Selected text machine translator for English to Sinhala

B Hettige[1], AS Karunananda[2], and G Rzevski[3]

[1, 2 & 3]Department of Computational Mathematics, University of Moratuwa, Sri Lanka

[1] budditha@dscs.sjp.ac.lk, [2] asoka@itfac.mrt.ac.lk, [3] rzevski@gmail.com ([1]correspondence author)

Abstract—*Nowadays people are used to read English text online and sometimes prefer to see the meaning of the portions of text in his/her mother tongue. Some computer-based solutions enable the reader copy the desired text to a machine translator and to view the translation. This process is tedious and considerably disturbs the reading. Our research work presents an approach that enables reader to get Sinhala translation of selected English texts without disturbing the process of reading. This research emerges as an updated version of our existing English to Sinhala machine translation system, BEES. The system has been designed with three modules namely BEES client, BEES server and BEES translator. Upon highlighting the text, BEES client reads the selected text and sends it to the BEES sever. The BEES Server reads the text and provides appropriate Sinhala translation using the BEES translator. Three dictionaries namely, English dictionary, Sinhala dictionary, and English-Sinhala bilingual dictionary are used as the lexical resources of the system. The selected text translator of BEES has been tested through human support. Experimental results show that, this system successfully works with more than 80% accuracy.*

***Keywords*— Rule-based Machine Translation**

I. INTRODUCTION

Electronic documents are increasingly becoming popular as a new means of acquiring the world knowledge. Many of such documents are available on the Internet in English Language. However, it is a well-known fact that entire world community is not conversant in reading the documents published in English. Researchers have contributed to overcome this language barrier by introducing machine translation systems. Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another (Wikipedia).

As a solution for the language barrier, many countries all over the world introduced machine translation systems from English to their native languages. To develop machine translation system, numbers of approaches are available including Human-assisted, Rule-based, Statistical, Example-based, Knowledge-based, Hybrid and Agent-based (Hettige & Karunananda, 2011a). By using these approaches hundreds of machine translation systems are developed. Google translator(2013), Yahoo Babel fish(2013), and Apertium(2013) are some popular machine translation systems. In the region, Indians have also developed variety of machine translation systems, including Anusaaraka (Chaudhury and et.al. 2010) and MaTtra (Anantthkrishnan and et.al 2013). However, Sri Lanka is relatively late comer in this research area. There few number of systems have been developed for the English Sinhala machine translation. Among others, BEES (Hettige 2011a) is an English to Sinhala rule-based machine translation system, which runs through the concept of Varanegeema (conjugation) in Sinhala language (Hettige 2011b).

Considering the existing machine translation systems most of the systems are capable to translate given text (couple of sentences) or a web page. Therefore, user requires to give attention for the translator to input a text to the translator and read the translated text while reading the document. This unnecessary user attention imposes a barrier for the user to understand *the document.* The Selected text translation system overcomes this issue by providing the appropriate translation for the user selected text while reading the document without any disturbance.

This paper presents an updated version of the existing English to Sinhala machine translation system, BEES, that enables users to get Sinhala translation of the selected English texts while reading a document. The system has been designed to run on typical client-server architecture on the Internet. The BEES client reads the selected text from a document in any format (e.g. Word, PDF, and HTML) and sends to the BEES server, which translates the English text into Sinhala text.

The rest of the paper is organized as follows. Section 2 reports brief summary of the existing approaches and systems for the machine translation. The Section 3 gives design of the selected text translation system with brief description of each module. The section 4 presents evaluation methodology and finally section 5 gives conclusion and further works of the project.

II. EXISTING APPROACHES AND SYSTEMS

This section briefly describes existing machine translation approaches and systems. According to the approach, machine translation systems can be classified into seven categories, namely, Human-assisted, Rule-based, Statistical, Example-based, Knowledge-based, Hybrid and Agent-based.

Human-assisted approach uses human interaction for the pre editing, post editing intermediate editing stages. Therefore translation systems become semi-automated systems or expert systems. Most of the Indian families of machine translation systems including Anusaaraka, ManTra etc. use this approach.

Anusaaraka (2013) is a Human-assisted machine translation tool for the Indian languages that makes text in one Indian language accessible to another Indian language. Anusaaraka system gives (appears in layers) sequence of steps that follow each other till the final translation is displayed to the user. Using Anusaaraka modules number of machine translation systems has been designed for the number of Indian family of languages including Punjabi, Bengali, Telugu, Kannada and Marathi linto Hindi.

MaTra is a human-assisted transfer-based translation system for English to Hindi (Ananthakrishnan 2006). This System uses general-purpose lexicons and applied mainly in the domains of news. The MaTra project (2013) aims to produce understandable output for the wide coverage.

The Rule-based approach provides the translation using set of rules. Rules are used to source language analysis and target language generation through the source language morphological analysis source language syntax analysis, source to target translation, target language morphological generation and target language syntax generation. Further, these type of systems use minimum of three dictionaries namely a source language dictionary, a bilingual dictionary and a target language dictionary as the lexical resource.

There are number of machine translation systems that have been developed through the rule-based approach. Among others Apertium (2013) is a rule-based Machine Translation system, which translates related languages.

Statistical machine translation approach is by far the most widely-used machine translation method in the field of natural language processing. This approach tries to generate translations using statistical methods based on bilingual text corpora. Using this statistical approach, large numbers of machine translation systems have been developed.

Babel Fish (2013) is a web-based application developed by AltaVista which translates text or web pages from one language into another, Babel fish translates among English, Simplified Chinese, Traditional Chinese, Dutch, French, German, Greek, Italian, Japanese, Korean, Portuguese, Russian, and Spanish. A number of sites have sprung up that used the Babel Fish service to translate back and forth between one or more languages.

Bing Translator (2013) is a service provided by Microsoft as part of its Bing services which allow users to translate texts or entire web pages into different languages. All translation pairs are powered by Microsoft Translation, developed by Microsoft Research; it uses Microsoft's own syntax-based statistical machine translation technology.

Google Translator (2013) translates a section of text, or a webpage, into another language. It does not always deliver accurate translations and does not apply grammatical rules, since its algorithms are based on statistical analysis rather than traditional rule-based analysis.

The example-based machine translation system uses bilingual corpus with the parcel text for the machine translation. These systems are trained through the bilingual parallel copra, which contain sentence pairs. The example based approach is more useful for detecting the context from the text. Also this approach uses translation memories Using this approach number of machine translation systems have been developed all over the world. Among others, OpenMaTrExis (2013) one of the open source Example-based machine translation systems which is freely available on the OpenMaTrEx web site.

Knowledge-based machine translation approach uses knowledge for machine translation. This is an extended idea of the example-based machine translation. This approach uses linguistic and computational instructions, which are supplied by a human. Numbers of commercial quality Machine Translation systems have used this knowledge-based approach. EDR (Electronic Dictionary Research), by Japanese, is the most successful machine translation system. This system has taken a knowledge-based approach in which the translation process is supported by several dictionaries and a huge corpus (Toshio, 1995). While using the knowledge-based approach, EDR is governed by a process of statistical machine translation. As compared with other machine translation systems, EDR is more than a mere translation system but provides lots of related information.

The Hybrid machine translation system uses combine method in two or more translation approaches (rule-based and Statistical).

The Agent based approach is a modern approach for the machine translation. There are a number of NLP systems that have been developed using multi agent system technology. Most of these systems use agents to handle semantics in the translation.

There are few systems available for the English to Sinhala Machine Translation. As a first attempt Weersinghe and others have developed Sinhala to Tamil machine translation system through the corpus based approach (Weersinghe, 2010). They have designed translation tool named OpenTM, which is based on the translation memories. In 2003, Vithanage (2003) and others have developed English to Sinhala machine translation systems for weather forecasting domain. Vithanage's translation system can translate simple sentences and works on the limited set of words and the limited sentence patterns. In 2008, Fernando and others have developed English to Sinhala machine translation system using Artificial Neural Networks (Fernando, 2008).

Among others, BEES, an acronym for Bilingual Expert for English to Sinhala is English to Sinhala rule-based machine translation system that works through the concept of Varanegeema (conjugation) in Sinhala language. System has been developed with seven modules, namely, English Morphological Analyser, English Parser, English to Sinhala Base Word Translator, Sinhala Morphological Generator, Sinhala Parser, Transliteration module and Intermediate Editor. In addition to the above, system uses four lexical dictionaries namely, English dictionary, Sinhala dictionary, English-Sinhala Bilingual dictionary and Concept dictionary. The BEES successfully translates English sentences with simple or complex subjects and objects. The translation system successfully handles most commonly used patterns of the tenses including active and passive voice forms. Using the existing BEES Modules, English to Sinhala selected text translator has been developed.

III. DESIGN

English to Sinhala selected text translation system has been designed with three modules namely BEES Client, BEES Server and a BEES translator. Figure 1 shows the design diagram of the system. The BEES client runs on the client machine and reads the selected text while user is reading a document. After free processing, selected text has been send to the BEES server. The BEES server works as a web server that accepts the BEES client requests. The BEES server reads the selected text form its clients and sends it to the BEES translator to get appropriate Sinhala translation.

The BEES translator consists of 3 modules, namely English Language System, English to Sinhala Translation system and Sinhala Language system. The English Language system uses to analysis the selected English text and Sinhala language system uses to generate appropriate Sinhala text. To analyse the English text, English language system consists of two modules namely English Morphological analyser and the English parser. To generate the Sinhala text, the Sinhala language system consist of two modules namely Sinhala Syntax generator and the Sinhala Morphological generator. The English to Sinhala Translation system work as a word level translator that translates appropriate Sinhala words for the given English words. The English to Sinhala selected text, translator system uses three dictionaries namely English dictionary, the Sinhala dictionary and the English–Sinhala bilingual dictionary as the lexical resources. The present system comprises of more than 20000 English words and 50000 Sinhala words. Figure 2 shows the design of the BEES translator. Brief description of the each module is given below.

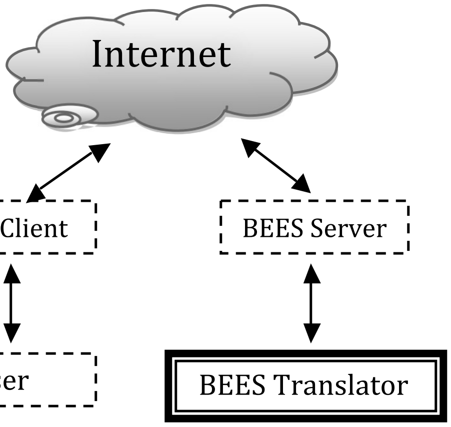


Fig 1. Top-level design of the system

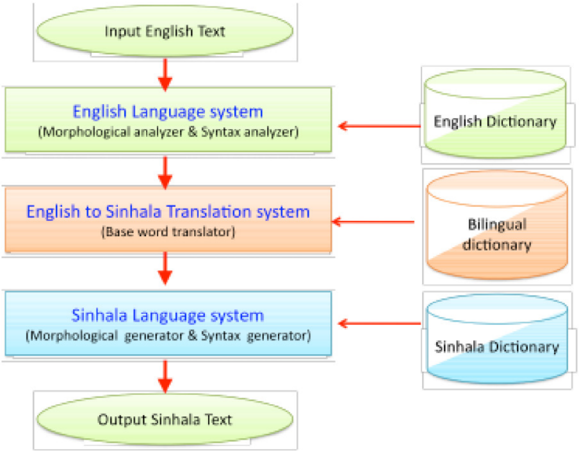


Fig 2. Design of the BEES translator

A. English Morphological Analyser

The English Morphological analyser reads a given English text, word by word and identifies morphological information for each word. The English Morphological analyser is a Prolog based system that can identify morphology of the given word. This Analyzer is capable to analyze all the English parts of speech with the irregular and regular word forms through the set of available grammatical rules. This system implements more than 50 essential rules to identify syntactic categories of English words.

B. The English parser

The English syntax analyser (English Parser) analyses the English sentence or a phrase and return the grammatical information such as subject, verb, complement etc. This parser has been implemented through the SWI-Prolog. To analyze the given English sentence it uses original English sentence and the result of the English morphological analysis.

C. The English to Sinhala Bilingual translator

The word level translator in BEES, translates English word into its appropriate Sinhala word with the support of the

English-Sinhala bilingual dictionary. English to Sinhala Bilingual translator is the prolog based module which used to get suitable Sinhala base-word for the given English base-word. The Bilingual translator uses output results of the English morphological analysis, output result of the English syntax analysis, and English-Sinhala-bilingual dictionary have been used to find the appropriate Sinhala base word.

D. The Sinhala morphological generator

The Sinhala Morphological generator generates appropriate Sinhala words for the given Sinhala grammar and makes the grammatically correct text. Sinhala Morphological generator implements more than 200 Sinhala morphological rules to generate a Sinhala word.

The Sinhala morphological generator is the key module of the system and it is implemented by using SWI-Prolog (2013). The Sinhala morphological analyser uses Sinhala dictionary and the result of the Bilingual translator.

E. The Sinhala syntax generator

The Sinhala syntax generator is used to generate Sinhala sentence or a part of a sentence according to the Sinhala syntax. This Sinhala sentence generator works as a Sinhala Parser and generates the grammatically correct sentence or phrase(s). The Sinhala syntax generator uses all the previous information for the sentence generation including Sinhala morphological generation, English sentence analysis, English morphological analysis and the English to Sinhala translation.

After the translation, BEES server sends the translated Sinhala text into the relevant BEES client.

IV. EVALUATION

The English to Sinhala selected text machine translation system has been evaluated through the human support. Translation system consists of three major modules namely BEES client, BEES server and the BEES translator. The system has been designed through the existing English to Sinhala machine translation system, BEES. All modules of the translator (English Morphological analyser, English syntax analyser, English to Sinhala word level translator, Sinhala Syntax generator and Sinhala Morphological generator) have already been tested (Hettige and Karunananda, 2010). Therefore only complete system has been tested through the human support. Same source (sample text file) is given for the five users and randomly sleeted text has been used to translate. Quality of the translation is tested through the knowledgeable people and ranks it as good, fair, accept, week or error for the each translation. The evaluation is done 50 times and number of words in the randomly selected text and result of the translation are recorded. The table 1 shows the results of the evaluation. Figure 2 shows the translation quality vs number of words available in the selection.

Number of words	Quality of the translation				
	Good	Fair	Accept	Week	Error
1	14	31	5	0	0
2	35	10	3	2	0
3	31	8	6	3	2
4	20	17	9	2	2
5	18	14	9	4	5
6	17	12	11	4	6

Table 1. Evaluation result

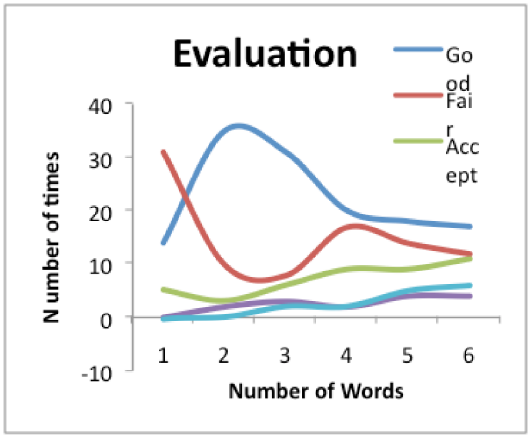


Fig 3. Evaluation of the Translation

Evaluation results show that, system gives more accuracy for the limited number of selected words. It also gives more accurate results for the multiple word (2 or 3) selection than a single word selection.

V. CONCLUSION AND FURTHER WORK

This paper provided a brief description of the English to Sinhala selected text translation system that enables users to get Sinhala translation of the selected English texts while reading a document. The paper briefly described existing approaches and systems for the machine translation, design of the selected text translation system and the evaluation methodology. The selected text translation system has been designed with three modules namely BEES client, BEES server, and the BEES translator. The client-server architecture is used to communicate and English to Sinhala translation is done through the BEES translator. Selected text translation system was tested with human support. The experimental results show that, BEES can be used to translate selected with on approximate accuracy of 80%.

Compared with Morphology and the syntax handling approaches, Semantic handling of the translation is the critical task in any machine translation. In addition, incomplete or incorrect text selection is one of the issues identified in the selection text translation. Rule-based top-down parsers are less efficient to handle incomplete sentences than the grammatically correct complete sentences. Therefore English parser and the Sinhala sentence generator in the translation system are needed to improve to handle incomplete sentences. As a solution for the above issue research has now been conducted to use the Multi-agent technology to improve the performance of BEES for handling the semantics in English to Sinhala translation.

REFERENCES

Anusaaraka (2013). http://anusaaraka.iiit.ac.in/ Accessed 28 July 2013.

Ananthakrishnan R et al. (2006). "MaTra: A Practical Approach to Fully-Automatic Indicative English-Hindi Machine Translation", In proc. MSPIL-06., 2006.

Apertium, (2013). Available: http://www.apertium.org/. Accessed 30 July 2013.

Bing (2013). < http://www.bing.com/translator>

Chaudhury S etal. (2010). "Anusaaraka: An Expert system based MT System", In proc. IEEE-NLPKE, China, 2010.

Fernando BTL et al. (2008). "English to Sinhala language Translator using Artificial Neural Networks", PSLIIT Vol2., SLIIT.

GoogleTranslator (2013). <http://translate.google.com>

Hettige B (2011). BEES homepage. [Online]., Available: http://www.dscs.sjp.ac.lk/~budditha/bees.html.

Hettige B (2011). "A Computational Grammar of Sinhala for English to Sinhala Machine Translation" M. Phil thesis, University of Moratuwa, Moratuwa, Sri Lanka.

Hettige B and Karunananda AS (2011). "Existing Systems and Approaches for Machine Translation: A Review", in Proc. the eight Annual Sessions Sri Lanka Association for Artificial Intelligence (SLAAI).

Hettige B and Karunananda AS, (2011), "Computational Model of Grammar for English to Sinhala Machine Translation",in Proc. ICTer2011, 2011, p 26 – 31.

Hettige B and Karunnanda AS (2010). "An Evaluation methodology for English to Sinhala machine translation", Proc. 6th International conference on Information and Automation for Sustainability (ICIAfS 2010), IEEE.

Hettige B and Karunananda AS (2008). "Web-based English to Sinhala Selected Texts Translation system",in Proc. the fifth Annual Sessions Sri Lanka Association for Artificial Intelligence (SLAAI).

KANT(2010). <http://www.lti.cs.cmu.edu/Research/Kant>

Moses (2013). <http://www.statmt.org/moses>

MaTra, <http://www.cdacmumbai.in/matra/> [Accessed 30 July 2013]

OpenMaTrEx, Available: http://www.openmatrex.org, [Accessed 30 July 2013]

Swi-Prolog (2013). <http://www.swi-prolog.org>

Toshio Y (1995). "The EDR electronic dictionary", Communications of the ACM, Volume 38, Issue 11, 1995.

Vithanage NVC, "English to Sinhala Intelligent Translator for Weather forecasting domain", Colombo: Thesis submitted BIT degree, University of Colombo, Sri Lanka, 2003.

Weerasinghe AR et al. (2010). "OpenTM: A Translation Memory System for Complex Script Languages", Proc. CLSA2010, Sri Lanaka, , p. 72-73.

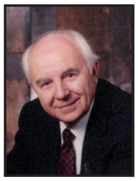
Wikipedia (2013). <http://en.wikipedia.org/wiki/Machine_translation>, Accessed 28 July 2013.

Yahoo Babel fish (2008). <http://babelfish.yahoo.com>

BIOGRAPHY OF AUTHORS

[1]B Hettige is a PhD Student of the Faculty of Information technology, University of Motratuwa, Sri Lanka. His research interests include Multi-agent technology, Machine translation and Sinhala Computing. He has produced more than 20 referred international and local publications to his credit.

[2]AS Karunananda is a Professor of Information Technology University of Motratuwa, Sri Lanka. At present he is the Dean of Research and Development of General Sir John Kotelawala Defence University. His research interests include Multi Agent Systems, Ontological Modelling, Machine Translation, and Theory of Computing.



[3]G Rzevski is a Visiting Professor in Multi-Agent Technology at Moratuwa University,
Sri Lanka and Professor of Complexity Science and Design at the Open University, UK. His research interests are in Applications of Complexity Science and Multi-Agent Technology to a variety of practical problems including morphological, syntactical and semantic processing.