Syntactic and Semantic Based Machine Translation from Tamil to English

Dr. S. Lakshmanapandian, K. Saranya, J. Durkadevi

Abstract

Machine Translation technology is the most feasible method which is used to translate large volumes of domain specific languages. Machine Translation is becoming a core competence for the future success of global business and this is done through a collaboration of machine translation technology and human translators. Mostly machine translation can be done by the English to some other natural languages. South Indian languages such as Telugu, Tamil and Malayalam have only few parallel corpora available and also Morphological rich. In such cases adopting only the statistical technique will not result in proper translation in the target language. Hence, they are adapting Hybrid method of translation, i.e., Combination of Rule Based and Statistical Based Machine Translation Process. In existing work, the time consumption for translating one natural language to another is high. The output of the text is not accurate. It can translate by means of word by word so the meaning of the sentence is differing from the input text. Natural language computation in Tamil language is difficult due to its inherent characteristic of morphologically richness and free word order. The system is mainly focus on Tamil-English Machine Translation. The proposed system can also possess chunking and naming entity. Chunking is combining many numbers of words. Naming entity can be used to identify the gender of the person who is performing actions. The main process involved in this Tamil - English translation is a dictionary and transliteration. The dictionary consists of many numbers of Tamil words with their meaning in English. Based on the dictionary the meaning of the input text is identified and it is translated. The named entities are transliterated.

INTRODUCTION

Natural Language Processing (NLP) is a computerized approach to analyze the text based on a set of theories and set of technologies. This active area of research and development involves the basic objective of Natural Language processing to facilitate human-machine interaction through the means of natural human language. Various research on NLP focused on various intermediate tasks that make partial sense of language structure without requiring a complete understanding which, in turn, contributes to develop a successful system.

Natural Language Processing (NLP) is a method to translate between computer and human languages. It is a method of getting a computer to understandably read a line of text without the computer being fed some sort of clue or calculation.

NLP was initially distinct from text information retrieval (IR), which employs highly scalable statistics-based techniques to index and search large volumes of text efficiently.

Machine translation is the translation of text by a computer, without human involvement. Pioneered in the 1950s, machine translation can also be referred to as automated translation, automatic or instant translation. Machine translation is abbreviated as MT is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another. In general, there are two types of machine translation system, namely rules-based and statistical based. Rules-based systems use a combination of language and grammar rules plus dictionaries of common words. Specialized dictionaries are created to focus on certain industries. Rules-based systems stereotypically 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 analyzing large amounts of data for each language pair. They can be trained for specific industries or disciplines using additional data relevant to the sector required. Classically statistical systems deliver more fluent-sounding but less consistent translations.

**Increased productivity – deliver translations faster:**
- Use freely available machine translation to pre-translate new segments that are not leveraged from translation memory.
- In reference to translator, you can connect to and use a customer's or supplier's trained SDL Enterprise Translation Server or SDL Be Global machine translation system to deliver even faster.

**Flexibility and choice – to suit all types of project:**
- Select from 3 different machine translation engines.
- Choose from over 50 languages and more than 2,500 language pairs to suit respective projects.
- Comparability rules-based and statistical machine translation engines.

2. **Related Work:**

Lakshmana Pandian et al (2012) describes, Machine Translation from English To Tamil using Hybrid Technique. Machine translation is done only at the word level, i.e. word by word translation. Later, as the technology grows many researchers have been carried out in this field and thus the translation of the phrase and sentence level has been implemented. The three major techniques involved machine translations are Rule Based, Statistical, and Example Based Technique. South Indian languages such as Tamil, Telugu, and Malayalam have only few parallel corpora available. In such cases adopting only the statistical technique will not result in proper translation in the target language.

Kanika, Ankur, Divyanjali and Shalini Mittal (2014) India is very rich when it comes to languages; hence entirely different rules are needed to be incorporated for each language. Also, Indian languages are free word order languages, which should be dealt accordingly. Anusaaraka is a kind of language translator that is designed to translate English to any Indian language. Outlines of some other Indian language translators developed during recent years are also given in short, to provide an overview of the progress in the field of translation. The key features of each translator are presented and analyzed to meet the requirements of an ideal system.

Paninian framework was designed for writing Sanskrit grammar decades ago. But as all the Indian languages have something in common with it, Paninian framework can be made as a base for the parsers. There are two main parts of the parsers. The lexicon is language dependent, but the parser does not depend on language and hence it is a great benefit that the same parser can be used for almost all Indian languages just by changing the lexicon. Hence it deals with many Indian languages. Sarah Ebling, AndyWay, Martin Volk, and Sudip Kumar Naskar (2011) described about The state-of-the-art approach in MT is a phrase based Statistical Machine Translation (SMT). Together with Example-Based Machine Translation (EBMT), SMT belongs to the Corpus-Based Machine Translation (CBMT) paradigm. Hence, both SMT and EBMT rely on a sententially aligned bilingual corpus. EBMT systems make use of the parallel corpus of consulting the training set (their example base) directly at runtime.

Loganathan Ramasamy Onďrej Bojar Zden´ek Žabokrtský (2012) Statistical machine translation (SMT), applying either pre-processing or post-processing to morphologically rich languages leads to better translation quality. In this work, it focus on the English- Tamil language pair. We implement suffix-separation rules for both of the languages and evaluate the impact of this pre-processing on translation quality of the phrase-based as well as the hierarchical model in terms of BLEU score and a small manual evaluation. The results confirm that our simple suffix-based morphological processing helps to obtain better translation performance.

R.Akilan and E.R.Naganathan (2012) described about Rule based model of parts of speech (POS) tag set for Classical Tamil Texts (CTT). The noun forms are represented as type pattern, verb forms are denoted as token pattern. This is done based on the form agreement method. CTT is very efficient and novel approach because Tamil Language has a built-in system of agreement/concord of the sentence.

B. Sasidhar, P. M. Yohan, Dr. A. Vinaya Babu, Dr. A. Govardhan, (2011) In the language of English it is observed capitalization is a major clue to identify NEs. Also, Indian languages are resource poor languages and gazetteers available are insufficient. Indian languages are agglutinative in nature the reason being more number of inflectional words.

3. **Proposed Method:**
The proposed system is divided into four modules

3.1 Morphological analysis
3.2 Part of speech tagging
3.3 Chunking
3.4 Named Entity Recognition
3.1 Morphological analysis:
The given input Tamil text can be analyzed by performing this morphological analysis process. First, the given text can be split by the means of word by word after that each and every letter is to be gets split based on the meaning of the word and also identify the corresponding tense in which the taken.

Then, we have to identify the root word of each and every word present in the given input Tamil text and then we have to identify the morpheme components.

3.2 Part of Speech Tagging:
Parts of speech tagging assigns a grammatical category of the languages that includes noun, Verbs, adjective, adverbs, and postposition etc.

Tamil is an agglutinative language. In Tamil, root or stem can be added with some suffixes that derivate new words. Those words are divided into two parts namely:
- Lexical word
- Grammatical word.

Thus nouns, verbs, adjectives and adverbs belong to open classes and pronouns, whereas conjunction belongs to closed classes. We have to identify the sentence pattern with the given word like Subject, Noun, Verb, Object, Preposition, Adjective, Adverb etc.

3.3 Chunking:
The words in one language will not have equivalent words in any another language. Also, in some cases a word in one language has to be expressed by groups of words in another. In such cases chunking of the source language has to be done.

Chunking is nothing but a grouping of words into a phrase. A noun phrase contains an adjective and a noun. Using bi-grams in a sentence are found and are translated into target language using the dictionary file. We have to identify the connected item of words.

3.4 Named Entity Recognition:
We have to identify the suffix words. These suffix words can be able to denote the person who is performing actions and it can also perform the transliteration process. It can also be able to find out the word which has a similar meaning in Tamil as well as in English language.

![Diagram](image_url)

Fig. 3.1: Proposed system design.

4. Experimental Results:
In Morphological analysis, we have to identify the rootword and its morpheme components. For example, அரண்மனையில் அனைவரும் அமர்ந்து ககொண்டிருந்தொர்கள். In this அரண்மனை represents rootword and பிலே represents morpheme components.
In part of speech tagging, consider

எந்தொளி அரண்மனையில் கெரிய விருந்து நடந்தது.

எந்தொளி - Noun

கெரிய - Adjective

விருந்து - Subject

நடந்தது - verb

In Chunking process, we have to identify the connected item of words. Consider,

அப்படியொ கொல்கிறொய் நூறு கொண் எந்தயம் கட்டுகிறறன்.

அப்படியொ? - Single word (Tamil)

Is it? - Compound words (English)

In Named Entity Recognition, we have to identify the word which possess different meaning in same language based on the meaning of the sentence.

Consider,

இயற்கை விலை கவள்ளி

துவாக்கீர்க்கு இயற்கை விலைக் கவள்ளி இரண்டொவது ரகொளொகு உள்ளது.

In first sentence இயற்கை represents silver and in second sentence இயற்கை represents planet venus.

And also identify Letters like ga and ka represent the same Tamil letter க and also Ti, Ty, Di, Dy represents டி.

We have to identify these words and their meaning.

Conclusion:

Machine Translation presents an effective methodology for Tamil to English translation. The work done based on this till now limits to translation of simple sentences from Tamil to English. The system can be further enhanced by translating complex sentences by creating new morphological reordering rules. Since a word in Tamil has multiple meaning in English, an effective word dictionary file is needed in order to achieve better results in translation.

This methodology for this machine Translation process from Tamil to English possesses accurate meaning. The output of the system is to be checked by means of Word error Rate i.e., No of words is to be substituted, deleted and also gets inserted. Its Precision can be measured in terms of no of correcting words to that of the total no of answers produced. Then the Recall can be measured by the means of correct answers to that of the possible correct answers produced. Based on these methods we easily translate the Tamil content into English contents successfully.

REFERENCES


Sasidhar, B., P.M. Yohan, Dr. A. Vinaya Babu, Dr. A. Govardhan, 2011. IJCSI International Journal of Computer Science Issues, 8(2).