An Accessible Translation System between Simple Kannada and Tamil Sentences

Rajaram B S R, A G Ramakrishnan and Shiva Kumar H R
Department of Electrical Engineering, Indian Institute of Science, Bangalore, India
rajaram.bsr@gmail.com, ramkiag@ee.iisc.ernet.in, shivahr@ee.iisc.ernet.in

Abstract* — *A first level, rule based machine translation system has been designed and developed for words and simple sentences of the classical Dravidian language pair Kannada – Tamil. Both grammatical and colloquial translations are made available. One can also give English input and the system returns both Kannada and Tamil equivalents. With accessibility to the visually or hearing challenged as the focus, the system has an integrated Text-To-Speech system and also gives transliterated output in Roman script for both the languages. The system has been tested by 5 native users each of Tamil and Kannada for isolated words and sentences of length up to three words and found to be user friendly and acceptable. The system handles sentences of the types: greeting, introduction, enquiry, directions and other general ones useful for a new comer.

Keywords* — *Machine translation, rule based translation, Kannada, Tamil, Text-to-speech, accessible.

I. INTRODUCTION

The Indian sub-continent, having the second largest population in the world, has people of multiple ethnicity, cultures and religions. Different parts of the country, having different influences, use various languages to communicate. The constitution of India recognizes 22 languages to be the official languages [1] and English language to be used for all official purposes [2]. Naturally, there is a need for, as well as challenges involved in inter-language machine translations (MT), transliterations and information retrieval (IR).

Karnataka and Tamil Nadu are neighboring states in southern India, where Kannada and Tamil, respectively, are the predominantly spoken languages. Both the states have a population of around 65 million [3] and about 6% of the population in Karnataka is either Tamil speaking or bilingual and about 2% of this population in Tamil Nadu [4] is Kannada speaking. Both the languages have a very rich literature and cultural history. However, the MT and IR activities with respect to Indian languages, starting in 1991 under Prof. R.M.K Sinha (ANGLABHARTI – an English to Indian Language translator [5]) to the more recent in 2009 (SAMPARK – a consortium mode Indian language to Indian language translator [6]) have not handled Kannada – Tamil pair, even though many other language pairs have been worked upon. Further, there is not a single Kannada-Tamil (K-T) dictionary or translator available either in the form of a book or as an online tool barring the 'Google Translate'.

Thus, the primary motivation is to have a medium or tool for the bi-directional translation between these two languages. The intended long term purpose of such a tool is to have a comprehensive word and sentence level translation system between the two languages with provision for adding support to other Indic languages in the future. However, currently our system is a triumvirate between the English-Kannada-Tamil (E-K-T) languages. This system can be used by non-native speakers, especially travelers/tourists, for everyday colloquial use, educational purposes and domain specific translation. The target audience are laymen devoid of any linguistic knowledge.

Section 2 gives an overview of the MILE SimpleMTS machine translation system (MTS). Section 3 describes the rules followed to translate and the integration of the TTS. The results are discussed in section 4.

II. DESCRIPTION OF SIMPLEMTS ENGINE

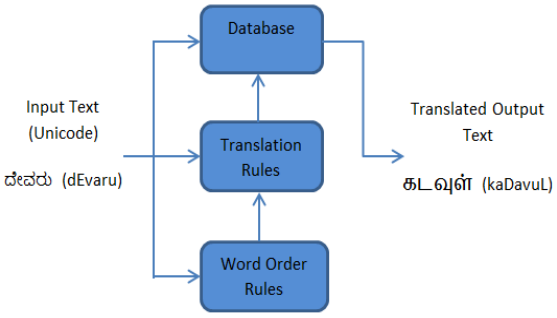

Fig 1: Block diagram of MILE MTS

SimpleMTS engine is built on a rule based translation framework. For the word level translation

and simple sentences of length up to 4 words, it is a straightforward one to one mapping between the words of the Dravidian language pair and hence there are no rules involved. However, some rules are involved when the translation is from English to Kannada or Tamil, which are discussed in Sec. 3. The block diagram of the SimpleMTS system is given in Fig. 1.

A. Database Creation

Our MTS engine is currently built on a parallel database having words from E-K-T languages with one-to-one mapping among the trio, along with the parts of speech (POS) tag for each of the word. The multiple meanings or words are intelligently handled using the translation rules. Currently we have about 9435 words in the database in the above mentioned format. The words for the parallel corpus have been collected over a period of time by the members of our laboratory. In addition to this, we also have a parallel corpus of about 200 sentences covering simple and interrogative conversations between two people.

B. Application Design

As shown in Fig 2, the current version of the tool has the provision to enter the word or a simple sentence in any of the E-K-T languages in the input field. After entering the input in the input field, on clicking the translate button, the translated words or sentences appear in the remaining languages on the remaining fields. Along with this, there is also transliteration of the target language word or sentence in Roman script. For simple sentences, if there are colloquial equivalents, then they will be displayed along with the grammatically correct word or sentence. If a word is not found in the database, the application intimates the user that the word is not present in the database and requests the user to help add the word into the database, by logging it in another field. However, it is an optional operation for the user.

III. TRANSLATION RULES

In linguistic typology, the Dravidian languages Kannada and Tamil are classified as Subject-Object-Verb (SOV) languages. This means that more often than not, the words in the simple sentences appear in the Subject – Object – Verb order. However, English is a Subject – Verb – Object (SVO) language. For example,

Kannada – **ಅವರು ದೇವರನ್ನು ಪೂಜಿಸುತ್ತಾರೆ.**
Subject – Object – Verb

Tamil – **அவர்கள் கடவுளை வழிபடுகின்றனர்**.
Subject – Object – Verb

English – **They worship God.**
Subject – Verb – Object

So, when the user enters simple sentences as the input, the translation rules come into play. If the input is a SOV sentence, then after the translation, the output in the English field will be in the SVO order for the corresponding words and vice versa.

A. Grapheme to Phoneme (G2P)

In layman's terms, grapheme is equivalent to a letter or two to three symbols that we write together and a phoneme is equivalent to a sound that we hear. In technical terms, they are the most basic structural units that can be written and spoken in a given language. These are abstract concepts that have no direct meaning. They combine together to form the written and verbal forms of words that have some meaning [7]. For the language pair Kannada and Tamil, the mapping between grapheme to phoneme is defined by clear rules. This means that finding grapheme is as good as finding a phoneme and hence the pronunciations can be accurately determined.

More often than not, the user, who wants to translate the word, does not know the script of the target language. For this purpose, we give the translated word in the target script and also in the Roman script. This is an effort to make the tool accessible to the hearing challenged.

Example:

English Input – what is the price?

Kannada Output –
ಇದರ ದರ ಎಷ್ಟು? [This price how much?]

Colloquially – ಇದು ಬೆಲೆ ಎಷ್ಟು? [This cost how much?]

Pronunciation -
Grammatically : idara dara eShTu
Colloquially : idu bele eShTu

Tamil Output –
இதன் விலை எவ்வளவு [This price how much?]
Colloquially - **இது என்ன விலை** [This what cost?]
Pronunciation -
Grammatically : idan vilae ewwaLawu
Colloquially : idu enna vilae

B. Integrated Text-to-Speech Engine

A text-to-speech engine as the name suggests converts a given language's script to intelligible audio. The work involved in the TTS systems is multidisciplinary, ranging from concepts of signal processing, natural language processing and acoustics. Our laboratory already has a popular web demo of its Text-to-Speech (TTS) conversion system for Tamil and Kannada languages [8].

In the context of this SimpleMTS, the spoken form of the translated texts for the language duo of Kannada/Tamil are also available with the integration of the TTS [9] on the same web page. This provides the users, who may not know the target script (Kannada/Tamil), with the spoken form of the output, thereby giving the pronunciation. Additionally, this also helps making the tool accessible to blind users.

IV. Evaluation and Discussion

Evaluation of the machine translation system is a complex and subjective task. They are subjective because evaluations or at the very least the evaluation metrics need to be different for users of the system and to the researchers and developers of the system. There are multiple factors to be considered in the evaluation like the semantic structures of the source and target languages, the specifications of the system as required by the end users, the qualitative adequacy and comprehensibility and many more. Furthermore, the sentences selected for testing and the background of evaluators used impacts the evaluation process [10]. There is no specific/universal standard based on which the MT evaluations can be done.

Based on the applications the MT systems are meant for, one can define the metrics to have a fairly pragmatic feedback on the system's quality and performance.

Keeping in mind the preliminary framework we have for the K-T MT system, our pool of evaluators has 6 native speakers and 1 bilingual speaker for Kannada. On the other hand, for Tamil we have 4 native speakers and 1 bilingual speaker. We evaluate the system based on the following metrics [11]:

- Translation Quality
- Level of Comprehensibility

A. Translation Quality

The translation quality is all about the extent to which the translation is good or bad and is independent of the user's ability to understand the intended meaning after translation. It is about the faithfulness of the translation from source to the target language. Scoring scheme of the translation quality is defined in Table I.

B. Level of Comprehensibility

This metric is about the user's capability to understand the meaning of the sentence. This is not a measure of the grammatical correctness. Here, the users are required to give scores based on their understanding of the translated sentences. To keep the user's mind unbiased, the users are not allowed to see the source sentence to begin with. They are instructed to see the translated sentence first and then optionally look into the source sentence to check if the gist of the translated sentence is the same as the source sentence. The scoring scheme is given in Table II.

TABLE I. Scoring scheme for translation quality

Score	Significance
1	Zero fidelity – *Absolutely wrong*
2	Barely faithful – More than 50% word and/or order errors.
3	Fairly faithful – Less than or equal to 50% word and/or order errors.
4	Acceptable – Less than or equal to 25% word and/or order errors.
5	100% fidelity – Perfect Translation

TABLE II. Scoring scheme for comprehensibility

Score	Significance
1	Unintelligible – Doesn't make any sense.
2	Barely Intelligible – the general idea is intelligible only after considerable study. There are a lot of grammatical inaccuracies and poor/wrong word choice.
3	Comprehensible – The general idea is clear and intelligible. Despite some grammatical errors and/or word order errors, the message is conveyed.
4	Fairly acceptable – The sentence has minor grammatical errors and completely understandable.
5	Good and accurate translation

The results of our system for its accuracy and intelligibility averaged over the scores of each tester, for the language pair Kannada-English are given in Table III. Here, each tester is asked to enter 5 simple sentences with their vocabulary bounded 'mostly' to the domains of (G) greeting, (I) introduction, (E) enquiry and (D) directions (GIED). With these restrictions enforced, we have a total of 35 sentences under the GIED domains for Kannada-English language pair.

TABLE III. RESULTS OF ACCURACY AND INTELLIGIBILITY FOR KANNADA-ENGLISH PAIR

Tester No.	Average Accuracy		Average Intelligibility	
	Eng – Kan	Kan - Eng	Kannada Set 1	Kannada Set 2
1	3.8	3.2	3.8	3.2
2	4.2	3	2.6	2.6
3	4	2.8	3.6	3
4	3.8	3.2	4	4.2
5	4.2	2.6	3	3
6	4	2.8	4.6	4
7	3.6	3	3	2.4

The average accuracies of the translated sentences from English to Kannada (E-K) and K-E are 3.94 and 2.94, respectively. These are equivalent to the percentages of 78.8% and 58.8%, respectively. Likewise, the numbers 3.51 and 3.2, respectively denote the average intelligibility of the translated Kannada outputs. These are equivalent to 70.2% and 64%, respectively for the set 1 and set 2 of translated Kannada outputs.

Similarly, the results of the system for its accuracy and intelligibility averaged over the scores of each tester, for the language pair English-Tamil is given in Table IV. Similar to the Kannada case, here too the testers are required to enter 5 sentences each, making a total of 25 sentences under the GIED domains for the Tamil-English language pair.

The scores 3.08 and 3.32 represent the average accuracy of the translated sentences from English to Tamil (E-T) and vice versa. This amounts to 61.6% and 66.4% respectively. In the same vein, scores 2.32 and 3.96 represent the average intelligibility of the translated Tamil outputs from sets 1 and 2. This amounts to 46.4% and 76% respectively.

The remaining percentage that accounts for inaccuracy for both E-K (21.2%) and vice versa (41.2%), and E-T (36%) and vice versa (32%) are accounted for, by the virtue of idioms in input or inappropriate word forms and in some cases word order errors. These factors affect the intelligibility percentages too.

As seen in Table III, the average intelligibility scores range from 2.6 to 4.6 because a couple of testers were not given the instructions properly. These two scores have skewed the results to have a higher average value than anticipated to give a very distributed range. However, since we have very few test data points, we have retained the results.

TABLE IV. RESULTS OF ACCURACY AND INTELLIGIBILITY FOR TAMIL-ENGLISH PAIR

Tester No.	Average Accuracy		Average Intelligibility	
	Eng-Tam	Tam-Eng	Tam – Set 1	Tam – Set 2
1	3.2	2.8	2.6	4.4
2	3	3.6	2.2	3.2
3	3.2	3.8	2.2	3.2
4	3	3.4	2.6	4.4
5	3	3	2	4.6

The web demo of MILE MTS is available at http://mile.ee.iisc.ernet.in:8080/SimpleMTS. A link for Indic Keyboard interface, an open source Indic script input software developed by our Lab [11] is also provided at the demo site, which enables the users to input Tamil and/or Kannada text in Unicode. The text can also be copied and pasted from any website supporting Unicode Tamil and/or Kannada text.

V. CONCLUSION AND FUTURE WORK

Since the application development is in its nascent stage, only word level and simple straightforward three or four word sentences are handled for translation. Also, our system has a small parallel corpus of 9435 words and 200 sentences in the bounded domains of GIED which is fairly low for the MT systems. Our system lacks morphological analyzers for Kannada and Tamil to have a full-fledged system. Our testing also involved only five sentences per person per metric, which is rather low.

Since currently a Kannada-Tamil translation system is virtually non-existent, venturing to develop it is a basic requirement. However, the highlights of our system include the accessibility features in the form of integrated TTS and the Romanized transliteration of the Tamil or Kannada outputs for the visually and hearing challenged. A Romanized transliteration not only helps in accessibility but also helps normal users with pronunciation of the unknown language.

This basic framework of our MTS engine can be used as a template to support other SOV Indic languages like Telugu, making it a multi-lingual MT system. However, machine learning concepts and/or morphological analysis and synthesis are needed to scale the system to support higher word count in and also complex sentences. A rigorous and comprehensive testing involving a large number of sentences needs to be performed to get more reliable results.

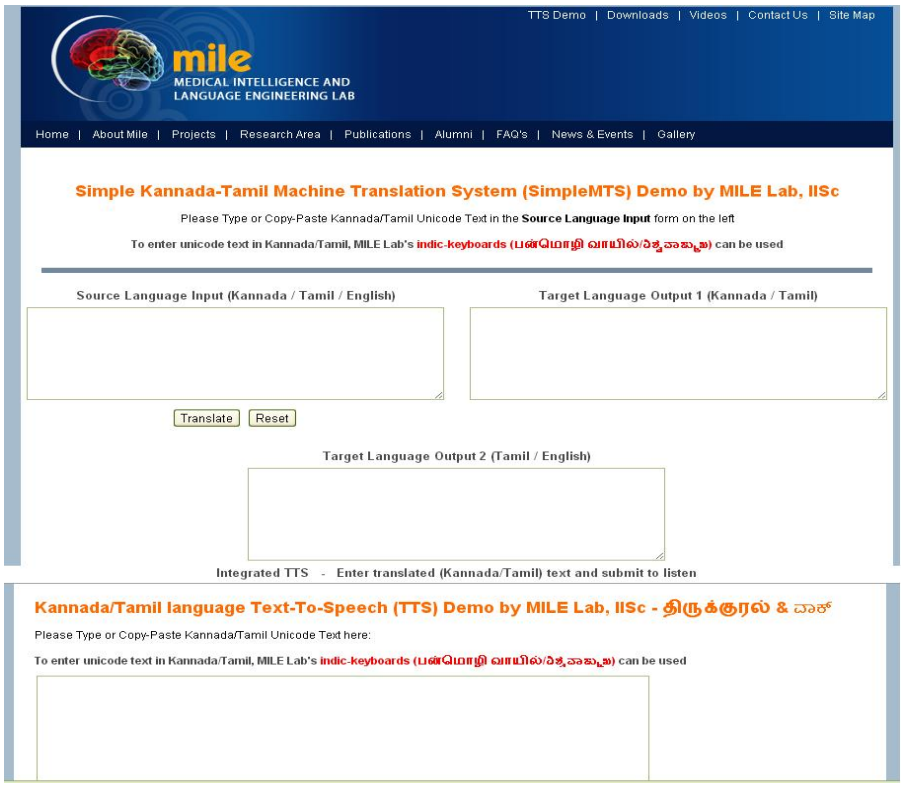

Fig 2: GUI of the accessible SimpleMTS Kannada-Tamil MTS web application.

REFERENCES

[1] 22 official languages in India – (http://www.constitution.org/cons/india/p17344.html)

[2] Official Languages Act 1963: Continued Use of English – (http://www.rajbhasha.gov.in/khand8-eng7.pdf)

[3] Population of Karnataka (http://censusindia.gov.in/ Census_ Data_2001/Census_Data_Online/Language/Statement3.htm)

[4] Kannadigas in Tamil Nadu (http://articles.timesofindia.indiatimes.com/2008-04-16/bangalore/27782927_1_tamil-nadu-malayalis-gujaratis)

[5] Kumar Sourabh,"An Extensive Literature Review on CLIR and MT activities in India", International Journal of Scientific & Engineering Research, Feb 2013.

[6] Consortium Mode Machine Translation System – (http://www.tdil-dc.in/components/com_mtsystem/CommonUI/homeMT.php)

[7] Paul Taylor, "Text-to-Speech Synthesis", Chapter 2, Cambridge University press, 2009.

[8] Shiva Kumar H R, Ashwini J K, Rajaram B S R, A G Ramakrishnan, "MILE TTS for Tamil and Kannada for Blizzard Challenge 2013", Proc. of Blizzard 2013 workshop, UPC, Barcelona, Sept 2013 (http://www.festvox.org/blizzard/bc2013/MILE_Blizzard2013.pdf)

[9] Bharati A, Moona R., Singh S., Sangal R., Sharma D.S., "MTeval: An evaluation methodology for Machine Translation Systems", Proc. SIMPLE Symp on Indian Morphology, Phonology and Lang Engineering 2004, IIT Kharagpur, INDIA

[10] G S Josan and G S Lehal, "Evaluation of direct machine translation system for Punjabi to Hindi", International Journal of Systemics, Cybernetics and Informatics, 2007, pp. 76-83.

[11] HR Shiva Kumar, Abhinava Shivakumar, Akshay Rao, S Arun, AG Ramakrishnan, Panmoshi Vaayil – a multilingual Indic keyboard interface", Information Systems for Indian Languages, 2011, pp. 278-283.