Application of Transfer Based Machine Translation from Sinhala to English

D. I. De Silva, P. K. D. A. Alahakoon, P. V. I. Udayangani, D. Kolonnage, M. H. P. Perera and S. Thelijjagoda
Sri Lanka Institute of Information Technology, Malabe

This research concentrates on a transfer based machine translation system that is capable of translating a grammatically correct Sinhala sentence in to its corresponding English sentence. In addition to the translator this system comes with features such as an inbuilt dictionary, a grammar tool, a Sinhalese grammar checker, an add word tool, and a debugging tool. This approach, Sinhala to English translation has never been addressed before. As the majority of the worldwide population carry out their day-to-day work in English, it has become a necessary language that people should learn. In addressing this need, the challenge of building a Sinhala to English language translator was taken up. At present, the system has achieved a success rate of 75% with a corpus of 150 grammatically well-balanced sentences.

***Index Terms*—English, Sinhala language, transfer based machine translation, translator.**

I. Introduction

Sinhalese is the mother tongue of Sri Lanka. According to ancient records, Sinhalese is the oldest language in the world, spoken by nearly sixteen million people, amongst whom nearly thirteen million are native speakers. It has originated from the Indo-Aryan group of languages, which in turn is a branch of Indo European languages. The origin of Sinhalese is from Sanskirt.

English, on the other hand is a West German language originating from England. Although it belongs to the Indo-European language family, it has much less in common with the Sinhalese language. A working knowledge of English has become a requirement in a number of fields, occupations and professions such as medicine, chemistry and consequently over a billion people speak English at least at a basic level.

Lack of knowledge in the English language has caused many Sri Lankans difficulties at job interviews, appointments of personnel for various positions and facing examinations in English etc. In addition, foreigners also face difficulties when trying to understand and communicate in Sinhalese when they visit Sri Lanka. To overcome such barriers this language translator was thought of.

This paper introduces a pilot machine translation system that translates grammatically correct Sinhalese sentences into grammatically correct English sentences in a consistent, quick, accurate and precise manner along with several other features to facilitate the users.

To perform a translation of non-similar languages is a far more difficult task than similar languages. For example, translation from English to French or German is easier than translation from Sinhalese to English, as English, French and German are closely related languages, and are based on a similar alphabet.

There are three main translation approaches available, namely Rule-based, Statistical and Example-based. This research focused on the Transfer-based machine translation approach, which is a Rule-based approach. By using this translation strategy, it is possible to obtain high quality translations, with accuracy within the region of about 90% (although this is highly dependent on the languages being considered).

Although a considerable number of Machine Translation projects have been carried out in Sri Lanka over the past decade, most have attempted to translate from English to Sinhalese. Translations from Sinhalese to English have received less attention, with no successful systems having been built until now, although there are existing systems that provide dictionaries.

All the three “persons” and all the tenses in the Sinhala language were considered when developing the system. There are thirty-two grammatical regulations in the Sinhalese Language to construct a grammatically correct Sinhalese sentence(s), out of that twenty are for active voice, and twelve for passive voice. Shown below are some example sentences from the active voice, which cover the grammatical rules of Sinhalese:

මම ගෙදර ගියෙමි.
නුඹ පාසලට යන්නෙහි.
මිනිස්සු වැඩ කරති.

This system handles sixteen out of the twenty regulations for the active voice. It has also been designed to handle sentences with adjectives, adverbs and nouns inclusive of days, months etc.

II. Methodology

As the first step in the translation process, the translation engine takes in the input text that may be a sentence or paragraph. Each sentence should be grammatically correct and separated with a full stop (.) or a question mark (?). This is essential for the translation engine to produce accurate sentences.

The translation engine then analyses the input text. Once the system notices a full stop or a question mark, it considers the text up to that full stop or question mark, as one sentence, and each word of that sentence is added to a sentence object, which is then added to an ArrayList. This procedure continues until all the sentences are included in to the ArrayList. If there is a space, comma, colon, semicolon, end of a double quote, or single quote then the translation engine identifies it as the end of a particular word.

Fig. 1 illustrates the steps carried out by the translation engine when converting the sentence සුදු ළමයා වේග යෙන් පාසලට දුවයි from Sinhalese to English.

Before adding a word to the data table (G in Fig. 1) the translation engine first checks whether the word is in English or Sinhala. If it is in English then the translation engine considers that word as an animate noun and adds it to the

Data table. If it is in Sinhalese, the translation engine first converts it into “Singlish” (C in Fig. 1). Then it will be stemmed (a process of removing one character at a time from the end of the Singlish word) until it is found in the database. When it finds that word, it will add the information of that word such as whether it is a noun/verb, animate/inanimate, person, sex, etc to the Data table. Eventually for each sentence there will be one table, in which each row will contain details of a particular word.

The translation engine then joins words (H in Fig. 1) as necessary – for example joining an adjective with the noun after it, or joining an adverb with the word after it. When joining certain types of words there will be no interchanging of the two words – for example, when displaying the output in the case of an adjective. In other cases, for example when an adverb is present, there will be an interchanging of the two words to comply with English grammar rules. Likewise the translation engine joins any words, other than nouns or verbs, such as prepositions, ownership nouns, definite plurals etc accordingly. The postfix of the final word – i.e. – the final letters of the Singlish word that are not included in the database – is extracted and used to determine the context of the sentence (I in Fig. 1). After that, if there are any objects in the sentence they will be formatted accordingly. eg: if the object is gedaraTA, in the output “to” will be added in front of home because the Sinhala word ends with”TA” likewise when the Sinhala word ends with “nTA”,”uTA”,“gee”“ee”, ”ehi”, ”ka”, “hi”, “sin”, “kin”, ”gen”, ”en” ,”n”, ”nak”, ”k”, ”aava”, ”uva”, ”nva”, ”vala”, ”aa” the object will be formatted accordingly (J in Fig. 1). Table I illustrates different English words that are used when translating the above postfixes of Sinhala words.

TABLEI ENGLISH WORDS FOR POSTFIXES OF SINHALA WORDS

Sinhalese	Singlish	English
ක්	k	a / an
කින්	kin	by a / by an
නක්	nak	on a / on an
උව	uva	a / an
එන්	en	from a/ from an
සින්	sin	from
න්	n	from
ගෙන්	gen	from the
වල	vala	in
එහි	ehi	in a
ක	ka	in a
ඒ	ee	on/ on the
ආ	aa	on
හි	hi	on
ට	TA	to
ගේ	gee	of / of the
ආව	aava	the
එන්ව	nva	the
න්ට	nTA	the
උට	uTA	a/an

The translation engine then makes the string pattern for the Sinhalese sentence, which it will then use to make the English string pattern of the sentence (k in Fig. 1).

The subject of the sentence is formatted next (L in Fig. 1) and will be added to a string. Next, the verb will be formatted according to the tense and the voice (active or passive) and once it is formatted it will be included in the same string that the subject was placed, according to the sentence pattern (M in Fig. 1).

Finally, if there are any more words other than the subject and verbs those will be put in to the same above mentioned string according to the sentence pattern (N in Fig. 1).

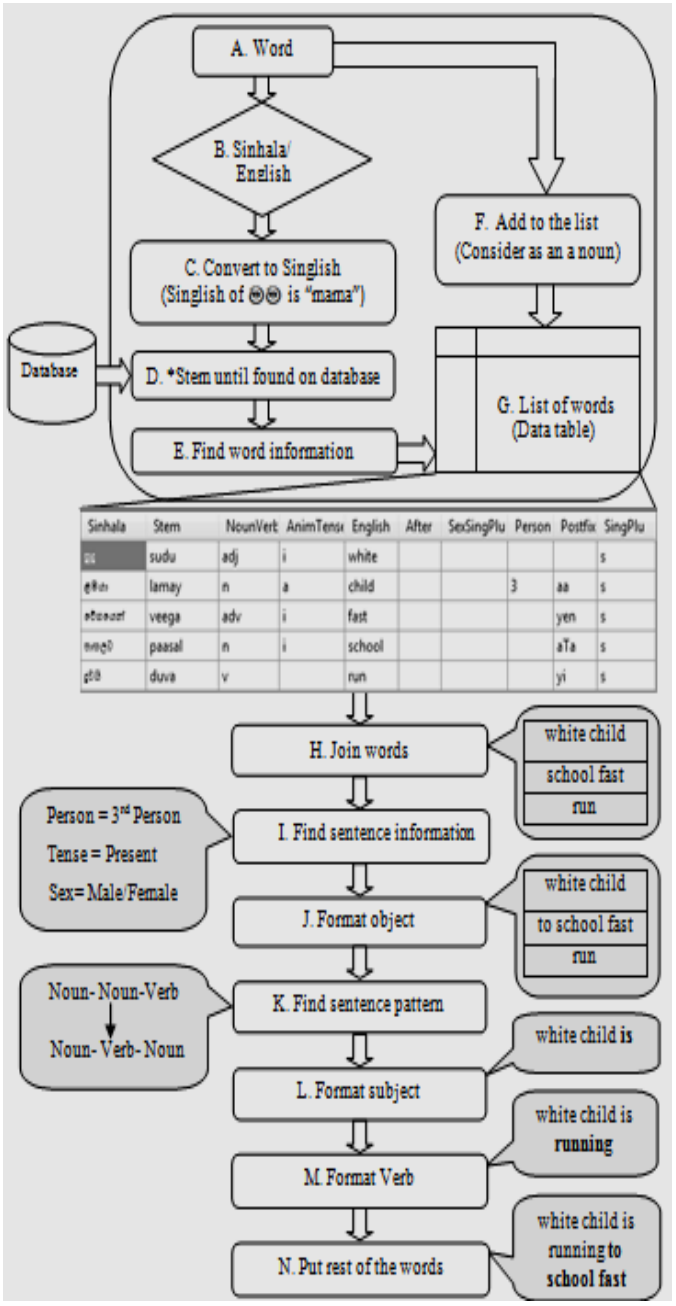

Figure 1. Translating the sentence
සුදු ළමයා වේග යෙන් පාසලට දුවයි.

Fig. 2 shows the main interface of the system. Apart from the translation process, this system contains features such as:

a *debugging tool* to assist future developers. Fig. 3 illustrates how the Sinhalese sentence සුදු ළමයා වේග යෙන් පාසලට දුවයි is processed in the debugging tool.

a *grammar tool* is also available which helps the user to find information about both Sinhala and English sentences – for example whether a particular word is a noun, verb, adjective - and provides all the materials for a user to do self studying of Sinhala and English languages.

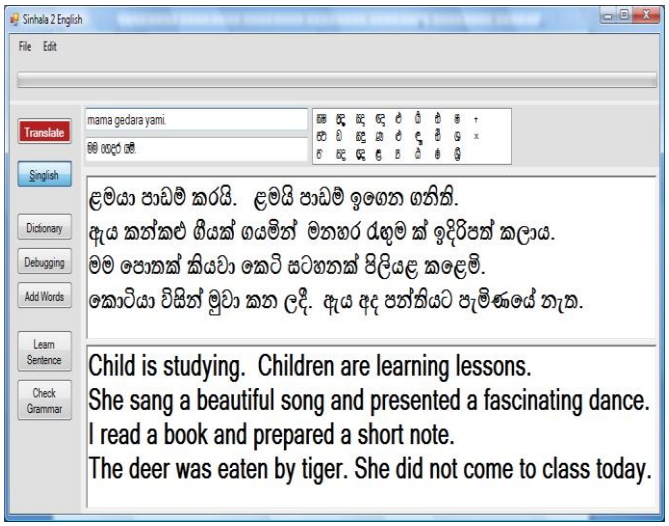

Figure 2. Main Interface of the Translator

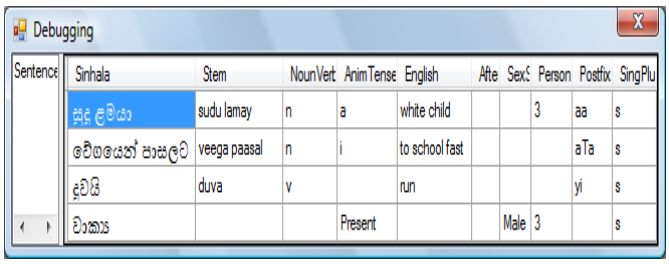

Figure 3. Debugging tool

a *grammar checker* to verify the grammar of an input Sinhalese sentence or paragraph(s). Fig. 4 illustrates the process of the grammar checker.

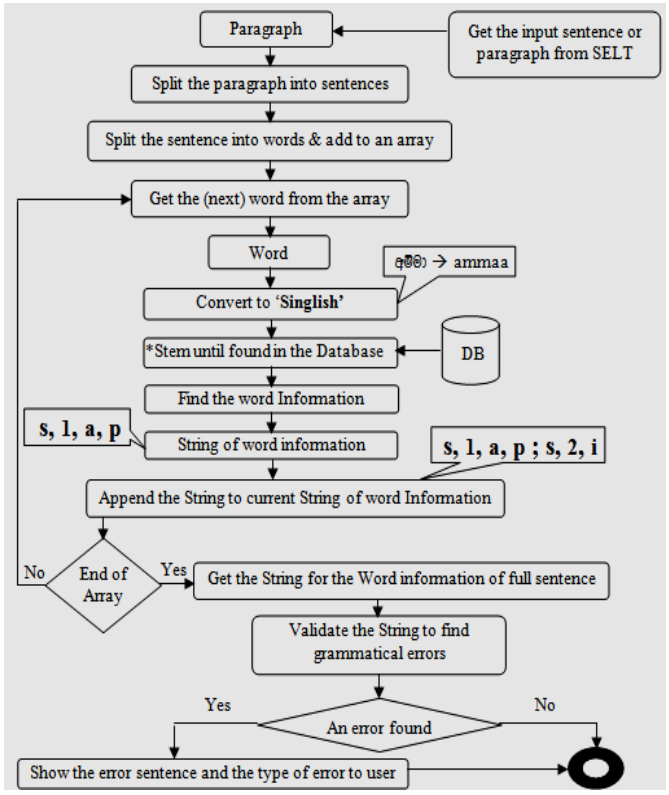

Figure 4. Process of the grammar checker

an *in-built dictionary* has been included, which gives similar words to the word that the user entered, in addition to providing meanings for both English and Sinhala words, and example sentences for each word. Apart from these, it also provides a way of pronouncing an entered English word. Fig. 5 illustrates the process of the dictionary.

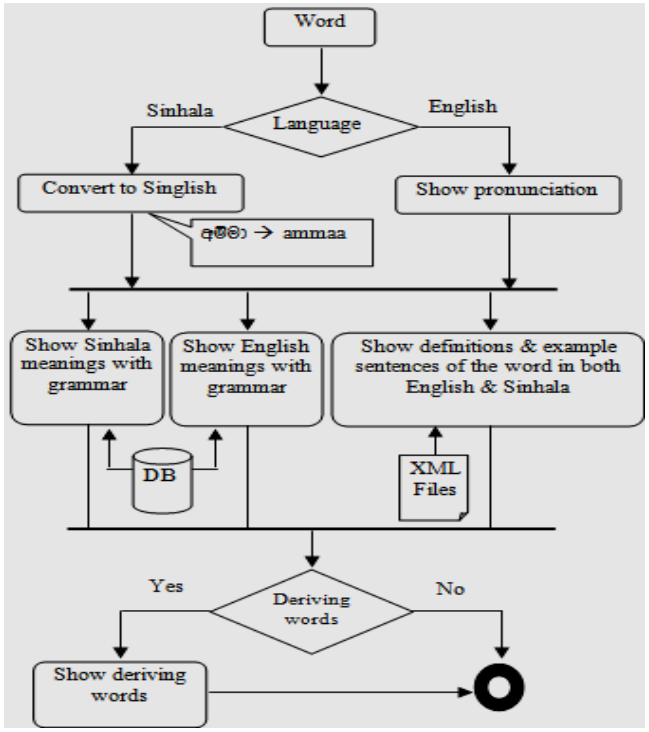

Figure 5. Process of the dictionary

an *"add-word" tool* which provides the user an easy way of adding words to the vocabulary through the front end

We have also utilized Unicode Sinhalese fonts in order to felicitate the copying or transferring of the original and translated text to other applications.

III. Results and Discussion

A. *Success*

A machine translator will never attain the overall quality of human translation because a machine does not have the human dimension required for dealing accurately with the subtle nuances of natural language. However, a translation test for 150 sample sentences was conducted using the system, and an acceptable translation result with 75 percent was achieved.

Table II contains some example sentence patterns that were correctly translated by the system. In the first sentence the female subject "ඇය" does two actions simultaneously. Since there are two verbs in this sentence, it will initially be divided into two parts and in the output those two parts will be joined with the "and" connector along with their English meanings. In the second sentence because of the conjunction "බව" initially the sentence will be separated into two parts. Then the English meanings of these two parts will be joined with the English meaning of the conjunction. When joining, the parts will be interchanged. The next set of sentences are

examples of a sentence that is in passive voice, a sentence that is in negative form, a sentence with an ownership noun, a sentence that has more than one noun connected with "ද", and a sentence with a group word, respectively.

TABLEII SOME EXAMPLES FOR CORRECTLY TRANSLATED SENTENCE

1. ඇය කන්කලු ගීයක් ගයමින් මනහර රැඟුමක් ඉදිරිපත් කළාය. She sang a beautiful song and presented a fascinating dance.
2. ඔහු නිතර පාසලට පැමිණි බව ගුරුවරයා දනී. The teacher knew that he came to school always.
3. හාවා නරියා විසින් කනා ලදී. Rabbit was eaten by fox.
4. මා අද පාසලට ගියේ නැත. I did not go to school today.
5. ඇය කුරුල්ලන්ගේ හඬට ඇහුම්කන්දී සිටියාය. She listened to bird's sound.
6. අම්මා ද තෑඹ ම්මද පන්සලට ගියෙමු. Mother, you and I went to temple.
7. ගුවන්යානා පෙළ අහසේ පියාසර කරයි. Line of aeroplanes are flying on the sky.

B. Future Research Directions

The system cannot get accurate translations for sentences with verbs such as "ඇත","නිබේ"and"සිටී", wishes (second sentence in table III), proverbs (third sentence in table III), question form sentences (fourth sentence in table III), orders such as the fifth sentence in table III. These are the areas that we hope to look into in the future. We have already begun work in some of them. In addition, more focus has to be placed on the passive voice, as the current translator is mainly focused on the active voice sentences in the Sinhala language.

TABLEIII EXAMPLE FOR INACCURATELY TRANSLATED SENTENCES

1.ලංකාව ඉන්දියන් සාගරයෙන් වටවී ඇත.
2.ඔබට දීර්ඝායුෂ ලැබෙවා!
3.පූහුල් හොරා කරෙන් දැනේ.
4.පරිගණකයට මිනිස් සංස්කෘතිය වෙනස් කළ හැකිද?
5.පාරේ දකුණු පැත්තෙන් යනු.

By increasing the number of words in the dictionary, users will have the benefit of being able to get more meanings and example sentences in Sinhalese and English. The speed of the grammar checker tool can also be optimized. The grammar tool can be further developed in a way that it would enable a foreigner to study Sinhalese grammar in the English medium.

IV. CONCLUSION

This system proposes a way of how technology based learning can enhance the quality of the current learning system for students who seek to understand the English language.

This Sinhala to English translator will help users to convert the Sinhala words, sentences or paragraphs to English language easily and to facilitate their learning.

ACKNOWLEDGMENT

We would like to express our utmost gratitude to Dr. Samantha Thelijjagoda and Mr. Jayantha Lal Amararachchi.

We would be failing in our duty if we do not mention Mrs. Chandra Perera and Mrs. Ranjani Senevirathne who were of great assistance, with regard to guiding us to the correct form of the Sinhala Language.

Last, but not least, we are deeply indebted to our parents who have contributed in innumerable ways towards the success of this project and have always been a source of inspiration.

REFERENCES

[1] S. Thelijjagoda, Y. Imai and T. Ikeda, "Japanese-Sinhalese machine translation system Jaw/Sinhalese," Journal of National Science Foundation Sri Lanka, vol.35, no.2, June 2007.

[2] W. S. Karunatillake, An Introduction to Spoken Sinhalese, M.D. Gunasena and Company Ltd, Colombo 11, Sri Lanka, 1990.

[3] S. O. Fernando, Sammana, Wara nonamena Nipatha, January 1994.

[4] S. O. Fernando, Sammana, Kriya pada igena ganimu, January 1994.

[5] S. O. Fernando, Sammana, Sinhala Nouns year 11, March 1994.

[6] L. R. H. Chapmen, English Grammar and Exercises, Longman group UK Ltd. London, 83rd ed., 1994.

[7] Machine Translation. (2008, October 19). [Online]. Available: *http://en.wikipedia.org/wiki/ Machine Translation*

[8] English Language. (2008, October 19). [Online]. Available: *http://en.wikipedia.org/wiki/English_language*

[9] English Grammar. (2008, October 19). [Online]. Available: *http://en.wikipedia.org/wiki/English_grammar*