AnglaHindi:
An English to Hindi Machine-Aided Translation System

R.M.K. Sinha (rmk@iitk.ac.in) **A. Jain** (ajain@iitk.ac.in)
Chief Investigator Co-Investigator
Indian Institute of Technology, Kanpur, India

Abstract

This paper presents a system overview of an English to Hindi Machine-Aided Translation System named AnglaHindi. Its beta-version has been made available on the internet for free translation at http://anglahindi.iitk.ac.in AnglaHindi is an English to Hindi version of the ANGLABHARTI translation methodology developed by the author for translation from English to all Indian languages. Anglabharti is a pseudo-interlingual rule-based translation methodology. AnglaHindi, besides using the rule-bases, uses example-base and statistics to obtain more acceptable and accurate translation for frequently encountered noun and verb phrasals. This way a limited hybridization of rule-based and example-based approaches has been incorporated.

1 Introduction

India is a highly multilingual country with eighteen constitutionally recognized languages and several hundred dialects & other living languages. Even though, English is understood by less than 3% of Indian population, it continues to be the de-facto link language for administration, education and business. Hindi, which is official language of the country, is used by more than 400 million people. Therefore, machine translation assumes a much greater significance in breaking the language barrier within the country's sociological structure. In this paper, we present a glimpse of our effort in this direction. Our work on machine translation started in early eighties when we proposed using Sanskrit as interlingua for translation to and from Indian languages (Sinha, 1984; Sinha, 1989). However, as English continues to be the link language, a machine translation system catering to English as the source language and the target language being all Indian languages, was considered to be a priority. Further, as the state of current technology is short of producing high quality automated translation and the human translators are unable to cope up with the volume, a machine-aided translation (MAT) system is an obvious answer. ANGLABHARTI (Sinha et.al., 1995) is a rule-based MAT system with source language as English and uses a pseudo-interlingua to cater to all Indian languages. Although, the design methodology of Anglabharti, is geared to achieve an 'acceptable' translation at the first instance, it is recognized that the system will have inherent weaknesses of being short of producing 'quality' translation thus requiring post-editing. AnglaHindi is an English to Hindi version of the ANGLABHARTI translation methodology with a mixture of some example-based translation methodology. AnglaHindi system has been web-enabled and is available at URL: http://anglahindi.iitk.ac.in for free translation. This is first such system designed to our knowledge. This paper presents an overview of AnglaHiindi system.

2 System Overview

As AnglaHindi is a derivative of Anglabharti, let us first look at the Anglabharti methodology. As pointed out earlier, Anglabharti is a machine-aided translation methodology specifically designed for translating English to Indian languages. English is a SVO language while Indian languages are SOV and are relatively of free word-order. Instead of designing translators for English to each Indian language, Anglabharti uses a pseudo-interlingua approach. It analyses English sentences only once and creates an intermediate structure with most of the disambiguation performed. The intermediate language structure has the word and word-group order as per the structure of the group of target languages. The intermediate structure is then converted to each Indian language through a process of text-generation. The effort in analyzing the English sentences is about 70% and the text-generation accounts for the rest of the 30%. Thus only with an additional 30% effort, a new English to Indian language translator can be built.

Anglabharti is a pattern directed rule based system with context free grammar like structure for analysis of English as source language. The analysis generates a `pseudo-target' applicable to a group of Indian languages. A set of rules obtained through corpus analysis is used to identify plausible constituents with respect to which movement rules for the 'pseudo-target' is constructed. The idea of using `pseudo-target' is primarily aimed at incorporating advantages similar to that of using interlingua approach exploiting structural similarity. Indian languages are verb ending, free word-group order, and a lot of structural similarity. Indian languages can be classified into four broad groups according to their origin and similarity. 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, there is a high degree of structural similarity. Paninian framework based on Sanskrit grammar using Karak (similar to 'case') relationship provides an uniform way of designing the Indian language text generators using selectional constraints and preferences.

A block schematic diagram of the Anglabharti methodology is depicted in figure 1. A brief description of some of the major building blocks of Anglabharti is given in the following paragraphs.
Rule-base**:** This contains rules for mapping structures of sentence from English to Indian languages. This database of pattern-transformations from English to Indian languages is entrusted the job of making a surface-tree to surface-tree transformation, bypassing the task of getting a deep tree of the sentence to be translated. The data base of structural transformation rules from English to Indian languages forms the heart of the Anglabharti system. The system is designed to cater to compound, complex, imperative, interrogative and other constructs such as headings etc. As mentioned earlier, by making a generic rule-base for Indian languages, Anglabharti exhibits a potential benefit while translating from English. This module is also responsible for picking up the correct sense of each word in the source language to the extent feasible using interleaved semantic interpreter. Further disambiguation and choice of right construct and lexical preferences are performed by the target language text-generator module. Many a time, multiple rules may get invoked leading to multiple interpretation of the input sentence. The rules are ordered in terms of their preferences and an upper limit is put on the number of alternatives produced. These multiple translations are available for further post-editing.
Multi-lingual dictionary/ Lexical data-base and Sense Disambiguator: The lexical database is the fuel to the translation engine. It contains various details for each word in English, like their syntactic categories, possible senses, keys to disambiguate their senses, corresponding words in target languages with their tags. A number of ontological/semantic tags are used to resolve sense ambiguity in the source language. Most of the disambiguation rules are in the form of syntacto-semantic constraints. We use semantics to resolve most of the intra-sentence anaphora/pronoun references. Alternative meanings for the unresolved ambiguities are retained in the pseudo target language. The lexical database is hierarchically organized to allow domain specific meanings and also prioritize meanings as per users' requirement.
Target text generators and Corrector for ill Formed Sentences (Sinha, Srivastava and Agrawal, 1995; Sinha and Sanyal, 1993): These form the tail end of the system. Their function is to generate the translated output for the corresponding target languages. A text generator module for each of the target languages transforms the pseudo target language to the target language. These transformations do lead to sentences which may be ill-formed. The ill-formed sentences are target language specific and are usually related to incorrect placement of emphasizers, negation and forms denoting cultural dependence (such as plurals being used for persons whom you pay respect). A corrector for ill-formed sentences is used for each of the target languages. Finally, a human-engineered post-editing package is used to make the final corrections. It is our experience that for more than 50% of the normal text, the human post-editor needs to know only the target language

as the humans use a lot of contextual information in making the right choice. For resolving the structural ambiguity, one needs to consult the source language. It may be noted that by having different text generators using the same rule-base and sense disambiguator, a generic MT system is obtained for a host of target languages. We have used Paninian framework with verb-centric expectation driven methodology (Sinha, 1989) with selectional restrictions/semantic constraints for synthesizing the Indian language text.

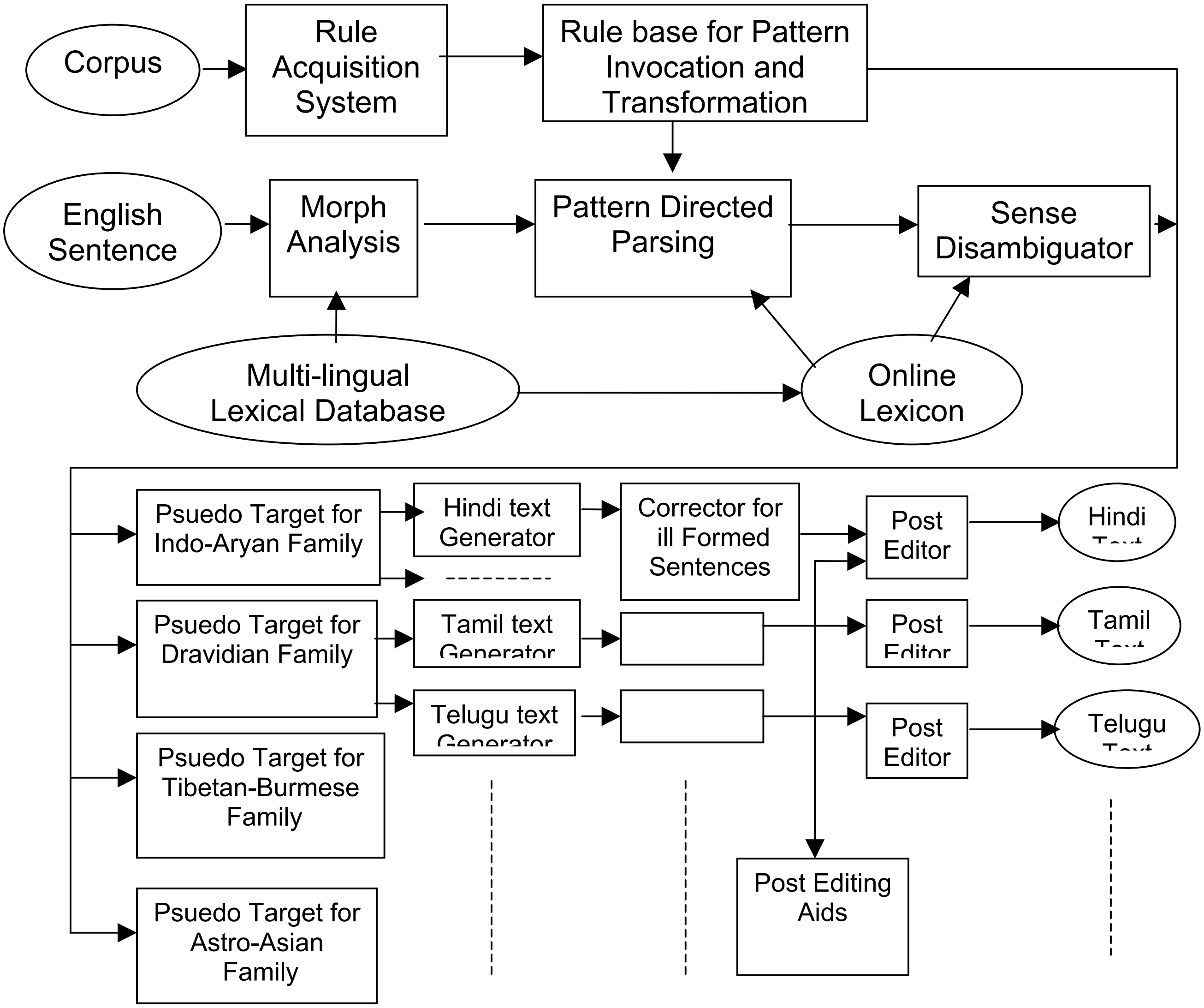
Corpus
Rule Acquisition System
Rule base for Pattern Invocation and Transformation
English Sentence
Morph Analysis
Pattern Directed Parsing
Sense Disambiguator
Multi-lingual Lexical Database
Online Lexicon
Psuedo Target for Indo-Aryan Family
Hindi text Generator
Corrector for ill Formed Sentences
Post Editor
Hindi Text
Psuedo Target for Dravidian Family
Tamil text Generator
Post Editor
Tamil Text
Telugu text Generator
Post Editor
Telugu Text
Psuedo Target for Tibetan-Burmese Family
Psuedo Target for Astro-Asian Family
Post Editing Aids

Figure 1: System Architecture of ANGLABHARTI

AnglaHindi besides using all the modules of Anglabharti, also makes use of an abstracted example-base for translating frequently encountered noun phrases and verb phrasals. The example-based approach developed by the author's group, named ANUBHARTI (Jain, Sinha and Jain, 1995,2001), is invoked before the rule-based approach is applied. The example-base is statistically derived from the corpus. Ambiguities in the meanings of the verb phrasals are also resolved using an appropriate distance function in the example-base (Bhandari, Sinha and Jain, 2002). AnglaHindi accepts unconstrained text (Jain, Sinha and Jain, 2002; Sinha, 2001). The text may be made up of headings, parenthesized texts, text under quote marks, currencies, varying numeral & date conventions, acronyms, unknowns and other frequently encountered constructs. The performance of the system has been evaluated by human translators. The system generates approximately 90% acceptable translation in case of simple, compound and complex sentences upto a length of 20 words.

Current version of AnglaHindi is not tuned to any specific domain of application or topic. However, it has user friendly interfaces which allows hierarchical structuring of the lexical database leading to preferences on lexical choice. Similarly, it has provisions for augmenting its abstracted example-base specific to an application domain. This not only eliminates the alterative translations but also generates more accurate and acceptable translation. Currently, the alternate translations are being ranked with respect to the ordering of the rule-base. This can be further enhanced by using domain specific information and target language statistics. The alternate translations can be ranked based on hidden Markov model of Hindi in the specific domain. For each alternate translation, the language model yields a figure of merit reflecting preferences for style and lexical choice.

Overall, the AnglaHindi system attempts to integrate (Sinha, 2000) example-based approach with rule-base and human engineered post-editing. An attempt is made to fuse the modern artificial-intelligence techniques with the classical Paninian framework based on Sanskrit grammar.

3 Related References :

Vartika Bhandari, R.M.K. Sinha and Ajai Jain. 2002. Disambiguation of Phrasal Verb Occurrence for Machine Translation, In *Proc. Symposium on Translation Support Systems STRANS2002*, March 15-17, Kanpur, India.

Ajai Jain, R.M.K. Sinha and Renu Jain. 2002. On Translating Unconstrained Text, In *Proc. Symposium on Translation Support Systems STRANS2002*, March 15-17, Kanpur, India.

R.M.K. Sinha. 2001. Dealing with Unknown Lexicons in Machine Translation from English to Hindi, In *Proc. of IASTED International Conference on Artificial Intelligence and Soft Computing*, May 21-24, Cancun, Mexico, pp 333-336.

R.M.K. Sinha, Renu Jain and Ajai Jain. 2001 Translation from English to Indian Languages: ANGLABHARTI Approach, In *Proc. Symposium on Translation Support Systems STRANS2001*, February 15-17, Kanpur, India.

Renu Jain, R.M.K. Sinha and Ajai Jain. 2001. ANUBHARTI: Using Hybrid Example-Based Approach for Machine Translation In *Proc. Symposium on Translation Support Systems STRANS2001*, February 15-17, Kanpur, India.

R.M.K. Sinha. 2000. Hybridizing Rule-Based and Example-Based Approaches in Machine Aided Translation System, In *Proc. International Conference on Artificial Intelligence IC-AI'2000*, June 26-29, Las Vegas, USA.

R.Jain, R.M.K.Sinha, A.Jain. 1997. Translation between English and Indian Languages, *Journal of Computer Science and Informatics*, pp 19 -25.

R.M.K. Sinha and others. 1995. ANGLABHARTI: A Multi-lingual Machine Aided Translation Project on Translation from English to Hindi, In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, Vancouver, Canada, pp 1609-1614.

Renu Jain, R.M.K. Sinha and A. Jain. 1995. Role of Examples in Machine Translation, In *Proc. IEEE International Conference on Systems, Man and Cybernetics*, Vancouver, Canada, pp 1615-1620.

R.M.K. Sinha, R. Srivastava and A. Agrawal. 1995. Designing Hindi Text Generator for Machine Translation, In *Proc. Symposium on Natural Language Processing, SNLP'95*, Bangkok, Thailand, pp 286-296.

Renu Jain, R.M.K. Sinha, A. Jain and R. Srivastava. 1995. HFSM: A Finite State Machine for Analyzing Hindi Sentences, In *Proc. Symposium on Natural Language Processing, SNLP'95*, Bangkok, Thailand, pp 317-324.

Renu Jain, R.M.K. Sinha and A. Jain. 1995. A Pattern Directed Hybrid Approach to Machine Translation through Examples, In *Proc. Symposium on Natural Language Processing, SNLP'95*, Bangkok, Thailand, pp 325-335.

R.M.K. Sinha and C. Sanyal. 1993. Correcting ill-formed Hindi sentences in machine translated output' In *Proceedings of Natural Language Processing Pacific Rim Symposium NLPRS'93*, Fukuoka, Japan, pp 109-119.

R.M.K. Sinha. 1989. A Sanskrit based Word-expert model for machine translation among Indian languages, In *Proc. of workshop on Computer Processing of Asian Languages*, AIT, Bangkok, Thailand, Sept.26-28, pp. 82-91.

R.M.K. Sinha. 1984. Computer processing of Indian Languages and Scripts - Potentialities and Problems, *Jour. of Inst. Electron. & Telecom. Engrs(India),* 30(6) pp. 133-149.