

Translation from English to Indian Languages: ANGLABHARTI Approach

R.M.K.Sinha
Deptt. of C.S.E.
I.I.T. Kanpur
Kanpur

Renu Jain
Deptt. of C.S.E.
I.E.T., CSJM University
Kanpu

Ajai Jain
Deptt. of C.S.E.
I.I.T. Kanpur
Kanpur

Abstract

An English to Indian languages machine aided translation system, named ANGLABHARTI, has been developed. It uses pattern directed approach using context free grammar like structures. A 'pseudo-target' is generated which is applicable to a group of Indian languages. Set of rules are acquired through corpus analysis to identify the plausible constituents with respect to which movement rules for the 'pseudo-target' are constructed. A number of semantic tags are used to resolve sense ambiguity in the source language. Alternative meanings for the unresolved ambiguities are retained in the pseudo target language code. A text generator module for each of the target languages transforms the pseudo target language to the target language. A human-engineered post-editing package is used to make the final corrections. The post-editor needs to know only the target language. The strategy used in ANGLABHARTI lies in between the transfer and the interlingua approach. It is better than the transfer approach, as the translation is valid for a host of target language sentences, but falls short of genuine interlingua, in the sense that it ignores complete disambiguation/understanding of the text to be translated.

1. Introduction

This paper presents the details of the ANGLABHARTI system, which is a multi-lingual translation system for translation from English to Indian languages. India is highly multi-lingual country with a large number of living languages. Indian languages are verb ending, free word-group order language with lots of structural similarity. Indian languages can be classified into four broad groups according to their origin. These are Indo-Aryan family (Hindi, Bangla, Asamiya, Punjabi, Marathi, Oriya, Gujrati etc.); Dravidian family (Tamil, Telugu, Kannada & Malayalam); Austro-Asian family and Tibetan-Burmese

family. Within each group the languages exhibit a high degree of structural homogeneity. We exploit this similarity to a great extent in our system.

ANGLABHARTI is a pattern directed rule based system defining patterns in context free grammar like structures for the source language and for the target languages applicable to a group of Indian languages. A set of hand-crafted rules are obtained through corpus analysis of source and target language patterns and are used to identify plausible constituents with respect to which movement rules for the 'pseudo-target' are constructed. The idea of using 'pseudo-target' is primarily to exploit structural similarity to obtain advantages similar to that of using interlingua approach. The 'pseudo-target' is not the intermediate language used in interlingua in the sense that no attempt is made to develop a meaning representation here.

Looking back at the history of machine translation (MT), we find that three major strategies have governed the design of MT systems over the last two decades [1]: Direct translation strategy, Transfer strategy and Interlingua strategy. The direct translation system is designed for a specific source and target language pair with no intermediate representations. No general linguistic theory or parsing principles are necessarily present for direct translation to work. These systems depend on well developed dictionaries, morphological analysis, and text processing software to get the translation. SYSTRAN [2] is an example of such a system.

In the transfer strategy, a source language sentence is first parsed into phrases according to the structure of the source language. Thereafter, a 'transfer' is made at the lexical level, at the grammar level or at the structural level built by the grammar, and so-forth into corresponding structures in the target language. In the third stage, translation is generated. This strategy was popularized by the system like SUSY [3], TAUM-METEO for translating weather reports from English to French in Canada.

The interlingua-based approach depends upon the theme that a suitable universal intermediate representation can be defined for the source text which is independent of the target language. This intermediate representation is presumed to have resolved all ambiguities, and so it should be possible to generate text in any language at the target from this representation. However, it is almost impossible to derive a truly natural language independent intermediate representation.

The Anglabharti system aims at providing a practical machine aided translation system. It assumes that the perfect translation is not feasible with the present day technology and resource constraints. However, it is observed that 90% of the commonly used/encountered sentences can be translated using rules and a partial knowledge-base, and only 10% require disambiguation/correction by post-editing. As the Indian languages are similar to each other, we have divided the task of translation into two distinct

stages: one which obtains an intermediate representation (called pseudo target) applicable to a group of languages, and the other is the generation of the text from this intermediate representation. The intermediate representation derives the appropriate word-group order for the target along with the host of information obtained from the lexical database. Presently, we have a working system for Indo-Aryan family with text generator for the target language of Hindi. We are aiming towards the development of a complete translation system specially in Health Campaign and Information Technology context. The following section presents the details of the system

2. System Design

Figure 1. depicts the architecture of ANGLABHARTI system. The system has six major processing modules: Morphological analyzer, Parser, Pseudo code generator, Sense disambiguator, Target language text generators, and Post-editor. The system makes use of mainly two databases and many paradigm files for the source language and for the target languages during the process of translation. One of the database is a multi-lingual [4] lexical database and the other is a rule-base for pattern invocation and transformation from English to the pseudo target.

2.1 Morphological Analyzer

Morphological analyzer takes as input the English sentence, reads each word, identifies the root word, and retrieves the necessary information from the lexical database about that word. Identification of different kinds of phrasals(group of words when used together exhibit entirely different meaning from their original meanings) is also done during morphological analysis. If a particular word is not found in the lexical database, system transliterates the word and also tries to guess the expected syntactic category of the word. An on-line lexical database is created, as we proceed with the process of analysis of the input text. Here, we store information about all the words encountered in the text, and the analyzer first examines the entries of this lexical database before searching in the large multi-lingual database. As a large number of words get repeated in a typical text, the on-line lexical database helps to reduce the search time to a very large extent.

Morphological analyzer makes use of the multi-lingual lexical database and different paradigm files for the source language to extract the root word from different forms of the word. For Example: the word 'play' as a verb has 5 forms(plays, play, played, played, playing), and as a noun it has two forms(play,plays). But, in the lexical database, only the root word 'play' is stored with syntactic and semantic knowledge and also for the target language only the root word meaning is stored.

2.2 Multi-lingual lexical database- Design issues

A lexical database is like an electronic dictionary(a collection of words along with the information about its different syntactic categories, its derivatives and uses under different possible contexts) with some additional information needed for the translation system or for some other NLP application. We are developing a multi-lingual lexical database for ANGLABHARTI system. This multi-lingual lexical database keeps most of the syntactic, semantic and other information needed for disambiguation, only for the source language. It has multiple fields for each of the target language where each field has TL meanings and some TL specific information needed by the target language text generator. To resolve multiple target meanings, certain patterns, semantic tags and some constraints for disambiguation are also stored.

lexicon	syntactic info	English meaning	verb pattern	semantic tags	First TL meanings	Second TL meanings
---------	-------------------	--------------------	-----------------	------------------	----------------------	-----------------------

The above table shows the general format of entry in our multi-lingual dictionary.

In the above figure, lexicon represents the root word of the source language. Syntactic information means category information of the word along with other information like number, person, gender etc. Each verb in any language exhibits certain expectations and according to that possible syntactic patterns are assigned to each verb. These patterns help us in selecting the correct meaning of the verb while usage.

Semantic tags try to represent the behaviour and perception of the objects represented by the lexicon. A list of semantic tags and their hierarchy has been defined in Figure 2.

This list of semantic tags has been designed keeping in mind the application of machine translation. This semantic tree is for nouns only and these nouns tags are assigned to the expected subject and object of the verb. However, any list of semantic tags can not be stated as complete and every designer decides based on the application and experiences gained through the analysis of the language pair.

There are many issues related to the development of such a lexical database in the context of machine translation or for any of the NLP work. Lexical database used for translation is basically a collection of hand-crafted limited knowledge. But the amount of knowledge we use for the translation work is much more than the knowledge put in the lexical database. We make use of so much implicit knowledge and explicit knowledge while translating that it really becomes difficult for us to first pin point the exact knowledge used for the translation and even if we are able to do so, it is difficult to represent the knowledge and then make use of the knowledge at the right place. The main issues related to the development of lexical database are identification of appropriate knowledge necessary for machine

translation(or for any NLP application), collection of knowledge, representation of knowledge and use of knowledge to be able to produce the correct translation.

Development of such lexical databases can not be done just by the data entry operators. We need people who have good feel of both source language and target languages. We also need help of linguists to identify the patterns, semantic tags and constraints needed for disambiguation.

2.3 Intermediate form generation

In this section, all the three processing modules have been discussed. A pattern directed parsing is performed on the source language, English. Parser makes use of the syntactic and semantic information of the words of the sentence provided by the morphological analyzer and a knowledge base called rule-base. Rule-base used by the parser is a collection of hand crafted patterns in Context Free Grammar like structures for the source language(English) and the corresponding target language structures applicable to a group of Indian Languages. It has been developed by examining the corpus of the source language and the target language sentences. A rule is constructed by substituting constant parts of the sentence to the variable parts as far as possible if the right hand side depicting the pseudo target code remains the same.

Information about the words of the input sentence is used to form patterns and these patterns are matched to the left-hand side of the rules stored in the rule-base. Top down approach is used for matching the rules. System recursively tries to unify with all the possible patterns and on finding the match, the corresponding rule is invoked, and the right-hand side of the rule yields the pseudo target. System tries all the possible combinations by back tracking and retains only the successful ones and due to that unification of multiple rules is also possible. In such a case, more than one pseudo target is generated, and post-editing is required to choose the most appropriate parsing on the basis of the context of the sentence. The idea of using the pseudo target is primarily to exploit structural similarity to obtain advantages similar to that of using interlingua approach. The pseudo target is not truly in the language used in interlingua in the sense that no attempt has been made here to develop a meaning representation.

A sample typical rule in the rule-base is as follows :

noun-phrase verb-phrase prep-phrase --- > target-noun-phrase target-prep-phrase target-verb-phrase

These are non-terminal symbols and these are further defined in terms of terminal or non-terminal symbols. For example:

basic-noun-phrase = noun/determiner noun/ adjective noun

noun-phrase = basic-noun-phrase/basic-noun-phrase 'and' basic-noun-phrase

and similarly prep-phrase and verb-phrase are defined.

After the identification of appropriate pattern/patterns for an input sentence, pseudo code generator generates the intermediate form according to the structure of the target language for all the parsings. Sense disambiguator makes use of semantic knowledge and verb patterns supplied by the morphological analyzer to discard some of the parsings and some of the meanings reducing number of translations. But, sense disambiguator is able to disambiguate only partially due to incomplete and inconsistent knowledge.

From the above examples, we can see that the same word is translated differently in target language depending upon the sentence type. In some cases meaning resolution is possible while in other cases it shows both meanings of the word 'boil'. Multiple translations are obtained mainly due to several valid rules being fired.

2.5 Target language sentence generation

Intermediate version generated by the rule-base is synthesized and target language translation is generated by the target language text generator. For each target language, we will have to have a separate text generator which will make use of language specific rules to supply correct post or pre positions in the sentence and to generate the correct form of words according to the information provided by the intermediate form. Each text generator takes care of the peculiarities of its language. A number of grammatical rules of the target language in the form of expectations and constraints are used to select the appropriate case markers, affixes, etc. A Paninian framework [5] is used for this purpose. At this stage, it is possible that some of the target sentences are still ill-formed. A Corrector for ill-formed sentences is utilized [6] for further improvement. The ambiguities which still remain unresolved, are taken care by human post-editing.

Here, we illustrate the process of generation of a Hindi sentence for a given sentence by Hindi text generator.

English Input Sentence: *The boys ate all the good apples.*

:Intermediate version < aff { sub_np (the det [] [anda] [A]) (boys noun masculine plural third [human] [ladZakA:m 11] [] []) } { obj1_np (allthe adjective any [NIL] [saBI] [] [])(good adjective any [NIL] [acCA] [] []) (apples noun neuter plural third [edible] [seba:m 6] [] []) } k1 { main_vp (eat verb_3 normal normal masculine plural third [KA] 1 [] []) } > . sviram

Hindi Output Sentence: *ladZakoM ne saBI acCe seba KAe.*

Intermediate form contains the hindi meaning of root word of 'boys' as 'ladZakA' which is transformed by the text generator into 'ladZakoM'. 'ladZakoM' is generated from 'ladZakA' on the basis of the number, person, gender of the noun and also looking at the form of the verb. In Hindi, nouns can have two cases 'oblique' and 'dative'. If a post-position follows the noun, noun is oblique otherwise it is taken as dative and then according to GNP of the noun, noun is transformed into correct form. The type of the verb and form of the verb decide the post-position 'ne' after the subject and accordingly subject 'ladZakA' is changed into 'ladZakoM'. Similarly, forms of adjective meaning are changed according to the Gender, Number and Person of the following noun. Root form of verb is transformed either examining the properties of subject or object and the tense of the sentence. In the above example, verb is in agreement with the object and if we change the object by a feminine object like 'medicines' then 'KAe' will be changed to 'KAIM' . Many hand crafted rules and different types of paradigm files for different syntactic categories are used. Paninian framework provides a convenient way of imposing constraints. However, it may not be possible to build such rules which can completely disambiguate with the limited information available, and requires context for resolution. These ambiguities are left for human post editing.

3. Conclusions

A multi-lingual machine aided translation system called 'ANGLABHARTI' has been designed. Based on this, a functional translation system for English to Hindi has been developed and still more work is going on to incorporate all the possible types of patterns and improve the quality of translation. Attempts are being made to get 90% of the task done by the machine and 10% left to the human post-editor. In order to get better translation quality, we are trying to incorporate context knowledge so that more appropriate meaning can be picked up. About 500 rules have been implemented, and it was found that system is able to translate most of the different kinds of encountered sentences. However, for a given source English sentence, the system sometimes produces more than one target sentences due to multiple meanings and due to multiple parsings. Work needs to be done in the direction of development of target language text generators for other languages and more rules have to be evolved to find the correct role of prepositions in the sentence and to minimize the number of translations.

References

- [1] A.B. Tucker., "Current Strategies in Machine Translation Research and Development", *Machine Translation, Theoretical and Methodological Issues*, (ed.) Sergei Nirenburg, Cambridge University Press, pp. 2-41, 1987.
- [2] P. Toma., "SYSTRAN as a Multi-Lingual Machine Translation System in Commission of European Communities:Overcoming the Language Barrier", *Munich: Dokumentation Verlag* ,pp. 129-160, 1977.
- [3] H.D. Maas., "The MT System SUSY", *paper presented at The ISSCO Tutorial on Machine Translation*, 1984.
- [4] R.M.K.Sinha, K.Sivaraman, Aditi Agrawal, T. Suresh, Chinmoyee Sanyal., " On Logical Design of Multilingual Lexicon for Machine Translation" *Technical Report, Department of Computer Science and Engineering, I.I.T.Kanpur*", TRCS-93-173, 1993.
- [5] G.Cardona., "PANINI: A Survey of Research" *Motilal Banarsidas, Delhi*, 1976.
- [6] R.M.K.Sinha and C.Sanyal., "Correcting Ill Formed Hindi Sentences in Machine Translated Output", *Natural Language Processing Pacific Rim Symposium NLPRS:93,Japan.*, 1993.
- [7] A.S.Hornby., "A Guide to Patterns and Usage in English", *Oxford University Press*, 1985.

