Study of various approaches in Machine Translation for Sanskrit Language

Deepak T. Mane
JSPM's Rajarshi Shahu College of Engineering

Study of Various Approaches in Machine Translation for Sanskrit Language

Prof. Deepak Mane[1], Aniket Hirve[2].

[1]Information Technology, Rajarshi Shahu College of Engineering, Pune, India; [2] Information Technology, Rajarshi Shahu College of Engineering, Pune, India;
Email: [1]dtmane@gmail.com, [2]anikethirve@gmail.com.

ABSTRACT

Machine translation is one of the most important applications of Natural Language Processing. Machine Translation is one of the most important branches of Artificial Intelligence. Artificial Intelligence is very useful in providing people with a machine, which understands diverse languages spoken around the world. Machine translation helps people from different places to understand an unknown language without the aid of a human translator. Machine translation translates the text from one language known as source language(SL) into the text of another language known as target language(TL).This paper gives a survey of the work done on various Indian machine translation systems either developed or under the development. Some systems are of general domain, but most of the systems have their own particular domains like particular languages and applications based on that.

Keywords : Sanskrit, Analysis, Machine Translation, Rule Based Model, English to Sanskrit Machine Translation, MT, Computational Linguistics, Morphology.

1 INTRODUCTION

India is a highly multilingual country and the largest democratic country in the world. Indian constitution recognizes eighteen languages with several hundred dialects. [1] Therefore, machine translation provides a solution in breaking the language barrier within the country. English is an important language both in India as well as the world. It is the de facto language for two key areas: administration and education. It is the main language for most of the information. Sanskrit acts as an intermediary language or Interlingua for translation to and from Indian languages. This is owing to the fact that Sanskrit is the mother of all Indian languages. It is an ancient language which is almost dead in today's time. But due to the cultural, literary and religious association of the language it is recognized in the Indian constitution of 1950. Sanskrit is still in use in some states of India like Haryana, Delhi, Rajasthan, Jammu and Kashmir and also in countries other than India such as UK, USA, UAE, Singapore, Kenya, Fiji, Bangladesh, Canada, and Malaysia [11]. The initial work on Machine Translation in India was performed at various locations like IIT Kanpur, NCST Mumbai, Computer and Information Science department of Hyderabad, CDAC Pune, Ministry of Communication, the Department of IT and IT Government of India. In the mid 90's and late 90's some more machine translation projects also started at IIT Bombay, IIT Hyderabad, department of computer science and Engineering Jadavpur University, Kolkata, JNU New Delhi etc.

A. General Terms

1) Statistical Machine Translation Technique (SBMT): The statistical machine translation (SMT) is a machine translation paradigm where translations are generated on the basis of statistical models whose parameters are derived from the analysis of bilingual text corpora [6]. The SMT is a corpus based approach. Here parallel corpora are maintained based on different probababilistic (i.e. translation and language models. The advantage of this is that the corpora can be maintained without any specific training. For translation from a source language to the target language, the corpora is maintained preserving the word order of the target language. [12]

2) Example-based machine translation (EBMT): Example based machine translation (EBMT) is a versatile, simple yet accurate approach for machine translation. The basic units of EBMT are sequences of words or phrases. The basic techniques involved can be categorized as the matching of input sentences with source examples; matching phrases from the database and the extraction of corresponding phrases and the extraction of corresponding translation (translation phrase) and the aggregation of the extracted phrases into correct translation sentences (of target language). [13]

3) Rule Based Machine Translation Technique (RBMT): The rule based model generates Sanskrit translation of a given input English sentence using rules that generate verbs and nouns for Sanskrit. The rule based approach mostly makes use of hand written transfer rules to the translation of substructures from the source language to target language, in this case from English to Sanskrit. The main advantage of rule based approach is the easy implementation and small memory requirement [8].

TABLE I

COMPARISON BETWEEN EBMT AND RBMT:

EBMT	RBMT	SBMT
Core process is the selection and extraction of TL elements corresponding to SL fragments.	Core process is the bilingual dictionaries and rules for converting SL structures into TL structures.	Core process is the 'translation model' taking SL words or phrases as input and producing TL words or phrases as output.
It is preceded by an 'analysis' stage for the decomposition of input sentences into appropriate fragments and their matching against SL fragments.	The preceding stage of analysis interprets input SL strings into appropriate 'translation units' (like canonical nouns and verb forms) and relations (like dependencies and syntactic units).	The preceding 'analysis' stage is represented by the process of matching individual words or word sequences of input SL text against entries in translation model.
Succeeding stage of synthesis adapts extracted TL fragments and combines them as output sentences.	Succeeding stage of 'synthesis' derives TL texts from TL representations produced by the core Process	Succeeding stage involves a 'language model' which synthesizes TL words as 'meaningful' TL sentences.
Robustness: Low; EBMT works on best match reasoning.	Robustness: High; works on exact match reasoning.	SBMT works on contextually matching surrounding more valuable in terms of understanding & memorizing.

2 Machine Translation Systems Till Date

2.1 DESIKA:

According to Ramanujan the morphological analysis of Sanskrit is the basic requirement for the processing of Sanskrit. The Nyaya (Logic), Vyakarana (Grammar) and Mimamsa (Vedic interpretation) is a suitable solution that covers syntactic, semantic and contextual analysis of Sanskrit sentence. In his paper, he describes a Paninian bases system called DESIKA which includes Vedic processing and shabda-bodha as well. There are three separate modules in the model: generation, analysis and reference.The user specifies the nominal or verbal class is generated by the user and the subsequent applicable rules are generated. In analysis, the syntactic identification and assignment of roles for every word is carried out using the Karaka-Vibhakti mappings. Morphological analyzers for Sanskrit are developed by Akshara Bharati Group at Indian Institute of Technology, Kanpur-India and University of Hyderabad-India. [3]

2.2 Sanskrit Karka Analyzer for Machine Translation:

In 2007, a Sanskrit Karka analyzer was designed by Sudhir K Mishra for making a translation tool for Sanskrit language at JNU Delhi. The methodology adopted was Rule Based MTS (RBMTS). But the system was not able to resolve the ambiguities and also not able to work on Sandhi and Samasa for Sanskrit language [5].

2.3 Constrained Based Parser for Sanskrit Language

1) ANGLABHARTI (1991): It is a machine-aided translation system specifically designed for translating English to other Indian languages. English is a Subject-Verb-Object (SVO) language while Indian languages are Subject-Object-Verb (SOV) and are relatively free of word-order. Instead of designing translators for English to every Indian language, Anglabharti employs a pseudo-interlingua approach which analyses English only once and creates an intermediate structure called PLIL (Pseudo Lingua for Indian Languages). This is the basic translation process which translates the English source language to PLIL with most of the disambiguation already performed. The PLIL structure is further converted to each Indian language through text-generation. The effort in analyzing the English sentences and translating them into PLIL is estimated to be about 70% and the text-generation accounts for the remaining 30%. Thus only with an additional 30% effort, a new English to Indian language translator can be built. [14]

2) ANGLABHARTI-II (2004): It addressed many of the shortcomings of the earlier architecture. It uses a generalized example-base (GEB) for hybridization besides a raw example-base (REB). During the development phase, when it was found that the modification in the rule-base was difficult and which might result in some unpredictable results, the example-base was grown interactively by augmenting it. At the time of the actual usage, the system first attempts to find a match in REB and GEB before finally invoking the rule-base. In AnglaBharti-II, provision was made for automated pre-editing and paraphrasing; generalized and conditional multi-word expressions, recognition of named-entities and also incorporated an error-analysis module and statistical language-model for automated post-editing. The purpose of automatic pre-editing module is to transform/paraphrase the input sentence to a form which can be translated easily. Automated pre-editing might even fragment an input sentence if the fragments can easily be translated and positioned in the final translation. Such fragmentation may be triggered in case of a failure of translation by the 'failure analysis' module. The failure analysis comprises of heuristics on speculating on the most likely causes of the failure. The entire system is pipelined with various sub-modules. All these have significant contributions to a greater accuracy and the general robustness to the system.

2.4 Etrans System

A string of English sentence can be translated into string of Sanskrit sentences. The method for the design and development is in the form of a software called "EtranS". The system comprises of user interface developed using .NET framework and the lexicon using MS-Access 2007.

The modules are developed on the basis of process engine e. The user interface is responsible for taking input and generating the output. It is heavily dependent on the tables created in the database for generating output and the programming done to extract the information based on the logic developed [6].

The software is comprised of the following modules:

1) *PARSE MODULE:*
- Input Module
- Sentence Analyzer Module
- Morphological Analysis Module
- Parse Module
- Parse Tree

2) *GENERATOR MODULE:*
- Mapping
- Output Module

2.5 Google Translater:

In 2007, Franz-Josef Ochs applied the statistical machine translation approach for Google Translate from English to Sanskrit Language and vice-versa. Hindi, Urdu and Sanskrit are the only Indian Languages present among the 57 Languages for which translation is provided. Accuracy of the system is good enough to understand the sentences after translation. The system provides both the robustness incase of failure and no loss of information while translating the text. The output of the system follows the grammar of the source language [4].

TABLE II
COMPARISON BETWEEN DESIKA, ANGLABHARATI, ETRANS & GOOGLE TRANSLATOR:

Basis	DESIKA	ANGLABHARTI	E TRANS	GOOGLE TRANSLATOR
Features	Desika is a Paninin grammar based analyss progrm.it includes Vedic processing and shabdabodha as well.	It analyses English only once and creates an intermediate structure called PLIL. The PLIL structure is then converted to each Indian language through a process of text-generation.	The Etrans system comprises of user interface developed using .NET framework and the lexicon using MS-Access 2007. The modules are developed on the basis of process engine.	Hindi, Urdu and Sanskrit are the only Indian Languages for which translation is provided.
Approach	Desika uses rule based approach.	Anglabharti uses a pseudo-interlingua approach.	Etrans uses complete framework of Rule Based approach.	Google Translator uses the statistical machine translation approach.
Advantages	Easy implementation and small memory requirement.	Significant contribution to greater accuracy and robustness to the system by including failure analysis module.	Translation of Simple and compound sentences from English to Sanskrit.	Robustness against failure and no loss of information while translation of the text.
Limitations	Rule Based approach not sufficient due to no morphological order of Sanskrit	Might produce unpredictable results initially.	Heavy dependency on databases.	Hindi, Urdu and Sanskrit only languages provided.
Future Scope	Can be combined with other approaches to make system faster and more robust.	Development of a more robust error analysis.	Process engine can be developed with the combination of a stronger rule base and other techniques.	Other languages can be included for the translation as target languages.

3 CONCLUSION

The current work is in the direction of building an English to Sanskrit Machine Translation system. Our aim has been to examine how the semantic information can be passed through the grammar resulting in appropriate Sanskrit form. We have shown some of the outputs for the simple sentences taken for translation. Our system handles English sentences of types: (i) simple subject, object and verb; (ii) subject, object, adverb and verb; (iii) subject, object, adjective and verb; (iv) subject, object, preposition and verb; (v) compound sentences.

REFERENCES

[1] Sinha, R.M.K. and Jain, Ajai, 'AnglaHindi: An English to Hindi Machine Translation System', MT Summit IX, New Orleans, USA, Sept. 23-27, 2003.

[2] Samantaray, S.D., 'Example based machine translation approach for Indian languages', In Proceedings of ICCS, pp 1-10, 2004.

[3] Ramanujan, P., 'Computer Processing Of Sanskrit', Computer Processing Of Asian Languages CALP-2, IIT Kanpur, India, 1992.

[4] Och, Franz Josef, Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning, Prague, Association for Computational Linguistics, pp. 858-867, June 2007, [Online] Available: http://www.translate.google.com.,http://translate.google.com/about/intl/en_ALL/.

[5] Sudhir K. Mishra,"Sanskrit Karaka Analyzer for Machine Translation", a Ph. D. Thesis, SCSS JNU New Delhi, 2007.

[6] Promila Bahadur, A.K Jain,D.S Chauhan, "EtranS-English to Sanskrit Machine Translation" ICWET 2012, Bombay, AC 2012.

[7] Mishra, Vimal and Mishra, R. B., '*Performance Evaluation of English to Sanskrit Machine Translation System: A Novel Approach*', In Proceedings of International Joint Conference on Information and Communication Technolog (IJCICT 2010),9th-10thJanuary,pp181-186,IIMT, Bhubaneswar,India,2010.

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

[9] S. Bandyopadhyay, "An Example Based MT System in News Items Domain from English to Indian Languages", Machine Translation Review 12, pp 7-10, 2001.

[10] [2008] Sanskrit Linguistic Resources,"http://sanskrit.inria.fr/DATA/XML/".

[11] D.T. Mane, P.R. Devale and S.D. Suryawanshi 'Rule based approach for English to Sanskrit machine translation and synthesizer system', In Oriental Journal of Computer Science & Technology.

[12] Uday C. Patkar, P. R. Devale, S. H. Patil, 'Transformation of multiple English text sentences to vocal Sanskrit using Rule-Based technique', International Journal of Computers and Distributed Systems, Vol. No.2, Issue 1, December 2012.

[13] Vimal Mishra and R. B. Mishra, 'Study of Example Based English to Sanskrit Machine Translation'.

[14] 'Advances in Machine Translation Systems'