzivassiloglou(1995) and Yamada(1996).

The Pangloss Mark III machine translation system translates unrestricted Spanish news-wire text into English. The work was carried out at the Center for Machine Translation, Carnegie Mellon University (CMT/CMU), ISI/USC and Computing Research Laboratory, New Mexico State University (CRL/NMSU). The translation from Russian News items into English was done at the CRL/NMSU. Details for this system can be obtained from Knight(1994b) and Brown(1996).

The NHK System in 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(1999).

The translation of names and technical terms is crucial in translating news items since these are not found in bilingual dictionaries. The results and examples of transliteration rules for names and technical terms from Arabic to English can be found in Knight(1997) and Stalls(1998).

Proper nouns like the place of news, month, date and the name of the news agency can be recognized since they occupy fixed places in a news item. The capitalization of proper nouns has been assumed in the present work. The Bilingual dictionaries include the proper noun, its classification and the target language representation. Proper nouns may also arise in a news item as acronyms. A separate Acronym Dictionary is used in the present system.

2. News Items : Structure & Classification

Short single paragraph news items are being sampled from the CALCUTTA Edition of the English News paper *The Statesman*. Generally these samples follow the structure :
< Headline >
<Place of News>, <Month>.<Date>. - <Body of the News> - <News Agency>.

The news items are classified based on the nature of the news. The different acronyms and proper nouns may vary for each classification and separate Bilingual dictionaries are kept for each of them. If the Example bases and the Knowledge bases are classified then the searching into them is focused, too.

3. Translation of the News Headlines

Translation of news headlines is being carried out using Example-based MT techniques. The Example base includes both specific and general examples. It includes general examples at the syntactic level also. The various phrases in the source language and their corresponding translation in the target language are stored. The present system thus can be termed as a Generalized Example based Machine Translation system.

The news headlines can be single word (e.g., *Repolling*), multiple word (e.g., *Cease fire plea*), an ungrammatical sentence generally without the auxiliary verb (e.g., *Kanika critical*) or a grammatical sentence (*Aatre is new Scientific Advisor*). The translation for the headline is first searched in the table, organized under each headline structure, containing specific source and target language pairs. If not found, the template representation of the headline is searched in another table. Headlines of similar structure are grouped into templates. For example, the two headlines *"Milk to be costlier in Punjab"* and *"Bread to be dearer in State"* follow the template structure : *<item> to be <costlier> in <place>*. The Bengali translation is stored as : *<place>-e <item>-er daam baarlo*.

If the headline still can not be translated, syntax directed translation techniques are applied. If it matches with any phrase of a sentence structure the translation is obtained using the Example base and the Bilingual dictionaries. Otherwise, word-by-word translation is attempted.

4. Dictionary Design

Root words in English are generally stored in a Dictionary. The *category* (e.g., proper, common etc.), *number* and *gender* are stored for a *noun word* along with its *semantic category*. Semantic categories are associated with words to identify the inflections to be attached with corresponding words in Indian languages as well as to identify the context in the sentence. The semantic category can be independent of the application domain, e.g., Tiger : *<Animate Object>* & Book : *<Inanimate Object>*. Some semantic categories can be dependent on the application domain, e.g., Microsoft : *<Company Name>*. The word *share* has different meanings in *Equal share* and *Microsoft share*. The *person*, *case*, *number* and *gender* are stored for *pronoun words*. For words of all other categories only the part of speech information is stored.

The design of the Bilingual Dictionary requires an understanding of the context in which a source language word is used. *Context templates* have been associated with words to identify the context in which the word is used so that its appropriate meaning in the target language can be retrieved from the Bilingual Dictionary. The meaning of a word in English may vary under different parts of speech. Further, the meaning of a word in English may be independent of the context (e.g. *boy*), may depend on the occurrence of a sequence of words (e.g., *run across*) or words with certain semantic categories (e.g., *run a <fever>* and the words *fever* & *temperature* associated with the semantic category *<fever>*) or may depend on the occurrence of certain keywords or keywords with certain semantic categories (e.g. *bank* along with the keyword *river* or *bank* along with keywords with semantic category *<Financial Institution>* have different meaning). These context templates are included in the Bilingual Co-occurrence Dictionaries along with appropriate pointers from the main Bilingual Dictionary which contains the root words in the source language and their basic meanings.

Context identification is also done by the recognition of Figure of Speech expressions. A separate Figure of Speech Dictionary stores such expressions in English along with their corresponding counterparts in the target language.

5. Different phases of the Translation System

The work is being carried out in the Visual Studio 6.0 environment with Visual C++ 6.0 as the programming language and Microsoft Access 2000 as the associated database management system. This section describes the different phases of the syntax directed translation in the present system. The general structure of a simple assertive sentence in English in active voice is represented as follows:
{ *Adverb* | *Preposition*} {*Noun Phrase1*} {*Verb Group*} {*Adverb* | *Preposition*}{*Noun Phrase2*}.

Other types of simple sentences are not possible in news items domain. Complex sentences are translated by identifying the *main clause* and the *dependent clause* and then translating the two clauses separately. Similarly,

the compound sentences are translated by identifying the conjunction or disjunction in the sentence and translating the two parts separately.

The Knowledge bases include the Suffix Table for Morphological Analysis of English surface level words, Parsing table for Syntactic Analysis of English, Bilingual Dictionaries for different classes of proper nouns, different dictionaries, different tables for synthesis in the target language. The Morphological Analysis is done using Suffix Tables and the English Dictionary. The Suffix Table stores the *suffixes in reverse order*, the *original end pattern* and the *changed end pattern* of the word. The Syntactic Analysis is done using a table driven parser. Auxiliary verbs in English, which do not directly translate into Indian languages but help in identifying the *tense* of the verb are represented as *null* after the analysis phase unless it is the main verb of the sentence. The tense information for a verb is obtained during suffix analysis. Special verb forms like *went* are stored in a separate table and no suffix analysis is required for them. Similar tables are present for irregular noun forms like *men* etc.. The *number* and *person* information for a verb are obtained from the immediately preceding noun or pronoun. Analysis of adjectives takes into account words like *more, most, less, least* etc. occurring before it. Idiomatic expressions are taken care of separately.

The root words in the target language are retrieved from the dictionaries. Each surface level word is then synthesized. Each phrase in the input sentence is now considered and its corresponding mapping in the target language is retrieved from the Example base. Special consideration have been made for prepositional phrases. Some inflections are added to the last noun in the prepositional phrase based on the matra of the last symbol of the word or on the semantic category of the word. For example, *on Monday* is translated as *sombaare* but *on the table* is translated as *tebiler upar*. The semantic category of the word *Monday* is <*day*>, its translation to Bengali is *sombaar* and the inflection that has been added to the word is *–e*. The semantic category of the word *table* is neither *date* nor *day*, its translation to Bengali is *tebil* and the inflection that has been added is *–er upar*.

Finally, the translation of the different phrases are assembled. The simple assertive sentences with multiple phrases are translated from English to Indian languages by the following rule :

- the order of all the phrases before the verb phrase are inverted.
- the order of all the phrases after the verb phrase are similarly inverted.
- the verb phrase is put at the end of the sentence.

For example, the different phrases of the sentence "*He has been cultivating the land with a spade in the garden since morning for better crop.*" and their translation into Bengali are as follows :
He – Noun Phrase – *Se*
has been cultivating - Verb Phrase – *chaas karchhilo*
the land – Noun Phrase – *jami*
with a spade – Prepositional Phrase (PP) – *kodaal diye*
in the garden – PP – *baagaane*

since morning - PP – *sakaal theke*
for better crop - PP – *bhaalo fasoler janya*

The Bengali translation of the sentence is as follows :
Se bhaalo fasoler janya sakaal theke baagaane kodaal diye jami chaas karchhilo.

6. Results and Discussion

The methodologies for a Knowledge driven Generalized Example based Machine Translation system from English to Indian languages in the news items domain have been developed. Currently, a prototype of the system in English-Bengali translation has been developed. The post-editing part of the system is not yet ready. The methodologies can be used for developing similar systems for other Indian languages.

Acknowledgements

The work is being carried out as part of a Research Award granted to the author by the University Grants Commission (UGC), Government of India in 1999 (UGC Research Award for the IXth Plan Period), F.30-95/98 (SA-III).

References

Brown Ralf D. (1996), 'Example-Based Machine Translation in the Pangloss System', in the Proceedings of the COLING-96.
Hatzivassiloglou V. and K. Knight (1995), 'Unification-Based Glossing', in the Proceedings of the 14th IJCAI Conference.
Hutchins J. (1999), 'The Development & Use of Machine Translation Systems and Computer-based translation tools' in the International Symposium on Machine Translation and Computer Language Information Processing.
Knight K. et. al., (1994a), 'Integrating Knowledge Bases and Statistics in MT' in the Proceedings of the 1st AMTA Conference.
Knight K. and S. Luk (1994b), 'Building a Large-Scale Knowledge Base for Machine Translation', in the Proceedings of the AAAI-94.
Knight K. and J. Graehl (1997), 'Machine Transliteration' in the Proceedings of the ACL-97.
Stalls B. and K. Knight (1998), 'Translating Names and Technical Terms in Arabic Text', in the COLING/ACL Workshop on Computational Approaches to Semitic Langauges.
Yamada K. (1996), 'A Controlled Skip Parser', in the Proceedings of 2nd AMTA Conference.
Yoshimi T. (1999), 'Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting', in the Proceedings of the MT Summit VII.