Issues while developing Sangam Tamil-English Bilingual Parallel Corpora for Statistical Machine Translation System

Dr.K.Umaraj
Assistant Professor, Department of Linguistics,
Madurai Kamaraj University, Madurai -625021 umarajk@gmail.com

1. Introduction

Machine Translation is the Translation of text by a computer, with no human involvement. Machine Translation can also be referred to as automated Translation, automatic or instant Translation. There are two types of Machine Translation systems. One is rules-based Machine Translation and another is statistical based Machine Translation. Rules-based systems use a combination of language and grammar rules plus dictionaries for common words. Rules-based systems typically deliver consistent Translations with accurate terminology when trained with specialist dictionaries. Statistical systems have no knowledge of language rules. Instead they "learn" to translate by analysing large amounts of data for each language pair. They can be trained for specific industries or disciplines using additional data relevant to the sector needed. Typically statistical systems deliver more fluent-sounding but less consistent Translations.

2. Statistical Machine Translation using parallel corpora

Statistical Machine Translation is mathematical model in which the process of human Translation statistically modelled. Statistical methods allow the analysis of parallel text corpora and the automatic construction of Machine Translation systems. In Statistical Machine Translation system, correspondences between the words in the source and the target language are learned from the bilingual corpora on the basis of alignment models. The engine uses state of the art statistical techniques which are presently gaining momentum in the Machine Translation community. There is a lot of work going on for building parallel of corpora of Indian languages. Few of them are as follows. E-ILMT, ILCI, MAT for English to Hindi, ELMT, Google Translating Corpora, Microsoft Bing Corpora and Yahoo bable fish Corpora. In Tamil Nadu, Dr.Kamakshi has discussed in detail about the parallel structure of English and Tamil and his data will be very useful for building Machine Translation system using transfer approach. G.Vasuki explained in her thesis about the parallel corpora of English and Tamil and her work will be very useful for building Statistical Machine Translation. However, so far no body developed Parallel corpora for Sangam Tamil and its English Translation. If we develop bilingual corpora for Sangam Tamil and its English Translation, definitively some issues will arise. Thus the present paper aims to identify the issues while developing Parallel Corpora for Sangam Tamil and its English version.

3. Statistical Machine Translation Model for Sangam Tamil

Statistical approach to Machine Translation generated Translations using statistical methods by deriving the parameters for those methods by analyzing the bilingual corpora. The following figure shows the block diagram of Statistical Machine Translation system for Sangam Tamil

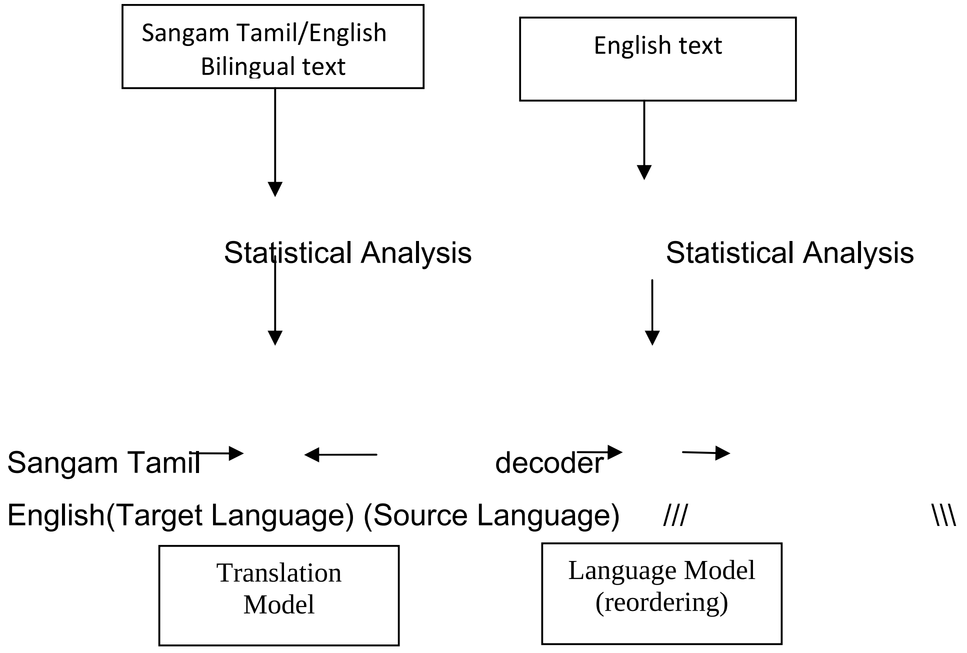

Sangam Tamil/English Bilingual text
English text
Statistical Analysis
Statistical Analysis
Sangam Tamil
decoder
English(Target Language) (Source Language)
///
\\\
Translation Model
Language Model (reordering)

Fig 1 Block diagram of Statistical Machine Translation

4. Translation Model

Translation models describe the mathematical relationship between two or more languages. It may be called as models of Translational equivalence because the main thing that they aim to predict is whether expressions in different languages have equivalent meanings. The role of Translation model is to find p(f/e) the probability of the source sentence (Sangam Tamil sentence) f given the translated sentence (English version of Sangam Tamil)

5. Language Model

A language model gives the probability of a sentence. The probability is computed using N-gram. For example it will give answer for the questions How likely is a string of English words is a good English? It will help to reorder a sentence and tells us what word will go will what word.

6. Decoder

The inputted Sangam Tamil corpus will be first decoded by Translation model. Language model will rearrange the word order of source language sentence into Target language word order. It is an important process for languages which differs in their syntactic structure. English and Sangam Tamil language pair has different syntactic structures. English word order is Subject-Verb_Object (SVO) whereas Sangam Tamil word order is Subject-Object-Verb (SOV). The main verb of a Tamil sentence always comes at the end but in English it comes between subject and object.

7. Data for the analysis

நிலத்தினும் பெரிதே வானினும் உயர்ந்தன்று
நீரினும் ஆரளவின்றே சாரல்
கருங்கோற் குறிஞ்சிப் பூக்கொண்டு
பெருந்தேன் இழைக்கும் நாடனொடு நட்பே

குறிஞ்சி மரத்தின் மலர்களைக் கொண்டு தேனடை செய்யும் நாட்டை கொண்ட தலைவனுடன் நான் கொண்ட நட்பானது நிலத்தின் அகலம் போலவும், வானின் உயரம் போலவும், கடலின் ஆழம் போலவும் பெரிது.

நிலத்தினும் பெரிதே	bigger than the earth
வானினும் உயர்ந்தன்று	higher than the sky
நீரினும் ஆரளவின்றே	deeper than the sea

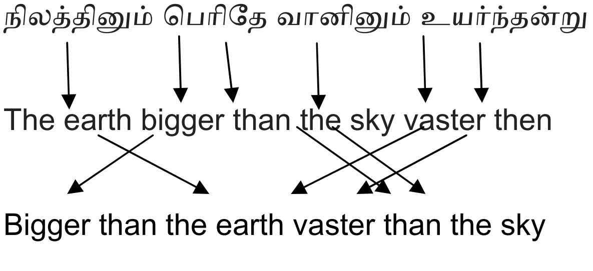
 கருங்கோற் குறிஞ்சிப் பூக்கொண்டு
பெருந்தேன் இழைக்கும் நாடனொடு நட்பே

is my love for him
from the hills
where the honeybees make
abundant honey
from the black-stemmed
kurinji flowers.

8. Bitext word alignment

Bitext word alignment is a NLP task of identifying Translation relationships among the words in a bitext, resulting in a bipartite graph between two sides of the bitext, with an arc between two words, if and only they are Translations of one another. Bitext word alignment is an important supporting task for most methods of statistical Machine Translation, the parameters of statically Machine Translation models are typically estimated by observing word aligned bitexts and conversely automation. Recent work begun to explore supervised methods which rely on presenting the system with a number of manually aligned sentences. In addition to the benefit of the additional information provided by supervision, these models are typically also able to more easily take advantage of combining many features of the data, such as context, syntactic structure , POS information etc..

8.1 Sample Bitext word alignment



நிலத்தினும் பெரிதே வானினும் உயர்ந்தன்று

The earth bigger than the sky vaster then

Bigger than the earth vaster than the sky

8.2 Use of Parallel corpora

Helps in teaching a particular language

Helps in Terminological studies

Helps to build Translation Machines and

It also helps to build cross language information retrieval engines.

9. Issues in developing parallel corpora

1. Identifying what is a word? What is not a word? in Sangam Tamil is a problem. whether all the compound words should be written as one word or not, whether all postpositions and clitics should be part of the word or it should be a separate word? There were no clear guidelines for segmenting the Sangam texts.

2. Assigning grammatical information to the head word is another issue in Sangam Tamil.

3. Identification of meaning for a particular word in Sangam Tamil is another issue.
4. In Sangam Literature one meaning can be represented in different words and in the same way one word may represent more than one meaning.

5. Word in languages may changes. So the language changes can be ranked with the help of the language model and the best can be selected.

6. Idioms don't have a direct meaning. While translating idioms we have to take care properly. A separate dictionary for idioms is necessary or manually we have translated those words.

7. In parallel corpora, single sentences in one language can be found translated into several sentences in the others and vice versa. Sentence aligning should be properly done.

10. Conclusion

The present paper discusses Translation of Sangam Tamil to English using parallel corpus. The accuracy of the system depends on the amount of parallel corpora available in the languages, addition of linguistic materials such as morphological information and Parts of Speech categorization can enhance the accuracy of the system.

Reference

1. Dr. Anand Kumar M "Morphology based Prototype statistical Machine Translation System for English to Tamil Language" Phd Thesis submitted at Amirtha University, Coimbatore April 2013

2. Latha R. Nair, David Peter S. "Machine Translation System for Indian Languages" , International Journal of Computer Applications Volume 39-1, February 2012