QUALITY TRANSLATION USING THE VAUQUOIS TRIANGLE
FOR ENGLISH TO TAMIL

M.Mayavathi (dm.maya05@gmail.com)
K. Arul Deepa (karuldeepa@gmail.com)
Bharath Niketan Engineering College, Theni, Tamilnadu, India

ABSTRACT

The aim of this work is handling complex sentences and alignments of words. Hybrid Machine Translation is automatically
acquires knowledge from large amounts of training data at different languages. The system is to translate complex
sentence structures to process able chunks and translating the text English to Tamil. The system is first separates the
source text word by word with POS category and searches for their corresponding target words in the bilingual dictionary.
Rule Based Reordering, Morphological Analyzing, and dictionary based translation to the Target language. The transfer
rules for reordering from English parse tree with respect to Tamil help us to get the output in the syntactic pattern of
target language. The reordered output after morphological generation of Tamil words is displayed as the final output of
the machine translation system and then errors in the translated sentences are corrected by applying Statistical
technique.

1. INTRODUCTION

Machine Translation is a process of translating the sentences from one language to the other based on the information in
the Knowledge Base without human intervention. There are three approaches to machine translation: Statistical, Example
based and Rule based machine translation systems. Synchronous Tree Adjoining Grammar associated aligned tree/string
training data and a method of converting these grammars to a weakly equivalent tree transducer for decoding.

Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and
representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like
language processing for a range of tasks or applications. Also, the contents of the documents that are being searched will
be represented at all their levels of meaning so that a true match between need and response can be found, no matter
how either are expressed in their surface form. There is typically a well-defined problem setting, a standard metric for
evaluating the task, standard corpora on which the task can be evaluated, and competitions devoted to the specific task.

In earlier years the machine translation is done only at the word level i.e. word by word translation. This project is carried
out at many places for years but still need of a good translation system. Any basic translation requires two main view
points: First is the linguistic point of view and second is the mathematical point of view. The three major techniques
involved machine translations are [1] Rule Based, Statistical, and Example Based Technique. The Statistical and Example
Based Techniques needs parallel corpora for translation. In such cases adopting only the statistical technique will not
result in proper translation to the target language.

2. SYSTEM FEATURES

A Hybrid technique is developed for a system that generates simple sentences translation with part of speech tagging,
chunking and morphological generator, segmentation is done. The preprocessing tool for machine translation that
simplifies the complex sentences into simple sentences. This system uses rule based technique for sentence simplification
and uses characters such as (,, ,?”) as delimiters for sentence separation. They have designed this system as a
preprocessing tool for English to Tamil translation.

Hybrid machine translation (HMT) is leverages the strengths of statistical and rule-based translation methodologies.
Several MT companies (Asia Online and Systran) are claiming to have a hybrid approach using both rules and statistics.
The approaches differ in a number of ways: Rules post-processed by statics: Translations are performed using a rules
based engine. Statistics are then used in an attempt to adjust/correct the output from the rules engine. Statistics guided
by rules: Rules are used to pre-process data in an attempt to better guide the statistical engine. Rules are also used to
post-process the statistical output to perform functions such as normalization. This approach has a lot more power, flexibility and control when translating.

3. CURRENT WORK

The given source sentence is parsed and tagged using POS tagger, the tagged information is stored in a separate file. The rule based reordering of the sentence has to be done in the above formulated order using the tagged information. Chunking of the source sentence has to be done using the bi-gram model, and the bi-grams are translated into Tamil language by means of a word dictionary file. Then the word by word translation has to be done with the bilingual dictionary and if a word does not exist in dictionary, it may be a proper noun which is to be transliterated to Tamil language. Then apply gender ending rules to get the target output sentence. The error in the target sentence is corrected statistically using the file which contains collection of Tamil verbs with proper tense and gender endings. Finally the Tamil sentence for the corresponding English sentence is generated.

![Overall System Architecture](image)

4. SYSTEM METHODOLOGY

This paper presents an effective methodology for English to Tamil translation. Hybrid Machine Translation is handled by mapping from input to output sentence. Input is the English sentence which is enriched with segmentation, parsing and bilingual dictionary information. Output is a Tamil sentence with statistical error correction. The purpose is to group sequences of words are translated from source sentence to target sentence using hybrid techniques. The system can be translating complex sentences by creating new morphological reordering rules. Since a word in English has multiple meaning in Tamil, an effective word dictionary file (4500 words from English to Tamil) is used in order to achieve better results in translation. Gender ending verbs for all possible tenses were created for the purpose of statistically correcting the errors in the output sentences.

4.1 HMT Process

The process is acquires knowledge from training data and also enhance the input text with POS tagging and morphological information. After applying the local word grouping rules to the Tamil sentence (s), based on their four methods to process and align. (1) **Dictionary lookup approach (DL)** is used verbs and other groups are processed with DL
approach; HWGs with categories such as proper nouns, city, job-title, location, and country are processed with TS approach. (2) Transliteration Similarity is transliteration system maintains a consistent correspondence between the alphabets of two languages, irrespective of sound. Given two words, each from a different language, we define “transliteration similarity” as the measure of likeness between them. This could exist due to the word in one language being inherited or adopted by the other language, or because the word is a proper noun. Named entities such as city, job-title, location, country and proper nouns, all recognized by the local word grouping algorithm are compared using a transliteration similarity approach.

Neighbors approach works on this principle and aligns one or more words with one of the English words. Considering one HWG at a time, we find the nearest Tamil word that is already aligned with one or more English word(s). We assume that the words in English-Tamil phrases follow a similar order and align the rest words in that group accordingly. The algorithm retrieves expected English word(s) from the HWGs and tries to locate them in the English sentence. This approach can be useful to locate one or more English words that align with one or more Tamil words.

INPUT

Ravi waited for the train but the train was late.

OUTPUT

Segmentation and Tagging

The segmentation and tagging of the source sentence is done using the Parts Of speech tagger. In our work we use Stanford POS tagger for the tagging purpose. The English sentences are taken as an input to the parts of speech tagger. The tagger tokenizes each word in a sentence and identifies the parts of speech information such as verb, noun, adjective etc. of that word. Then the words and their tagged information are stored in a separate file which is used for reordering of sentences. Characters are arranged in document lines following some type setting conventions which we can use to locate characters and find their style. The above complex sentence can be split into simple sentences.

Segmentation: 1. Ravi waited for the train, 2. but the train was late

Tagging: Ravi // waited // for // the // train // but // the // train // was // late.

NNP  VBD  IN  DT  NN  CC  DT  NN  VBD  JJ

Rule Based Reordering

The tagged words are stored separately for the purpose of reordering according to the morphological structure of the Tamil language. The tagged words are arranged according to the order which is mentioned below UH/ PP/ WP/WRB/WD/T/ NNP/ PRP/ RB/ DT/ CC/ JJ/ PPS/ WP/JJR/ JJS/ IN/ NN/ NNS/TO/ VB/ VBD/ VBG/ VBN/ VBP/ VBZ/ MD. The above mentioned order suits all most all the types of simple sentence when reordering it from English to Tamil language.
Handling Complex sentence:

Noun, adjective and adverb clauses are considered.

Step 1: Conversion of complex to minimal sentence by grouping the clauses
Step 2: Minimal simple sentence can be analyzed as mentioned earlier
Step 3: Integration of clauses into the minimal simple sentence

4.2. Morph Analyzer/ Morph generator:

The source text is passed to the morphological analyzer. Morphological analyzer extracts the root word and its feature equations. These feature equations will be used in the later part to generate or add proper inflections to the target language. The sole purpose of this module is to handle the morphology of the target language. Features stored in the target structure might be needed for producing the proper inflected target lexical form.

4.3. Transliteration

The transliteration is the process of labeling the text in one language with other. In English to Tamil transliteration, the English text is replaced with the Tamil text by preserving the spell. The SVM based Multilingual Amrita English- Tamil Transliteration tool is developed by Amrita – CEN and we use the same in the machine translation system. First the corpus of English words are collected and preprocessed. The preprocessing involves two level Romanization, segmentation and alignment. The English words are romanized into Tamil words, by English- Tamil mapping. The romanized Tamil words are again romanized back to English, by Tamil - English mapping.

4.4. Statistical Error Correction Method

Even though we write gender ending rules, in some cases accurate verb with proper gender ending cannot be obtained for Tamil language. Particularly when writing rules for past tense sentences many contradictions arises. In such cases there occurs an error in the target sentence. For example consider a source sentence “Ravi waited for the train but the train was late” for that we may get the target sentence .we have around 70 base verbs in Tamil language with all possible gender ending and tenses.

Wrong Sentence

Correct Sentence
5. APPLICATIONS

The list of some of the most commonly researched applications of machine translation. There is typically a well-defined problem setting, a standard metric for evaluating the task, standard corpora on which the task can be evaluated, and competitions devoted to the specific task. The problem of sentence understanding deals with understanding individual sentences, and determining their meaning in the context of preceding sentences. The problem is divided into three stages: semantic parsing, semantic classification, and discourse modeling. Information retrieval (IR) is the area of study concerned with searching for documents, for information within documents, and for metadata about documents, as well as that of searching relational databases and the World Wide Web. The FST based morphological analyzer and generators are widely implemented for many languages. Automatic summarization involves reducing a text document or a larger corpus of multiple documents into a short set of words or paragraph that conveys the main meaning of the text. Summarizations of multimedia documents are used in education, website and etc. Greatly speedups syntactic analysis Tagging is local and No need to process the whole sentence to find that a certain tag is incorrect. The application of POS Tagging is Senses depend on semantic context and less structured, longer distance dependency.

6. CONCLUSIONS

The overall design, architecture, functions, and translation methodologies are presented and reviewed in details. The system applies Translation Corresponding Tree structure for annotating bilingual and Constraint Synchronous Grammar for analyzing the syntax of bilingual texts. A major drawback with the statistical model is that it presupposes the existence of an aligned parallel corpus. The work done based on limits to translation of complex sentences from English to Tamil. The sentences are subdivided into words using Word-Based Translation models and words are aligned corresponding to translation models. Tree can be constructed based on word alignment. for example, are generally rather loosely translated - one sentence in the source language is often split into multiple sentences, multiple sentences are clubbed into one, and the same idea is conveyed in words that are not really exact translations of each other. If tokenization creates a one-to-one mapping, the number of tokens in both languages should be the same by adjusting this parameter. The system with all the necessary modules are in place, scalability is a key to improve its performance. Transliteration, Morph-synthesizer and extracting features are on its own a big task and these have to be enhanced as well to improve the overall performance of the system along with the root word lexicon and the reordering rules. The bilingual dictionary lacks the word sense information, so the semantic ambiguity arises in the system for many words.

7. FUTURE ENHANCEMENTS

In future works increasing the re-ordering rules; increase the database entries, fine tuning the Morph generator, and scalability. The system can be translating complex sentences by creating new morphological reordering rules. An effective word dictionary file (4500 words from English to Tamil) is used in order to achieve better results in translation. Gender ending verbs for all possible tenses were created for the purpose of statistically correcting the errors in the output sentences. The system is implemented using java codes. Multiple parse trees are used handled by the Stanford parser and the dependency parser is also used in the translation system. To handling the verbal phrases is possible by the system. The transliterator is limited to the Indian place names so performance of transliterator is very low when it’s used for vocabulary words which are not present in the database. The morph generator is implemented for certain cases but the dependency information of many inflectional categories is given by the parser, such cases works well in morph generator and translation of sentences. The reordering rules are confined to the nodes of the branches and same rule could be handled for different cases with same syntactic structure. Question type of sentences are handled it is also one of the limitation of system.

REFERENCES

7. “Machine Translation System for Indian Languages”, (Latha R. Nair, David Peter S- 2012)
10. “Segmentation and Alignment of parallel text for statistical machine translation”, (Yonggangdeng.-2007)
11. “Synchronous Tree Adjoining Machine Translation”, (Steve DeNeefe and Kevin Knight-2009)
12. “Semantic Role based Tamil Sentence Generator”, (S. Lakshmana Pandian, V.Geetha-2010)