Rule Based Statistical Hybrid Machine Translation

S. R. Priyanga, AP, A. AzhaguSindhu, AP

Abstract- Language is the main form of human communication. Translation is essential for co-operation among communities that speaks different languages. Machine Translation refers to the use of computers to automate the task of translation between human languages. Machine Translation performs its operation based on the available examples, which failed to overcome certain ambiguities like mapping of multiple words, idiomatic usages, phrasal verbs, structural ambiguity thus a statistical approach to the machine translation was proposed. Systems are designed either for two particular languages (bilingual systems) or for more than a single pair of languages (multilingual systems). Bilingual systems may be designed to operate either in only one direction e.g. from Tamil into English, or in both directions. Multilingual systems are usually intended to be bidirectional; most bilingual systems are unidirectional. This system is designed as bilingual system, i.e., converting English sentence to Tamil sentence using particular rules.

Keywords- idiomatic usages, phrasal verbs, structural ambiguity

I. INTRODUCTION

Rule Based Statistical Hybrid Machine Translation focuses on development of Statistical Machine Translation (SMT) system. Parallel corpuses of both the Languages are provided. Intuitively, for translation between English to any other Indian language, the linguistic knowledge about the correlations must be known because of the vast structural and lexical differences between the two languages. Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability. The key problems in statistical MT are: estimating the probability of a translation, and efficiently finding the sentence with the highest probability. The objective of this project is to translate the given English sentence into Tamil language. Although a small number of English to Tamil MT systems are already available, the outputs produced by them are not of high quality all the time, therefore through this work we intend to analyze the difficulties that lead to this below par performance, and try to provide some solutions for them. The most frequently cited benefits of statistical machine translation over traditional paradigms are better use of corpus; there is a great deal of natural language in machine-readable format.

The translation process converts a text in one human language to another which preserves not only the meaning, but also the form, effect and style.

Manuscript received on April, 2013.

S.R.Priyanga, AP, Department of Computer Science, Info institute of Engineering, India.

A.Azhagu Sindhu, AP, Department of Computer Science, Info institute of Engineering, India.

There are some countries in which more than one language is spoken but there is not enough human translators are available. So a scheme for automatic translation between two languages is very desirable for social and political interactions. Therefore, SHMT systems are not tailored to any specific pair of languages. There are several other major languages (e.g., Bengali, Punjabi, and Gujrathi) in the Indian subcontinent. Demand for developing MT systems from English to these languages is increasing rapidly.

II. RELATED WORK

Machine Translation basically performed by using four approaches namely, Transfer Approach , Interlingua Approach, Direct Approach, Corpus-based Approach.

1. Transfer Approach

Transfer model involves three stages: analysis, transfer and generate. Analyze the source sentence, transfer the structure of source sentence to the structure of target sentence finally translate the word, number, gender in the target words. But in this approach n generating components, n analysis components and n(n-1) transfer components are needed for n language translation, it will increase memory and working principle.

2. Interlingua Approach

The interlingua approach considers MT as a two stage process: Extracting the meaning of a source language sentence in a language-independent form, and, Generating a target language sentence from the meaning. In this approach burden on the analysis and generation components increases. We have to choose between various possible parses for a sentence, identify the universal concepts that the sentence refers to, and understand the relations between various concepts expressed in the sentence.

3. Direct Approach

Direct approach involves in four stages to translate any language to other language. Morphological analysis can be done i.e., identified the tense for the verb then Identify the constituents and Reorder the constituents based on target. Replace the source words to target with the help of dictionary. But direct approach in not a minimal structure and semantic analysis also won't produce a long term solution for MT.

4. Corpus-based Approach

Corpus-based Approach uses training corpus to guide the translation process. A parallel corpus consists of two collections of documents: a source language collection,and a target language collection. This approach processed in two ways, i.Example Based Machine Translation i.e., Split the sentence into sub-sentence, use the already translated sentences as example and fetch the sub-sentences from the example sentences.

ii.Statistical Machine Translation i.e., Statistical MT models take the view that every sentence in the target language is a translation of the source language sentence with some probability. The best translation, of course, is the sentence that has the highest probability.

III. PROPOSED SYSTEM

A. Approaches

The *RBSHMT approach* combines the three concepts of

1.Statistical Rule based approach

The rule based approach requires a language expert frame linguistic rules according to the grammar of the source and target languages, but we proposed this statistical approach that dynamically generates the grammar rules without language expert and bilingual dictionary.

2.Example Based Machine Translation (EBMT)

The phrasal part from the given sentence is extracted and matched with the existing sentence thus reducing the ambiguity.

s3.Probability estimation model (statistical approach)

This is implemented from the given amount of bilingual corpus.

Source side corpus can be processed with pre-processor, pos-tagged and removal of special character, which was not processed in target side corpus. Working of the system is, the text given as input is pre-processed (pos-tagged, identification of phrase and removal of special characters). This is done by giving the input sentence to an application interface. Then the pos-tagged sentence is aligned according to the sentence structure of the target language (Tamil). Translation involves replacement of the identified phrasal part, the word with highest probability. Words are picked for matching according to the order in which they are tagged. Matched words with highest probability is chosen from the database and replaced. Words which are not contained in the database are transliterated.

B. System Architecture

System Architecture encompasses the following, Sentences from the Bilingual-Corpus can be taken as a input for both language and translation model. Language model processed N-gram models. Outputs of language model can be stored in database and it can be taken as a input for decoding module. Translation model contains 3 models namely, T-Parameter, Distortion and Fertility. Output of these 3 models are stored in database and it can be taken as a input for decoding module. Decoding module combine the language, translation modules outputs and also fetch sentences from closed word table that is bag of words.

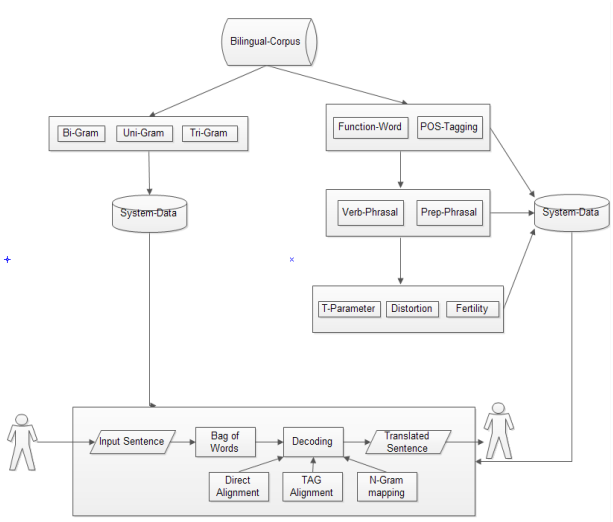

Figure 1 : System Architecture

C. Module Discription

The process of RBSHMT is divided into two stages containing seven modules through which a translated text in Tamil can be obtained for the entered English text.

Training Stage

1.Language Module

Language modelling is the task of assigning a probability to each unit of text. In the context of statistical MT, as described in the previous chapter, a unit of text is a sentence. N-gram can be processed in language modeling. In an N-gram model the probability of a word given all previous words is approximated by the probability of the word given the previous N-1 words. The approximation thus works by putting all contexts that agree in the last N-1 words into one equivalence class. With N = 2, we have what is called the bigram model, and N = 3 gives the trigram model. That is, given a sentence e, our task is to compute P(e). For a sentence containing the word sequence w1w2...wn, we can write P(e)without loss of generality,

$$P(e)=P(w1w2...wn)=P(w1)P(w2|w1)P(w3|w1w2)$$
$$... P(wn|w1w2....wn-1)$$

2. *Pre-Processing*

It is in this stage the corpus is refined from the unwanted characters it could even be the special charters like comma, semi-colleen, apostrophe, full-stop, quotation (single and double) and hyphen. Irrespective to the language, both the language pair is given as input in this process.

Input:- Raw input sentence that is to be trained

Output:- Pre-processed sentence after removing multiple spaces and symbols.

After clean-up the corpus Pre-Processing contains two more Processes,

i.Functional words – Data base can be maintained for those function words are called open class words or lexical words or auto semantic words and helps create meaning in sentences.

ii.Pos-tagging - Part-of-speech tagging is harder than just having a list of words and their parts of speech, because some words can represent more than one part of speech at different times, and because some parts of speech are complex or unspoken.

There are 9 parts of speech in English: noun, verb, article, adjective, preposition, pronoun, adverb, conjunction, and interjection. However, there are clearly many more categories and sub-categories. Stanford tagger and tokyo tagger are used for pos-tagging.

3.Phrasal Module

The phrasal module is a process where the phrasal verbs and phrasal prepositions in the English are tagged into single words. This work is done as the combined phrasal words in English give a single meaning or at times does not containing any meaning in Tamil. The basic principle behind the idea to combine the words is based on the EBMT. There are two types of phrasal are performing namely Phrasal Verb, Phrasal Preposition

i.Phrasal Verb - The technique to develop a trained corpus should contain a comparable knowledge of grammatical usage of the language pair, thus helping the system in building the rules to combine the verbs. A pattern is set; so as to check in sequence for the words ending with /VBD + /VBN, /VBP+ /VBN and /VBZ+/VBN.

ii.Phrasal Preposition - The concept of case in languages refers to the phenomenon of expressing reciprocal relations of prepositions by means of nouns present next to it. Check in sequence for the words with /IN+/NN, /IN+/NNP, /IN+/PRP$

Before phrasing:

Sisodiya/NNP Rani-Ka/NNP -/: Bagh/NNP was/VBD built/VBN by/IN Sawai/NNP Jai/NNP Singh/NNP II/NNP for/IN his/PRP$ Queen/NNP

After phrasing:

Sisodiya/NNP Rani-Ka/NNP Bagh/NNP was@built/TYPE Sawai/NNP Jai/NNP Singh/NNP II@by/INNN his@for/PRIN Queen/NNP

4. Translation Module

The role of the translation module is to find, the probability of the source sentence F(English) given the translated sentence E(Tamil) using the t-parameter, information about the number of target words for one source word using the fetility and the position of a word in a sentence using the distortion.

i.T-Parameter - T-parameter is responsible in finding the P (f|e) - the probability of the source word given the translated word. Intuitively, P (e|f) should depend on two factors:

1. The kind of sentences that is likely in the language E. This is known as the language model - P (e).
2. The way sentences in E get converted to sentences in F. This is called the translation model - P (f|e). Using the Bayes rule, e1 is reduced to e1 => arg max ((P (e) P (f|e))/ P(f))

Input :- Pre-processed sentence

Output :- A source words with combination of target words are stored in database with their probabilities.

ii.Distortion – Distortion is the process of finding the position of target for the particular position occurrence of source word.

Input:- Pre-processed tagged sentence and t-parameter table

Output:-This parameter will model the bilingual data and outcome is stored in distortion table.The distortion probability will tell you that source word in position 2 (of a source sentence) will generate a target word that winds up in position 5(of a target translation).

iii.fertility - Fertility of a source word gives the probability that a particular source word will produce how many target words whenever that source word appears in sentence for translation.

5. Post-Processing

Translated words for a particular English word can be up to 20 records. But these 20 records are not taken into consideration. Therefore a threshold is fixed in this process to mark the limit of records for consideration.

Development Stage

1.Translitration

The development stage starts with the transliteration followed by the decoding part. Transliterations are used in situations where the original script is not available to write down a word in that script, while still high precision is required. Transliteration attempts to use a one-to-one correspondence and be exact, so that an informed reader should be able to reconstruct the original spelling of unknown transliterated words.

2.Decoding

Decoding is the stage in which the user input is processed; it is here we perform the user obtain the translated output. Every word from the user given sentence is processed using the available references like direct alignment, TAG alignment and finally using the n-gram mapping.

i.Direct alignment - The usage of closed words for the given input sentence is done in the direct alignment along with the heuristic rules for direct alignment. The input sentence is tagged using the Pos-tagger, using these tagged words, the preposition, conjunctions and the delimiters are identified to replace them from the closed words table.

ii. Tag alignment - The remaining tagged words are matched using the t-parameter reference; a translated word with the highest probability is obtained from the table.

iii. n-gram mapping - The translated words are further checked from the n-gram mapping to verify the next occurrence of the target word.

IV. CONCLUSION AND FUTURE WORK

Traditional MT techniques require large amounts of linguistic knowledge to be encoded as rules. RBSHMT provides a way of automatically finding correlations between the features of two languages from a parallel corpus, overcoming to some extent the knowledge bottleneck in MT. A major drawback with the statistical model is that it presupposes the existence of a sentence-aligned parallel corpus. For the translation model to work well, the corpus has to be large enough that the model can derive reliable probabilities from it, and representative enough of the domain or sub-domain (weather forecasts, match reports, etc.) it is intended to work for. The system at times is limited to complex sentence. Data insufficiency is one of the major problem in this system as it is totally dependent on the corpus. Some the example are For e.g.:- 1.“parvati is standing there”, the translated output from our system is “பார்வதி அங்கு . நிற்றல்”, but the expected output is “பார்வதி அங்கு நிர்க்கிறாள்”- contains grammatical mistake. 2.

"my name is sachin", the output from our system is "என்னுடைய பெயர் சச்சிண்-", but the expected output is "என்னுடைய பெயர் சச்சின்"- contains error due to insufficient data. Another issue is that most evaluation of statistical MT has been with training documents that are very rigid translations of each other.

News articles and books, 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. In such situations, sentence-alignment itself might be a big challenge, let alone word-alignment. Since RBSHMT is sensible to tag alignment (with probabilities), it can be used for lexicon acquisition also, apart from the larger goal of MT. Statistical MT techniques have not so far been widely explored for other Indian languages like Sanskrit. It would be interesting to find out to what extent these models can contribute to the huge ongoing MT efforts in the country.

REFERENCE

1. Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Frederick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin, A Statistical Approach to Machine Translation, Computational Linguistics, 16(2), pages 79–85, June 1990.
2. Peter F. Brown, Stephen A. Della Pietra, Vincent J. Della Pietra, and Robert L. Mercer, The Mathematics of Statistical Machine Translation: Parameter Estimation, Computational Linguistics, 19(2), pages 263–311,June 1993.
3. Shachi Dave, Jignashu Parikh, and Pushpak Bhattacharyya, Interlingua-based English-Hindi Machine Translation and Language Divergence, Journal of Machine Translation (JMT), 16(4), pages 251–304, 2001.
4. Harold L. Somers, Example-based Machine Translation, Machine Translation, 14, pages 113–157, 1999.
5. Ananthakrishnan Ramanathan, Pushpak Bhattacharya, Statistical Machine Translation, pages 6-10, 2002.