Using Human-Assisted Machine Translation to overcome language barrier in Sri Lanka

B. Hettige[1], A. S. Karunananda[2]

[1]Department of Statistics and Computer Science, Faculty of Applied Science,
University of Sri Jayewardenepura, Sri Lanka.

[2]Faculty of Information Technology, University of Moratuwa, Sri Lanka.
budditha@yahoo.com[1] , asoka@itfac.mrt.ac.lk[2]

Abstract

Automated machine translation has faced with many issues regarding handling of semantics. In general, this issue can be addressed by computer-assisted machine translation at the pre-editing and post-editing stages. Our research has gone further and introduced an intermediate editing stage just before morphological analyzer of the target language. This approach detects semantic issues before hand and allows addressing those by human intervention. As a result the final translation will be more realistic and cut down the need for human intervention at the post-editing stage. The above approach has been used to develop English to Sinhala machine translation system. The system has been developed using Prolog and Java to run on a standard PC.

1. Introduction

Language barrier has been a major cause for not being able to disseminate world knowledge for rural communities those who do not use English as the mother tongue. The obvious solution to this issue is the use of modern computing technologies to translate from English to local languages. Many European and Asian countries have already taken steps to develop machine translation systems. In the Asian region, Indians have developed a variety of machine translation systems, including Mantra [5], Matra, Anusaaraka [2], Angalabharthi [4] and Angalahindi [3]. Among others, EDR [15] by Japanese is one of the most successful machine translation systems in the world. These translation systems use various approaches to machine translation, including, Human-Assisted Translation, Rule based Translation, Statistical Translation, Example-based and Knowledge-based Translation. However, due to various reasons associated with complexity of languages, for more than last fifty five years, Machine Translation (MT) has been identified as one of the least achieved area in computing. Most of these issues are associated with semantic handling in MT systems. Obviously, one approach to solve this issue is the use of post-editing[12] by humans. However, we argue that tedious work at the post-editing stage can be reduced by introducing an intermediate-editing stage just before running the morphological analyzer of the target language. This is mainly because; the intermediate-editing can ensure that only the appropriate words of the target language are sent forward.

With the above philosophy, we have been working on the development of English to Sinhala machine translation system. As per this system, we have already developed the Sinhala parser [6], Sinhala morphological analyzer [7], Transliteration module [9] and seven dictionaries [10]. The Sinhala parser and morphological analyzer have been tested through various applications such as Sinhala Chatbot [8]. This paper presents our approach to intermediate-editing as an expansion to the on going machine translation project. The system has been developed using Prolog [13] and Java to run on a standard PC.

The rest of this paper is organized as follows. Section 2 describes the overview of some existing Machine Translation systems. Section 3 describes design of the Human-Assisted English to Sinhala Machine Translation System. Section 4 presents how system works. Finally, Section 5 concludes the paper with a note on further work.

2. Some Existing MT Systems

Machine Translation (MT) is a translation process that translates one natural language into another [12]. In general, any machine translation system contains a source language morphological analyzer, a source language parser, translator, target language morphological analyzer, target language parser and several lexicon dictionaries. Source language Morphological analyzer analyzes a Source language word and provides Morphological information. Source language parser is a syntax analyzer that analyzes source language sentence. Translator is used to translate a source language word into target language. Target language Morphological analyzer works as a generator and it generates appropriate target language words for given grammatical information. Also target language parser works as a composer and it composes a suitable target language sentence. Further more, any Machine Translation system needs minimum of three dictionaries such as the source language dictionary, the bilingual dictionary and the target language dictionary. Source language morphological analyzer needs a source language dictionary for Morphological analysis. Bilingual dictionary is used by the Translator for translating source language into target language; and the target language morphological generator uses the target language dictionary to generate target language words. Regarding English to Sinhala machine translation point of view, the Machine Translation system needs an English dictionary, an English-Sinhala bilingual dictionary and a Sinhala dictionary.

Machine translation (MT) is a complex and a difficult task. However a large number of MT Systems have been developed for many languages all over the world. Sinhala is an Indo Arian language and Some Indian languages such as Pali, Sanskrith and Tamil are closer to Sinhala language. Therefore we need to study some existing MT systems especially the ones developed for Indian languages. At present Indians has developed a variety of machine translation systems. Below is a brief description on them.

The Anusaaraka [2] is a popular machine-aided translation system for Indian languages that makes text in one Indian language accessible to another Indian language. Also this System uses Paninian Grammar (PG) model [1] to its language analysis. The Anusaaraka project has been developed to translate Punjabi, Bengali, Telugu, Kannada and Marathi language into Hindi. The approach and lexicon is general, but the system has mainly been applied for children's stories.

MaTra [5] is yet another Human-Assisted translation system for translating English to Indian languages. This approach uses a 'tag' system to represent grammatical information of the language at hand. MaTra has been developed for the domain of gazette notifications pertaining to government appointments [5].

Angalabharti [4] is also human-aided machine translation system used in India. Since India has many languages, there are a variety of machine translation systems. For example, Angalahindi [3] translates English to Hindi using machine-aided translation methodology. Human-aided machine translation approach is a common feature of most Indian MT systems. In addition, these systems also use the concepts of both pre-editing [15] and post-editing[12 as the means of human intervention in the machine translation system.

Among others, Electronic Dictionary Research (EDR) [16] is the most successful machine translation system. This system has taken a knowledge-base approach in which the translation process is supported by several dictionaries and a huge corpus. While using the knowledge-based approach, EDR is governed by a process of statistical MT. As compared with other MT systems, EDR is more than a mere translation system but provides lots of related information. This ability has been possible since EDR has a huge reservoir of knowledge in the form of corpus and dictionaries. For instance, EDR has a word dictionary, concept classification dictionary; concept description dictionary; co-occurrence dictionary and bilingual dictionary. Due its rich reservoir of linguistic knowledge EDR has become a considerably automatic translation system. Therefore, in comparison with most Indian MT systems, human-aided machine translation is not necessarily encouraged in EDR. It is evident from the discussion and human-aided machine translation is more practical to consider for MT projects which are at their early stages. Therefore, the proposed English to Sinhala MT system has also taken the approach of human-assisted translation. However, we go beyond pre-editing and post-editing, and introduce an intermediate-editing stage to a MT system.

3. Design of English to Sinhala MT System

We have designed the proposed English to Sinhala MT system with seven modules, namely, English Morphological analyzer, English parser, Translator, Sinhala Morphological analyzer, Sinhala Parser, Transliteration module and Lexical dictionaries. Figure 1 shows Design of the English to Sinhala MT System. Brief descriptions of each component are given bellow.

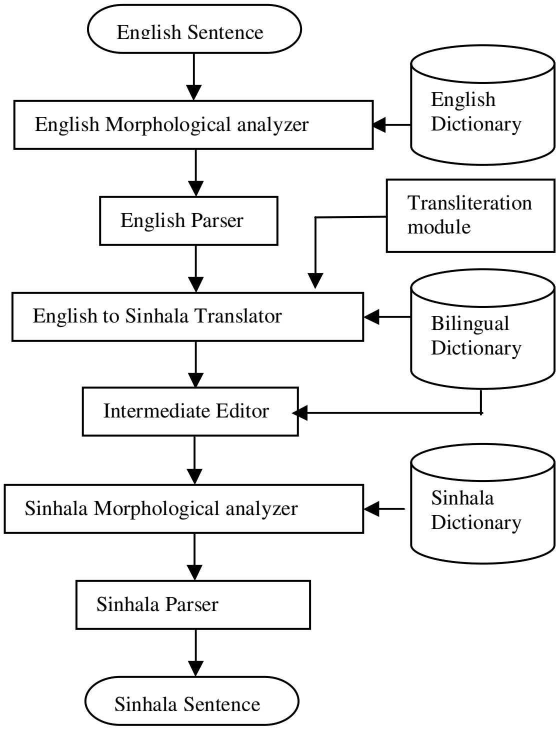


Fig 1: Design of the English to Sinhala Machine Translation System

3.1 English Morphological analyzer

English Morphological analyzer reads a given English sentence word by word and identifies morphological information for each word. There are many Morphological analyzers available for English language. Therefore, in this development, we have customized an existing English morphological analyzer. At this stage of the project, we assume that the sentences input to the MT system, has no spelling and grammatical mistakes. As such we can use a simple morphological analyzer for the English language at this stage. The morphological analyzer in our MT system has linked up with an English dictionary to get grammatical information of the words in the input sentence. SWI-Prolog has been used to implement this morphological analyzer.

3.2 English Parser

English parser receives source English sentences and tokens from English Morphological analyzer. This parser works as a syntax analyzer. Since there are many English parsers, we have customized an existing parser for our purpose. The current version of the parser used in our MT system mainly concerns only about the simple sentences. The parser has also been implemented using SWI-PROLOG.

3.3 Translator

Translator is used to translate English base word into Sinhala base word with the help of bilingual dictionary. This translator is a simple one and it does not automatically handle semantic of sentences. We argue that this stage can be supported by human intervention to generate the most appropriate translation for some words in a sentence. As such handling semantic, pragmatic and Multiword expressions must be addressed with the support from humans, for which we introduce and intermediate-editor.

3.4 Intermediate-Editor

English to Sinhala Human-Assisted Machine Translation system uses Intermediate Editor to handle ambiguities in semantic, pragmatic and Multiword expressions before proceeding to Sinhala linguistic modules in the MT system. Intermediate Editing facility is provided as a human interface for the MT system. This editor provides with facilities such as showing synonyms, anti-synonyms, related words, etc. The intermediate-editor is linked up both English and Sinhala dictionaries in the MT system. The process of intermediate-editing, before composing a Sinhala sentence, drastically reduces computational cost for running Sinhala morphological analyzer and parser. In addition, requirement for post-editing can be reduced by the process of intermediate editing. On the other hand, intermediate-editing can be used as means of continuous capturing of human expertise for machine translation. This knowledge can be reused for subsequent translations. As such the concept of intermediate-editing can be introduced as an approach to automatic knowledge management in a MT system. It should be noted that the knowledge used for pre-editing and post-editing cannot be readily captured by a MT system, as these process can be done even out side MT system. In contrast, intermediate-editing will be a integral part of the MT system, in which human directly interact with the system. The intermediate-editor of the MT system is a Java-based implementation.

3.5 Sinhala Morphological Analyzer

The Sinhala Morphological analyzer [7] works as a Morphological generator. This Morphological analyzer reads the words from Translator (as improved by a human when necessary) word by word. For each word, the morphological analyzer generates the appropriate word with full grammatical information such as nama (nouns), kriya (verb) and nipatha (preposition) in Sinhala language. This morphological analyzer works with the help of three dictionaries, namely, Sinhala Rule dictionary, Sinhala Word dictionary and Sinhala Concepts dictionary. All these databases and the morphological analyzer are implemented using Prolog.

3.6 Sinhala Parser

The Sinhala parser [6] works as a Sentence composer. It receives tokenized words from the morphological analyzer and composes grammatically correct Sinhala sentence. In generally, a Sinhala sentence contains 5 components, namely Ukktha vishashana (adjunct of subject), Ukkthya (Subject), karma vishashanaya (attributive adjunct of object), karmaya (object) and akkyanaya [8]. These five components of a Sinhala sentence are the building blocks for design and implementation of a Sinhala parser. The parser is also one of the key modules of this Human-Assisted English to Sinhala Machine Translation System and it is also implemented using SWI-PROLOG.

3.7 Dictionaries

Translation system uses six dictionaries such as English word dictionary, English concepts dictionary, English-Sinhala bilingual dictionary, Sinhala word dictionary, Sinhala rule dictionary and Sinhala concept dictionary [10]. English word dictionary contains English words and the lexical information. English concept dictionary contains synonyms, anti-synonyms and general knowledge about English words. English to Sinhala bilingual dictionary is used to identify appropriate Sinhala base word for a given English word and it contains relation between English and Sinhala words. Sinhala word dictionary stores Sinhala regular base words and lexical information. Same as English dictionary, Sinhala Concept dictionary stores Symantec information. The Sinhala rule dictionary stores rules required to generate various word forms. These are the inflection rules for formation of various forms of verbs and nouns from their base words. The rule dictionary also stores vowels, consonants, upasarga (prefix) and vibakthi (postfix).

3.8 Transliteration module

MT system needs to solve Out-of-vocabulary problems and handle technical terms. Machine transliteration can be used as a resalable solution for that. Transliteration is the practice of transcribing a word or text written in one writing system into another writing system [12]. In other words, Machine transliteration is a method for automatic conversion of words in one language into phonetically equivalent ones in another language. At present we have developed two types of transliteration models. One of these models transliterates Original English text into Sinhala Transliteration and the other transliterate Sinhala words that are written in English which transliterate into Sinhala. Finite State transducers are used to develop these two modules [9]

4. How System works

In this section we describe how translation system works for a given input sentence. For example, assume that the system reads “Saman eats red rice for his lunch” as the input sentence. Then the English Morphological analyzer identifies each word and returns the following Prolog predicates.

unknown([un001], ‘Saman’).

everb([ev01], ‘eats’).

eadj([ea01], ‘red’).

enoun([en02], ‘rice’).

eprep([ep01], ‘for’).

epnoun([en03], ‘his’).

enoun([en04], ‘lunch’).

Now the English Parser reads the original English sentence together with the output of the Morphological analyzer. After this analysis, the parser returns the following information.

subject ([un001]).

object([ea01, en02]).

objectp([ep01, en03, en04]).

verb([ev01]).

Tokenized ID of English words are then forwarded to the translator. The translator identifies Sinhala base word for each English word in the sentence, with the help of bilingual dictionary. It should be noted that, the first word ‘Saman’ is an unknown word in the dictionary. Therefore, it is out-of-vocabulary and translator cannot translate the word. As a result, the translator uses Transliteration module to get an appropriate Sinhala Transliteration. Then the output of the translator is as follows.

spronoun(un001, ‘සමන්’).

sverb([sv01], ‘කනවා’).

snoun([sn02], ‘රතු’).

snoun([sn03], ‘බත්).

snoun([sp01], ‘සඳහා’).

snoun([sn04], ‘ඔහුගේ’).

snoun([sn05], ‘දිවා ආහාරය’).

At this point human can change the above translated output by using Intermediate-editor. For example, the English word ‘rice’ contains several Sinhala meanings such as ‘ගොයම්’, ‘සහල්’, ‘වී’, ‘බත්’, ‘හැල්’ etc. Now human can select the most suitable Sinhala word for the word ‘rice’. Also the English word ‘for’ has several meanings such as ‘ට’, ‘සඳහා’, ‘වෙනුවට’, ‘ගැන’ ‘නිසා’, ‘පිණිස’ etc. Through human intervention the word ‘සඳහා’ can be selected. After that, the Sinhala Morphological analyzer reads all these words and generates appropriate Sinhala words with grammatical information. Output of the Sinhala Morphological analyzer is as follows.

snoun([sn01], ,3,1,3,0,1,v1, ‘සමන්’).

sadjn([sn04], ‘දිවා’).

snoun([sn01], 3,1,3,0,1,v2, ‘ආහාරය ’).

sprep([sn03], ‘සඳහා’).

sadjn([sn04], ‘රතු’).

snoun([sn04], 3,1,3,0,1,v5, ‘බත්’,…).

sverb([sv01] 3,1,3,0,1,1, ‘කයි’).

All the above information are sent forward and reserved by Sinhala parser. The Sinhala Parser identifies the following linguistic details in Sinhala language and it generates appropriate Sinhala Sentence.

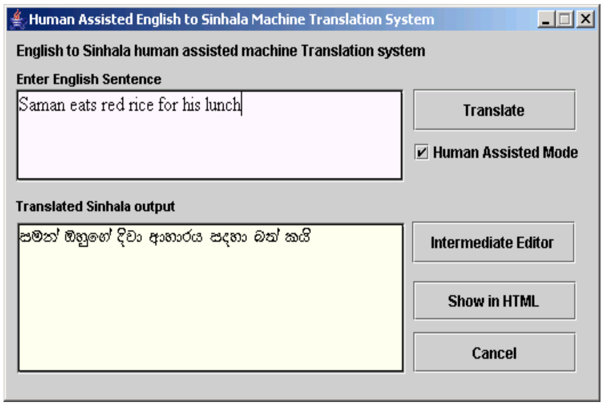
Subject (‘සමන්‘).

Object (රතු බත්).

Objectp(‘දිවා ආහාරය සඳහා’).

Verb(කයි).

As the last step of the translation process, Sinhala parser composes the corresponding Sinhala sentence ‘සමන් දිවා ආහාරය සඳහා රතු බත් කයි’. How System translate the given sentence, by using Intermediate editor is described bellow. Figure 2 shows a user Interface of the English to Sinhala Machine Translation System.

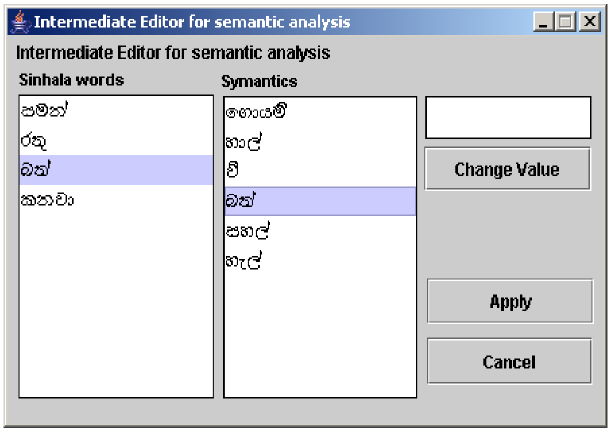

Human Assisted English to Sinhala Machine Translation System
English to Sinhala human assisted machine Translation system
Enter English Sentence
Translate
☑ Human Assisted Mode
Translated Sinhala output
Intermediate Editor
Show in HTML
Cancel

Saman eats red rice for his lunch

සමන් ඔහුගේ දිවා ආහාරය සඳහා බත් කයි

Fig. 2: User Interface of the English to Sinhala Machine Translation system

After Starting the Translation process the system automatically shows Intermediate Editor for selecting the suitable words. Using this editor, assistant can easily select the most suitable word. Figure 3 shows Intermediate Editor.


Intermediate Editor for semantic analysis
Intermediate Editor for semantic analysis
Sinhala words
Symantics
Change Value
Apply
Cancel

Sinhala words	Symantics
සමන්	ගොයම්
රතු	හාල්
බත්	වී
කනවා	බත්
	සහල්
	හැල්

Fig 3: Intermediate Editor for Symantec analysis

It shows numbers of synonyms for the word 'බත්'. Note that, Sinhala Morphological analyzer needs appropriate Sinhala base word and all the required grammatical information to generate Sinhala words [8]. By using this intermediate editor the user can select appropriate base word. However, some grammatical information does not need to change. These information are automatically generated by the previous stages such as Translation, Sentence and word analysis.

5. Conclusion and Further Works

Our objective of this project was to develop English to Sinhala Translation System. In this sense, we have designed and developed a Human-Assisted English to Sinhala translation system with a particular emphasis on an intermediate-editing through human intervention. This approach has reduced the workload at the post-editing stage of human-assisted machine translation. In addition it brings higher level of accuracy towards a meaningful translation.

Improvements and expansions to dictionaries will be an essential further work of this project. In addition, we intend to develop a capacity for the system to learn from intermediate-editing results to enable evolution of the translation system towards an automated system.

6. References

[1] Askhar B, Chaitanya V, Sangal R, "Natural Language Processing: A Paninian Perspective", Prentice Hall of India, New Delhi, India, 1995.

[2] Bharathi A, Chaitanya V, Kulkarni A. P, Sangal R., "Anusaaraka: Overcoming language barrier in India", to appear in "Anuvad: Approaches to Translation", Rukmini Bhaya Nair, (editor), Sage, New Delhi, 2001.

[3] Sinha R.M.K, Jain A., "AnglaHindi: an English to Hindi machine-aided translation system", MT Summit IX, New Orleans, USA, 23-27 September 2003; pp.494-497.

[4] Sinha R.M.K, "Integrating CAT and MT in AnglaBhart-II architecture", 10th EAMT conference, May. 2005, pp. 235-244.

[5] Durgesh R., "Machine Translation in India: A Brief Survey", National Centre for Software Technology, Mumbai, India. http://www.elda.org/en/proj/scalla/scalla2001/scalla2001Rao.pdf

[6] Hettige B., Karunananda A. S., "A Parser for Sinhala Language – First Step Towards English to Sihala Machine Translation", To appear in the proceedings of International Conference on Industrial and Information Systems(ICIIS2006), IEEE, Sri Lanka, 2006.

[7] Hettige B., Karunananda A. S., “A Morphological analyzer to enable English to Sinhala Machine Translation”, Proceedings of the 2nd International Conference on Information and Automation (ICIA2006), Colombo, Sri Lanka, 2006 pp. 21-26.

[8] Hettige B., Karunananda A. S., "First Sinhala chatbot in action", Proceedings of the 3rd Annual Sessions of Sri Lanka Association for Artificial Intelligence(SLAAI), University of Moratuwa, 2006.

[9] Hettige B., Karunananda A. S., “Transliteration System for English to Sinhala Machine Translation”, proceedings of second International Conference on Industrial and Information Systems(ICIIS2007), IEEE, Sri Lanka, 2007.

[10]Hettige B., Karunananda A. S., “Developing Lexicon Databases for English to Sinhala Machine Translation”, proceedings of second International Conference on Industrial and Information Systems(ICIIS2007), IEEE, Sri Lanka, 2007.

[11] A. M. Gunasekara, A Comprehensive Grammar of the Sinhalese Language”, Asian Educational Services, New Delhi, Madras, India., 1999.

[12] Jeff L., Christopher H, “Toward the development of a post-editing module for Machine Translation raw output”, Presented at the Third International Controlled Language Applications Workshop (CLAW2000), Washington, 2000.

[13] wikipedia, the free encyclopedia, http://en.wikipedia.org.

[14] SWI-PROLOG home page; http://www.swi-prolog.org

[15] Toshio Y., “Improvement of Translation Quality of English Newspaper Headlines by Automatic Preediting”, in Proceedings of the MT Summit VII, 1999.

[16] Toshio Y, “The EDR electronic dictionary”, Communications of the ACM, Volume 38, Issue 11 (Nov. 1995), pp. 42 – 44.