International Journal of Computer Applications (0975 – 8887)
Volume 136 – No.1, February 2016

Sanskrit Machine Translation Systems: A Comparative Analysis

Jaideepsinh K. Raulji
Lecturer, Ahmedabad University, Ahmedabad, Gujarat, India.
Research Scholar, Dr. Babasaheb Ambedkar Open University, Ahmedabad, Gujarat, India

Jatinderkumar R. Saini, PhD
Professor & I/C Director, Narmada College of Computer Application, Bharuch, Gujarat, India
Research Supervisor, Dr. Babasaheb Ambedkar Open University, Ahmedabad, Gujarat, India

ABSTRACT
Machine Translation is area of research since six decades. It is gaining popularity since last decade due to better computational facilities available at personal computer systems. This paper presents different Machine Translation system where Sanskrit is involved as source, target or key support language. Researchers employ various techniques like Rule based, Corpus based, Direct for machine translation. The main aim to focus on Sanskrit in Machine Translation in this paper is to uncover the language suitability, its morphology and employ appropriate MT techniques.

Keywords
Sanskrit, Bilingual Dictionary, Interlingua, Machine Translation (MT), Natural Language Processing (NLP).

1. INTRODUCTION
Machine Translation melts the language barrier so that humans can transform information, share ideas, know each other cultures, technological discussions etc hence it is vital application of Natural Language Processing. Using MT one natural language can be translated to other. India is a multilingual country with as many as 22 scheduled languages of which Sanskrit is one of them and it is official language of state of Uttarakhand, India. It is considered as the oldest Indo-European language. It is sacred and philosophical language in Hinduism, Buddhism, and Jainism. The Sanskrit is mother of most Indian languages. Sanskrit works includes extensive epics, Vedas, Upanishads, philosophical, mathematical, scientific, dramatic, poetic texts. MT helps to unite the world socially, culturally and technologically. There is big requirement for inter-language translation for transfer of information and sharing of information and ideas.

2. APPROACHES TO MACHINE TRANSLATION
There are main 4 approaches to Machine Translation [1]. They are Direct, Rule Based, Corpus Based and Knowledge based. It is represented diagrammatically in Fig. 1.

In Direct MT [1] there is no intermediate representation of codes. Using bilingual dictionary there is word by word translation with help of bilingual dictionary followed by some syntactic rearrangement. This method of translation is only feasible for one language pair. It requires little analysis of text and without parsing. Here analysis method like morphological analysis, preposition handling, syntactic arrangement (so

as to reflect correct word order in target language), and morphological generation can be performed.

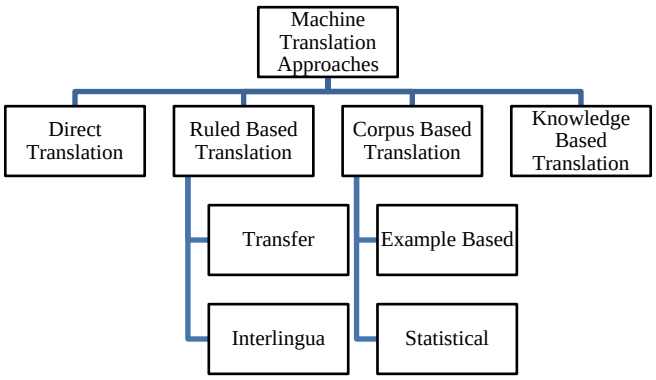

Fig. 1 Approaches to MT [1]

In Rule based MT [1], intermediate representation may be produced like parse tree. It rely on rules for morphology, syntax, lexical selection and transfer, semantic analysis and generation therefore known as Rule based. Rule based can be of 2 types as Transfer based and Interlingua. In Transfer based, SL to TL is without intermediate representation while in Interlingua some intermediate code representation is made through which SL is translated to TL via inter language codes.

In Corpus based Machine Translation [1] it requires sentence aligned parallel text for each language pair. It cannot be used for language pairs for which such corpora does not exist. It can be further classified into Statistical MT and Example Based MT.

In Knowledge based MT systems [1], semantic based approach to language analysis is introduced by Artificial Intelligence researchers. It requires large knowledge base that includes both ontological and lexical knowledge. The basic AI approaches includes semantic parsing, lexical decomposition into semantic networks and resolving ambiguities.

The drastic increment in computational power at personal computer systems and the increasing availability of written translated texts allowed the development of statistical and other corpus-based MT approaches also [1].

3. MACHINE TRANSLATION SYSTEMS IN WHICH THE SANSKRIT LANGUAGE IS INVOLVED.

3.1 English to Sanskrit Translator and Synthesizer (ETSTS).

ETSTS [2] [3] is a Rule based and Example based approach of MT. Using speech synthesizer as a plug in module it converts target sentence to speech output. The design of the system is modularized to Text Input, Grammar and spell check, Token Generator, Translator, Parser Generator Module, RBMT/EBMT engine, its bilingual database, Text output and a Waveform Generator.

Here bilingual dictionary is used for translation so that input sentence can be transformed into output sentence by carrying out the parse. Here they have used Stanford parser, then replaced source language words to target language words and finally rearrangement of words is carried out using grammatical rules of target language.

The important modules in the system are

3.1.1 Token generation process – Here sentence is decomposed to words applying space as delimiter. The resultant strings may be single word or compound word. Further by applying rules of English grammar, parse tree is generated to categorize words elements into noun, verb, noun phrase, verb phrase, etc.

3.1.2 Morphological Analysis process – This process takes the tokens as input and gathers grammatical information on that token.

3.1.3 Translator process – The actual translation is done using RBMT (Rule Based MT) and EBMT (Example Based MT) technique. It uses any one of these method. This module takes input as parse tree generated from Stanford parser. Using proper tagging methods of parser from RBMT or EBMT database the word is replaced to target language with proper rearrangement and finally presents the sentence in correct order.

3.2 ESSS (English Speech to Sanskrit Speech) using Rule based Translation [4][5].

It uses Rule based translation technique. After converting English speech to text, the sentences formed are first decomposed into words. The resultant words are matched in database, it also extracts Parts of Speech information using database to classify each word for noun, verb, adjective etc, then applying grammar rules, and Sanskrit text gets generated rearranging the sentence generating target language text. These text is given as input to wave form generator where it gets converted to Sanskrit speech.

3.3 E-tranS (English to Sanskrit Translator)

E-Trans [6] is rule based machine translation tool. It is primarily based on formulation of Synchronous Context Free Grammar (SCFG), a sub set of Context Free Grammar (CFG). Language representation of syntax is done using SCFG. The process engine developed works in two phases, the Top Down approach and Bottom to Top analysis.

After parsing and tokenization of input sentence, it matches information within dictionary for exact match word. Besides this it also gathers information related to parts of speech. It checks rules defined in file to form a target language sentence. Proper morphological analysis is done on target language sentence, Sanskrit words are aligned to get desired output. The result is quite promising for small and large sentences.

3.4 Sanskrit to English Translator [7] by Subramaniam A.

It performs *'Sandhi Viccheda'* of Sanskrit words and then translates them to English. It has two components – Morphological parser and translation generator. The main part of parser is a *'Viccheda'* module which applies reverse *'Sandhi'* rules to split the combined Sanskrit words into separate basic words. The translator is also mainly divided into two parts where first one structures the English language sentence according to grammar using parse information. The second generates equivalent English words according to the morphological details. Combining all these modules generates required translated text.

3.5 English to Sanskrit Machine Translation System by Mishra and Mishra [8] [9].

It implements Example Based MT technique.

The system is divided into different modules like

3.5.1 Sentence Tokenizer Module-The module split the English sentences into tokens (words)

3.5.2 POS Tagger Module –The POS (Part-of-Speech Tagging) is the process of assigning a part-of-speech (nouns, verbs, pronoun, preposition, adverb and adjective) to each word in a sentence.

3.5.3 GNP detection Module –This module detects the gender (G), Number (N) and Person (P) of the noun in English sentence.

3.5.4 Tense and Sentence detection module –Using rules Tenses of English language is detected.

3.5.5 Noun and object detection Module –This module returns noun for Sanskrit of the corresponding English noun.

3.5.6 Root *'Dhatu'* detection Module –This module gives verb for Sanskrit of the equivalent English verb. It uses ANN method for the selection of verb for Sanskrit.

3.5.7 Adverb Conversion Table –It maintains database of equivalent Sanskrit and English adverb.

3.6 English to Sanskrit SMT (Statistical Machine Translation) with Ubiquitous Application developed by Warhade S [10], et al.

It is Phrase-based Statistical machine translation system. It has following features/modules like Phrase translation probability, Inverse Phrase translation probability, Lexical weighting probability, Inverse lexical weighting probability, Phrase penalty, Language model probability, Distance based distortion model, Word penalty.

Translation examples from the respective bilingual text corpus are aligned in order to extract phrasal equivalences and to calculate the bilingual feature probabilities. Monolingual features like the language model probability are trained on mono lingual text corpora of the target language.

3.7 English to Sanskrit Machine Translation and Synthesizer system – A Rule Based approach by Mane D.T. [11], et al. –

It uses dictionary based MT technique. The input text sentence is converted into output sentence by carrying out the parsing, replacing source word with target language equivalents from bilingual dictionary and then using grammar rules of target language for re-arranging and aligning their order.

The system contained important modules like

3.7.1 Token Generator – This module splits the sentence into chunks of strings delimited by spaces. It is parsed using parser rules define in the system for English language and finally a parse tree is generated.

3.7.2 *'Vichcchceda'* Module – This module identifies root and complex words using sandhi rules.

3.7.3 Translator – It finds English words to Sanskrit words. Rearranging of Sanskrit words using grammar rules and does the actual translation.

3.7.4 Finally translated text is sent to speech synthesizer.

3.8 Sanskrit to Hindi MT system at JNU [12] –

The project is in development at Jawaharlal Nehru University (JNU) headed by Kulkarni A. It is focused in development of domains like children stories, building multimedia and e-learning contents for kids. The system will also undergo updating by incorporating modules like Word Sense Disambiguation module, Anaphora Resolution module, and Default Prose Order generator. The system will be extended for domains like Yoga and Ayurveda.

The Sankrit Machine Translation Systems are listed in Table-1.

Table – 1. Sanskrit MT Systems

MT System	Approach	Source-Target Language Pair	Features
ETSTS[2][3]	Rule and Example based	English to Sanskrit	Converts target sentence to speech output, Use of Bilingual dictionary, Modular design.
ESSS[4][5]	Rule based	English to Sanskrit	Converts English Speech to Sanskrit speech via English and Sanskrit words
E-tranS[6]	Rule based	English to Sanskrit	Formulation of Synchronous Context Free Grammar (SCFG), Lexicon used for Morphological analysis
Sanskrit to English Translator by Subramania m A.[7]	Rule based	Sanskrit to English	Focus on *Sandhi Vichheda* ,Morphological Analysis.
English to Sanskrit MT by Mishra and Mishra[8][9]	Example based	English to Sanskrit	POS tagger Module, Uses ANN for verb selection, GNP Module.
English to Sanskrit MT by Warhade S[10], et al	Statistical based	English to Sanskrit	Phrase based
English to Sanskrit MT by Mane D.T.[11] , et al	Rule based	English to Sanskrit	Use of bilingual dictionary and grammar rules file.
Sanskrit to Hindi MT by JNU[12].	Rule based	Sanskrit to Hindi	WSD module, Anaphora Resolution module.

4. SANSKRIT INVOLVED IN OTHER NATURAL LANGUAGE PROCESSING ACTIVITIES

4.1 Desika [13]

A Natural understanding system, a software package, developed by Indian Heritage Group, C-DAC, Bangaluru led by Ramanujan P. It can generate and analyze plain and accented written Sanskrit texts using grammar rules of Panini's *'Ashtadhyayi'* with database of *'Amarakosa'* and processing from *'Nyaya'* and *'Mimamsa Sastras'*. It can also analyze Vedic Text. It is general purpose Sanskrit parser which can identify the compound and combined word forms.

4.2 Sanskrit WordNet [14]

Sanskrit wordnet is based on idea of English WordNet. It is more than conventional Sanskrit dictionary. It gives different relations between synsets or synonym sets which represent unique concepts. It is developed by Bhattacharyya P., Centre for Indian Language Technology (CFILT), Computer Science and Engineering Department, IIT Bombay, Mumbai.

4.3 Morphological analysis of nominal inflections in Sanskrit[15][16]

It is an online system for *'Subanta'* analysis. Database for Sanskrit *'Subanta'* and verbs is built. Uses the *'Vibhakti'* information as well as the *'Subanta'* formulations of Panini and later grammarians to parse a text for *'Subanta'*. It can be used for MT from Sanskrit to other languages and understanding of Sanskrit words.

4.4 Dependency Parser for Sanskrit Language[17]

Dependency parser for Sanskrit Language uses deterministic finite automata (DFA) for morphological analysis and *'Utsarga Apavaada'* approach for relation analysis[18]. It uses *'Ashtadhayayu'* (a book of Sanskrit grammar) for its implementation. Parsing is the process of analyzing a string of symbols either in natural language according to rule of formal grammar[18]. The parser takes as input a Sanskrit sentence and using the Sanskrit rule base from a DFA analyzer. It analyzes each word of the sentence and returns the base form of each word along with their attributes[18].

4.5 *'Anusaaraka'* [17] [16]

The MT system which converts English to Hindi language has been architecture around Panini's *'Ashtadhyayi'* (Grammar rules on Sanskrit language). At present the system is applied to children's stories. The *'Anusaaraka'* MT approach mainly consists of two modules. The first module is known as Core *'Anusaaraka'*, which is based on language knowledge, and the second one is domain specific which is statistical based.

4.6 POS tagger for Sanskrit language developed by Chandrasekhar R. [19] at JNU.

4.7 Bilingual dictionaries are also developed like Sanskrit-Hindi dictionary developed by JNU under supervision of Jha G.N. [20]. Also Williams M. [21] Sanskrit-English dictionary, Apte's [22] Sanskrit-English dictionaries are available on web.

5. CONCLUSION

Though Sanskrit is considered as important language in Indo-European language family, still lot of work is required to explore the potential of this language to open vistas in computational linguistics domain. Machine Translation systems has been developed or in developing stage using Sanskrit as source or target language, but still some systems are particular to specific domain, confined to short sentences and phrases. Due to rich morphological nature of Sanskrit language, it use becomes challenging in Machine Translation application using Corpus based MT techniques.

There are system available which converts English to Sanskrit language using different translation techniques and are suitable for particular domain. The need for system which can cover all domains with acceptable accuracy is highly desired. Additionally work is also done on Sanskrit wordnet, dictionaries and POS taggers, morphological analyzers.

6. REFERENCES

[1] Siddiqui T. And Tiwary U.S., "Natural Language Processing and Information Retrieval", Oxford University press, 2008.

[2] Rathod S.G., "Machine Translation of Natural Language using different Approaches: ETSTS (English to Sanskrit Translator and Synthesizer)", International Journal of Computer Applications, Vol 102-No.15, September 2014.

[3] Rathod S.G., Sondur S., "English to Sanskrit Translator and Synthesizer (ETSTS)", International Journal of Emerging Technology and Advanced Engineering, Volume-2, Issue-12, December 2012.

[4] Shukla P., Shukla A., "English Speech to Sanskrit Speech (ESSS) using Rule using Rule Based Translation", International Journal of Computer Applications (0975-8887) Vol 92 – No. 10 ,Apr 2014.

[5] Shukla P., Shukla A., "A Framework of Translator from English Speech to Sanskrit Text", International Journal of Emerging Technology and Advanced Engineering. Vol 3, Issue 11, Nov 2013.

[6] Bahadur P., Jain A.K., Chauhan D.S., "Etrans – A complete framework for English to Sanskrit Machine Translation", International Journal of Advanced Computer Science and Applications (IJACSA) from International Conference and workshop on Emerging Trends in Technology, 2012.

[7] Aparna S. "Sanskrit to English Translator", Language in India, Vol 5:1 Jan 2005. Also available on website http://www.languageinindia.com/jan2005/aparnasanskrit dissertation1.html

[8] Mishra V., Mishra R. B., "Divergence patterns between English and Sanskrit Machine Translation", INFOCOMP Journal of Computer Science, Volume 8 (3), pp 112 – 138, 2009.

[9] Mishra V., Mishra R. B., "ANN and Rule Based Model for English to Sanskrit Machine Translation", INFOCOMP Journal of Computer Science, Volume 9 (1), pp 80-89, 2010.

[10] Warhade S. R., Devale P. R., Patil S. H., "English-to-Sanskrit Statistical Machine Translation with Ubiquitous Application", International Journal of Computer Applications, Volume 51 – No. 1. August 2012.

[11] Mane D. T., Devale P. R., Suryawanshi S.D., " A Design towards English to Sanskrit Machine Translation and Synthesizer system using Rule Base Approach", International Journal of Multidisciplinary Research and Advances in Engg. (IJMRAE), ISSN 0975-7074, Vol 2, No.11. pp 405-414 July 2010.

[12] "Sanskrit-Hindi JNU Sanskrit Hindi MT", Available on http://sanskrit.jnu.ac.in/shmt/index.jsp, visited November 2015.

[13] Desika (Natural Language Understanding System), http://tdil.mit.gov.in/download/Desika.htm.

[14] "Sanskrit Wordnet" Available on http://www.cfilt.iitb.ac.in/wordnet/webswn/english_version.php, Accessed November 2015.

[15] Subhash, Jha G.N. "Morphological analysis of nominal inflections in Sanskrit", Special Centre for Sanskrit Studies, Jawaharlal Nehru University, NewDelhi.

[16] "Information on Anusaaraka system", Available on http://anusaaraka.iiit.ac.in/, Accessed Nov 2015.

[17] Antony P.J., "Machine Translation Approaches and Survey for Indian Languages", Computational Linguistics and Chinese Language Processing. Vol 18, pp 47-78, March 2013.

[18] Shashank S., Raghav A., "Sanskrit as a Programming Language and Natural Language Processing", Global Journal of Management and Business Studies (2248-9878), Vol – 3, Number 10 (2013), pp. 1135-1142., Research India Publications.

[19] "Sanskrit POS Tagger" Available on http://sanskrit.jnu.ac.in/post/post.jsp, Accessed November 2015

[20] "Sanskrit-Hindi dictionary at JNU", Available on http://sanskrit.jnu.ac.in/student_projects/lexicon.jsp, Accessed November 2015

[21] "Monier Williams Sanskrit-English Dictionary" Available on http://www.sanskrit-lexicon.uni-koeln.de/scans/MWScan/2014/web/index.php, Accessed November 2015.

[22] "Apte's Sanskrit Dictionary" Available on http://www.sanskrit-lexicon.uni-koeln.de/aequery/index.html Accessed November 2015..