SATS: A Computer Aided Translation Engine

INTRODUCTION

Translation has become an important part of our day to day life. Developers have worked to bring translation services to our finger-tips.

Google Translate app supports 103 languages, now including speech and camera translations.

PROBLEMS OF GOOGLE TRANSLATION WITH RELATED TO LOCAL LANGUAGES

English	Sinhala	English	Tamil
barking dogs seldom bite	ලැජ්ජයි læjjayi	give me 1 kg of Ladies fingers	எனக்கு 1 கிலோ மகளிர் விரல்களை கொடுங்கள் Eṉakku 1 kilō makaḷir viralkaḷai koṭuṅkaḷ
English	**Tamil**	**English**	**Sinhala**
barking dogs seldom bite	குங்குமப்பூ நாய்கள் எப்போதாவது கடி Kuṅkumappū nāykaḷ eppōtāvatu kaṭi	give me 1 kg of Ladies fingers	Ladies ඇඟිලි 1 ක් දෙන්න Ladies æñgili 1 k denna
Sinhala	**Tamil**	**Sinhala**	**Tamil**
බුරන බල්ලා සපා නොකයි burana ballā sapā nokayi	நாய் வெடிக்கக் கூடாது Nāy veṭikkak kūṭātu	තේ වක්කරන්නද්? tē vakkarannada?	தேயிலை சமைக்க வேண்டுமா? Tēyilai camaikka vēṇṭumā?

CAUSE FOR THE PROBLEM

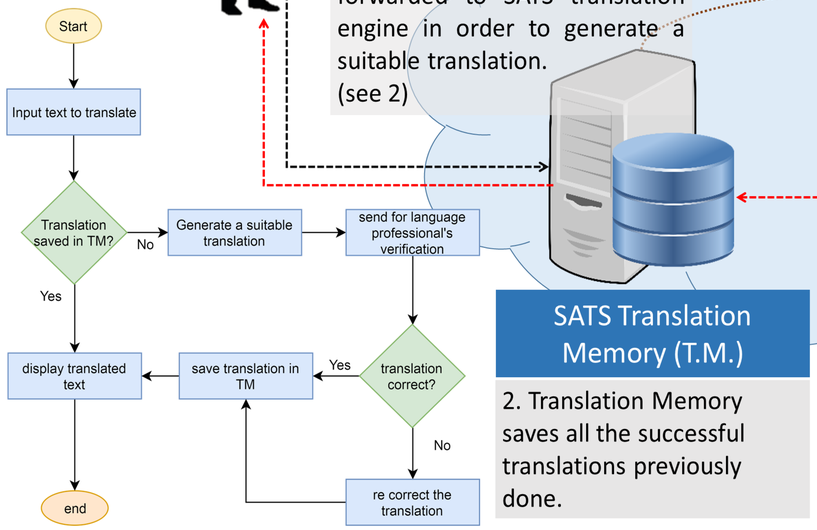
Lack of parallel corpora in local languages makes the system to predict inaccurate translations.

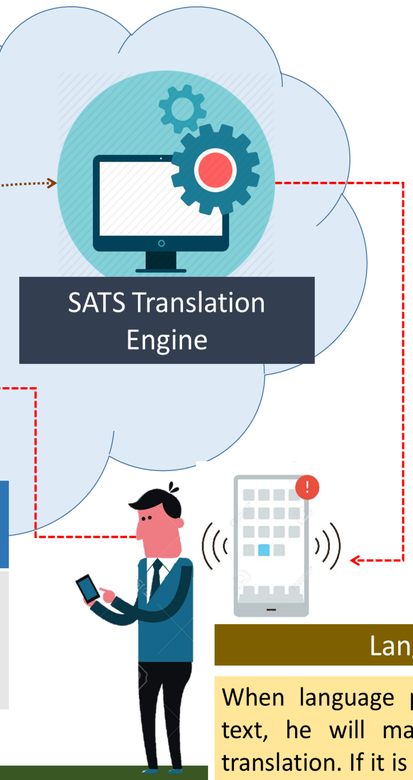
As most of the Tamil corpora are created by Indians, the translated text is more biased to Indian Tamil slang.

SATS Translation Engine was developed by considering the metrics accuracy and regional slang.

HOW SATS WORKS ?

1. A translation request is sent by the user through the SATS user interface. If the translation is available in TM, an immediate reply is sent to the user by the system. If not the request is forwarded to SATS translation engine in order to generate a suitable translation. (see 2)





2. Translation Memory saves all the successful translations previously done.

3. SATS Translation engine which is powered by Moses, IRSTLM and GIZA++ generates a suitable translation for the input text. This system is trained with a number of parallel corpora in local language. After generating the out put, rather than it is sent back to TM, the translated text is forwarded to the SATS (Sinhala And Tamil Speakers) group to check the correctness of the translation.

Languages supported:
- **Sinhala**
- **Tamil**
- English

Language Professional

When language professional receives a translated text, he will manually check the correctness of translation. If it is incorrect, then he will edit the text into a correct answer. The corrected text is sent to the user through the TM. The system provides accurate translations due to the involvement of experts.

TERMS AND THEIR DEFINITIONS

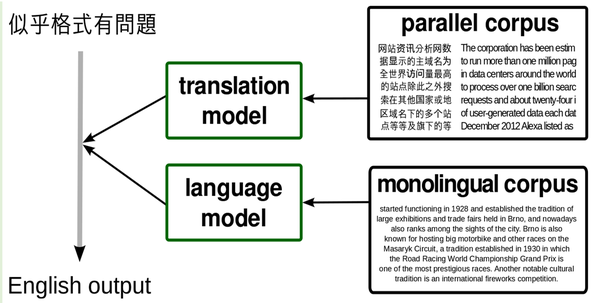
Statistical Machine Translation (SMT)

Statistical machine translation (SMT) is a **machine translation** paradigm where **translations** are generated on the basis of **statistical** models whose parameters are derived from the analysis of bilingual text corpora.

SATS uses **Moses** statistical machine translation system that allows you to automatically train translation models for any language pair.

Language Modelling (LM)

Building a statistical language model that can estimate the distribution of natural language as accurate as possible. **IRSTLM** tool is used for language modeling.



TRANSLATION MEMORY (TM)

A translation memory (TM) is a database that stores "segments", which can be sentences, paragraphs or sentence-like units (headings, titles or elements in a list) that have previously been translated, in order to aid human translators.

FAQ

What is SATS?
SATS was originally a group of people known as **S**inhala **A**nd **T**amil **S**peakers, who are experts in Both Sinhala and Tamil (English as well) who contributed for the development of this system. This group has members representing all the ethnic groups of Sri Lanka. Apart from this project, they are engaged in similar projects to reduce the language barriers in Sri Lanka.

Is SATS another trilingual dictionary?
No. SATS is not only a trilingual dictionary. It also allows users to translate sentences and phrases.

How the corpora is fed?
The English-Sinhala corpus is fed with thousands of Sinhala subtitles obtained from www.baiscopelk.com. For English-Tamil corpus we have obtained articles from the internet, localization strings and Wikipedia articles. As we do not have enough corpus to feed the Sinhala-Tamil corpus for now, we take the assistance of Google's translation engine until we eventually develop our own Sinhala-Tamil corpus.

How accurate the system is?
The system provides accurate translations due to the involvement of experts. The accuracy of the translation is up to the accuracy of the translator who is working in the back end. However, with the time the system will evolve as an expert language translation system, and will be able to provide accurate translations on its own with the help of Translation Memory and the self-developed corpora.

KUMARA A.I.U, NASEER MOHOMAD

Poster ID: XXXX