BILINGUAL TRANSLATION SYSTEM

(FOR ENGLISH AND TAMIL)

Dr. S. Saraswathi
Associate Professor
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

P. Kanivadhana
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

M. Anusiya
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

S. Sathiya
Department of Information Technology
Pondicherry Engineering College
Puducherry 605014, India

***Abstract*--- The project aims in developing Bilingual Translation System for English and Tamil using hybrid approach. We use Rule Based Machine Translation (RBMT) and Knowledge Based Machine Translation (KBMT) techniques. Since it's a bilingual translation system both English to Tamil and Tamil to English translation are possible. The source text is analyzed. The simple sentences are translated using the rules in RBMT. The complex sentences are split into simple sentences using KBMT and translated using RBMT and then processed to get text in target language. It is restricted to the domain Weather Report and can be expanded to other domains in future.**

Keywords- Machine translation, Win tree tagger, RBMT, KBMT

1. INTRODUCTION

India has a linguistically rich area—it has 18 constitutional languages, which are written in 10 different scripts. Tamil is the most commonly used language of the south. English is very widely used in the media, commerce, science and technology and education. Many of the states have their own regional language, which is either Tamil or one of the other constitutional languages. Only about 5% of the population speaks English.

Use of mother tongue is inevitable for creative thinking and acquisition of world knowledge. As such, in the recent past, people have given more attention to develop machine translation systems from English to other languages than learning of English by everybody.

In such a situation, there is a big market for translation between English and the various Indian languages. Currently, this translation is essentially manual. Use of automation is largely restricted to word processing. Two specific examples of high volume manual translation are—translation of news from English into local languages, translation of annual reports of government departments and public sector units among, English, Tamil and the local language.

II. EXISTING MACHINE TRANSLATION SYSTEMS

Machine Translation is a field where a text/speech from one natural language is translated into another. The advantages of MT are that is it's automated as there is no human intervention in translating the sentences to target language. There are many techniques in MT system, where in different types of sentences are translated into the target language avoiding to their structure and meaning.

In broader sense, machine translation approaches can be classified into three categories, namely, statistical approach, example based approach and rule-based approach. The Statistical approach uses some statistics such as mean, variance on bilingual text corpora to find the most appropriate translation. The Example-based approach is often characterized by its use of a bilingual corpus with parallel texts as its main knowledge base.

A hybrid translation system uses more than one technique for translating different types of sentences to the target language. A hybrid MT system also includes pre editing and post editing.pre editing is done before the

input sentence is actually translated into the target language. The input sentence is formatted so that it can be fed to the actual MT system for actual translation. Post editing is done after the translation. This is done in order to bring the sentence to a correct format in the target language. Hence hybrid MT system involves more than one technique.

A large number of machine translation systems have been developed under above three broader heading. Thenmozhi D and Aravindan C [1] , translates Tamil to English using statistical machine translation system. It developed a CLIR system in Agriculture domain for the Farmers of Tamil Nadu which helps them to specify their information need in Tamil and to retrieve the documents in English.

For instance, Apertium [6], [7] is rule-based MT system that translate related languages. This is an open – source system that can be used to translate any related two languages. This MT engine follows a shallow transfer approach and consists of the eight pipelined modules, such as de-formatter, A morphological analyzer, A part-of-speech(PoS) tagger, A lexical transfer module, A structural transfer module, A morphological generator, A post-generator, and A re-formatter.

Anglabharati [5] deals with machine translation from English to Indian languages, primarily Hindi, using a rule-based transfer approach. The primary strategy for handling ambiguity/complexity is post-editing—in case of ambiguity, the system retains all possible ambiguous constructs, and the user has to select the correct choices using a post-editing window to get the correct translation. The system's approach and lexicon is general-purpose, but has been applied mainly in the domain of public health.

Electronic Dictionary Research (EDR) [8], by Japanese, is the most successful machine translation system. This system has taken a knowledge-based approach in which the translation process is supported by several dictionaries and a huge corpus. While using the knowledge-based approach, EDR is governed by a process of statistical machine translation. As compared with other machine translation systems, EDR is more than a mere translation system but provides lots of related information.

III. PROPOSED SYSTEM

The proposed system is a hybrid machine translation system. We consider Rule Based and Knowledge Based machine translation techniques. In the proposed system the source language text is given as input. It is then given to the morphological analyzer, where the morphological analyzer returns the part of speech of each word in the sentence.

If the input sentence is complex, then it is given as input to KBMT and where it is split into simple sentences which are then given as input to RBMT. It is then translated to the target language using RBMT. If the input sentence is simple, then it is directly translated using RBMT.

In the Rule Based Machine Translation System, the tagged text is given to the sentence type is analyzer. The rule database consists of various sentence patterns. The text and matches with the patterns found in the rule database.
The individual words are then mapped into the target language, present in the Bilingual Dictionary. Bilingual Dictionary consists of equivalent words in the source and target languages. In Bilingual Dictionary, the verb and nouns are stored in separate tables so as to avoid ambiguity. For instance, in the sentence above, the root words of both word are same. i.e. one we store the verbs and nouns in separate tables, we can easily resolve the conflict in searching the words and can thus reduce ambiguity.

If the sentence is complex, the sentence is first given to the Knowledge Based Machine Translation system. Here, the sentence is changed into its equivalent simple sentence. The simple sentence is then translated to the target language using the Rule Based Machine Translation System. The complexity of the sentence is decided by the test cases that we used for analyzing our proposed system. Possible Test cases are sentences for analyzing the KBMT contains sentences with "and", "due to", "because of", "in the case of", etc.

In the Knowledge Based Machine Translation System, the source language text is first subjected to the process of tokenization. In this process, each word in the sentence is extracted. These words are then tagged so as to find the part of speech of each word in the input sentence. These tagged words are then given as input to the process of Lemmatization. In Lemmatization the meaning of the sentence is analyzed and is split/converted to form simple sentences based on its syntactic meaning.

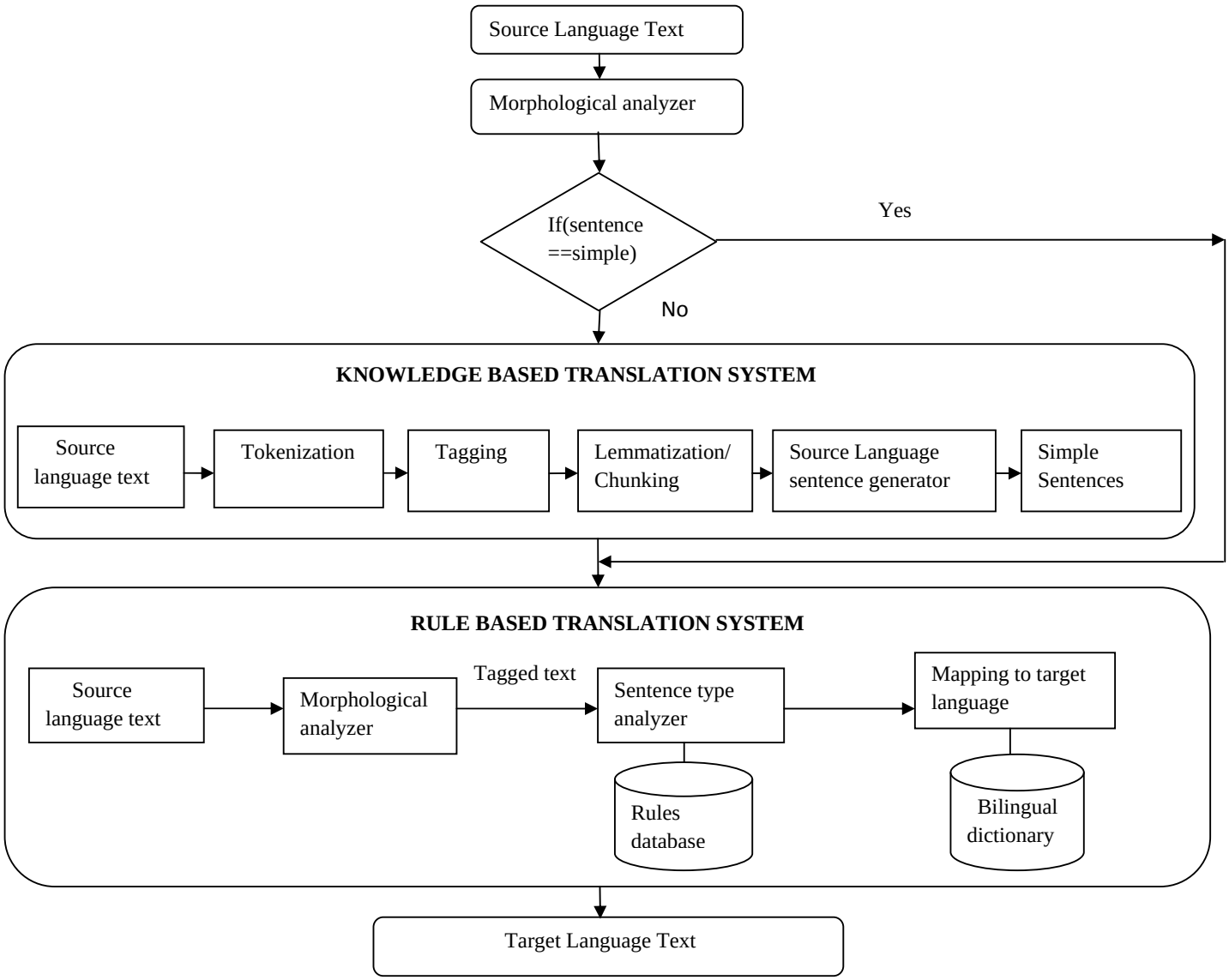


Figure.1 The architecture of the proposed system

4. Rule Based Machine Translation System:

4.1 English to Tamil Machine Translation:

English sentences which are to be translated are stored in a text file as given below.

File name-Input.txt:

Heavy rain affects land transportation across Tamil Nadu.

Heavy snow grounds 800 flights across Kashmir.

This file is given as input to the Morphological analyzer. The output of the Morphological analyzer is again a text file. This output file specifies the words with its parts of speech. The output file is shown below:

Filename-moroutput.txt:

```
Heavy       JJ      heavy
rain        NN      rain
affects     VVZ     affect
land        NN      land
transportation      NN      transportation
across      IN      across
TamilNadu           NP      <unknown>
.           SENT    .

Heavy       JJ      heavy
snow        NN      snow
grounds     NNS     ground
800         CD      @card@
flights     NNS     flight
across      IN      across
Kashmir     NP      Kashmir
.           SENT    .
```

This output file is read by the java program. Each sentence is stored in two dimensional array.

Heavy	*JJ*	*heavy*
snow	*NN*	*snow*
grounds	*NNS*	*ground*
800	*CD*	*@card@*
flights	*NNS*	*flight*
across	*IN*	*across*
Kashmir	*NP*	*Kashmir*

Here the pattern of the sentence is JJ NN NNS CD NNS IN NP. Similar patterns of sentences with same type of output pattern in Tamil are chosen. And the rules are written for those sentences.

Example:

Heavy snow grounds 800 flights across Kashmir.

In the above example, the words are translated as follows.

Heavy → பலத்த
Snow → பனி
Grounds(present tense)
Flight → விமானங்கள்
Kashmir → காஷ்மீர்

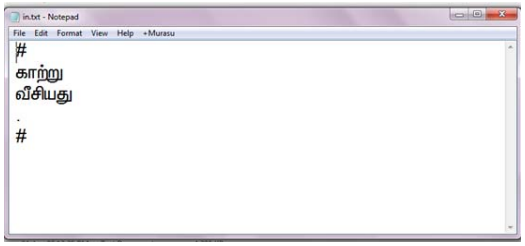
Here the noun heavy snow is translated as பலத்த பனி. The verb ground is translated as . தரையிறக்கு . Since the tense is present the verb is modified as தரையிறங்கியது
Since the preposition is 'across', Kashmir is translated as காஷ்மீரில். Therefore the output sentence is காஷ்மீரில் பலத்த பனியால்| 800 விமானங்கள் தரையிறங்கியது.

The words in English and corresponding meaning in Tamil are stored in the Database. The words are split depending on its nature as noun and verb. The verbs are given meaning in Tamil by considering its tense.

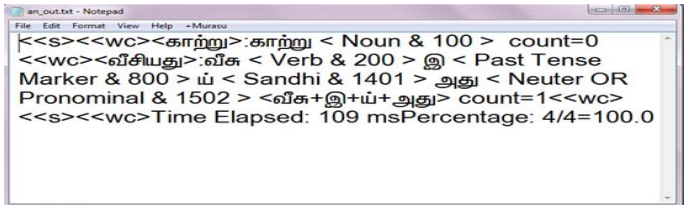
By using jdbc connection the meanings in Tamil are retrieved and the output sentences are framed and stored in a string variable

4.2 Tamil to English translation system:

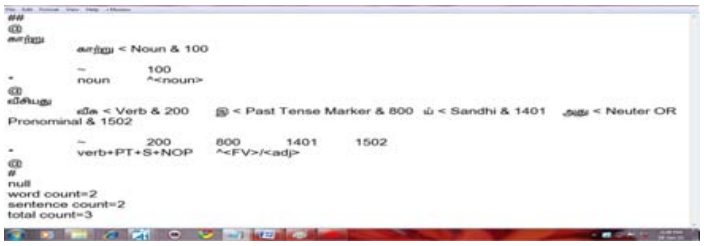
The input tamil sentence is given as input as follows



Input files is given as input to Tamil Morphological analyzer. Tamil analyzer generates three output files. *File1:*



File2:

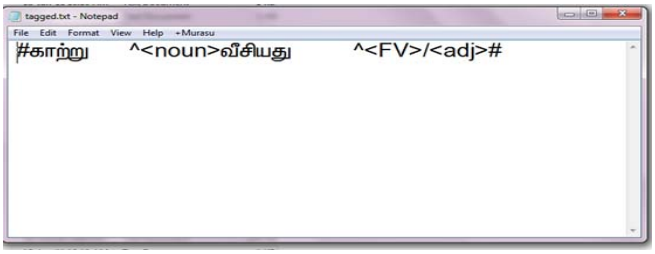


The root word and its tense is identified from file2 as

வீசியது → வீசு(root word)

Verb+PT+S+NOP indicates the verb is past tense.

File3 :



The input sentence pattern is recognized from file3 as

<noun><FV>/<adj>.

Similar sentences with same patterns are grouped and rules are written.

5. Knowledge Based Machine Translation System:

Input sentence:

Due to heavy rain, land transportation is affected in Chennai.

The output of the Knowledge Based Machine Translation System is

Heavy rain affects land transportation.

Since, the above sentence pattern is present in RBMT, it can be translated into target language sentence using RBMT.

IV RESULTS AND ANALYSIS:

The following table summarizes the results obtained from the proposed work. The parameters that we use for analyzing the result are Precision and Mean opinion score.

MOS(Mean Opinion Score):

Mean of the remarks obtained from the users of the system.

$$MOS = (OS_1 + OS_2 + OS_3 + ...OS_n) / n$$

Where OS is Opinion Score.
MOS and OS range from 0 to 10.

Precision:

No. of sentences translated correctly divided by the no. of sentences given as input.

Precision= (number of sentences translated correctly) / (total number of sentences given as input)
It ranges from 0 to 1.

TABLE I. PERFORMANCE MEASURE FOR SIMPLE SENTENCE

Sentence type	Total no. of sentences given	Output obtained	Precision (%)	Mean opinion score
PP VBD VVG	15	14	93.3	9.2
PP VBZ VBG	10	8	80	8.5
DT NN VHZ VVN	14	10	71.42	8.4
NN MD VV	17	15	88.23	8.7
DT NN VVD	12	9	75	8.9
NN NN IN DT NP NN	10	8	80	9.0
NN VHP VVN RB	16	15	93.75	8.4
DT JJ JJ NN NN VHP VVN	8	6	75	8.2
DT JJ NN NN VHP VVN RB	17	14	82.3	9.2
DT NN VBD JJ	14	12	85.71	
JJ NN NNS CD NNS IN NP	10	8	80	8.0
NP VVD NN IN NP	15	13	86.6	8.3
NP VVA JJ NN	12	11	91.66	9.0
NP VVD JJ NN	11	8	72.72	8.9
JJ VVD VVN IN NP	10	9	90	9.4
JJ NN VBZ VVN IN NP	12	10	83.33	8.6
NP VVZ NN	10	8	80	8.0
JJ NN VBZ VVN IN NP	8	6	75	9.7
DT JJ NN VVN VVD CD NN	6	5	83.33	8.6
DT JJ NN VVN VBZ CD NN NN	4	3	75	8.8
NN VVD VBD CD NN NN	6	5	83.33	9.8
NN VVN VBZ CD NN	10	9	90	9.2
JJ NN VVD DT VBD CD NN	12	10	83.33	8.5
NN VVN VBZ CD NN IN NNS	5	4	80	9.3
JJ NN VVD DT NN VBD CD NN	10	8	80	8.2

Precision of RBMT = 82.36.

TABLE II PERFORMANCE MEASURE FOR COMPLEX SENTENCE

Sentence type	Total no. of sentences given	Output obtained	Precision (%)	Mean opinion score
NN VVN VBZ CD NN NN CC DT DT JJ NN NN VVD DT NN	5	4	80	8.5
JJ TO JJ NN , NN NN VBZ VVN IN NP	6	5	83.33	8
NN VVD VBD NN CC DT VBZ DT JJ NN IN NN VVD DT NN	5	4	80	7.9

Precision of KBMT = 81.11.

PRESION OF ENTIRE SYSTEM = 81.735.

V CONCLUSION:

We presented a Bilingual Translation System which translates given input sentence in source language into target language using hybrid approach. New rules have been added to the proposed system in order to make the system more efficient. This work can be extended to other domains with the addition of new rules.

REFERENCES

[1] Thenmozhi D and Aravindan C, "Tamil-English Cross Lingual Information Retrieval System for Agricultural Society", Department of Computer Science & Engineering, SSN College of Engineering Chennai, India,2009.

[2] Pattabhi R. K. Rao and Sobha L, "AU-KBC FIRE2008 Submission - Cross Lingual Information Retrieval Track: Tamil-English", First Workshop of the Forum for Information Retrieval Evaluation (FIRE), Kolkata. pp 1-5 , 2008.

[3] Saraswathi S, Asma Siddhiqaa M, Kalaimagal K and Kalariarasi M, "Cross Lingual Information Retrieval System for English, Tamil and Hindi", Department of Information Technology, Pondicherry Engineering College, 2009.

[4] Manoj Kumar Chinnakotla and Om Damani P, "Experiences with English-Hindi,English-Tamil and English-Kannada Transliteration Tasks at NEWS 2009",Department of Computer Science and Engineering, IIT Bombay, Mumbai, India.

[5] R. Mahesh K. Sinha: Integrating CAT and MT in AnglaBharti-II architecture, 10th EAMT conference "Practical applications of machine translation", May 2005, pp235-244.

[6] Apertium Machine translation system, http://www.apertium.org/

[7] Francis M.Tyers, Linda Wiechetek, & Trond Trosterud: Developing prototypes for machine translation between two Sámi languages.EAMT-2009: Proceedings of the 13th Annual Conference of the European Association for Machine Translation, ed. Lluís Màrquez and Harold Somers, 14-15 May 2009, Universitat Politècnica de Catalunya, Barcelona, Spain; pp.120-127

[8] Toshio Y, "The EDR electronic dictionary", Communications of the ACM, Volume 38, Issue 11, 1995, pp. 42 – 44.

[9] Vithanage N. V. C. T., English to Sinhala Intelligent Translator for Weather forecasting domain, Thesis submitted BIT degree, University of Colombo, Sri Lanka, 2003.

[10] Manoj Kumar Chinnakotla, Sagar Ranadive, Pushpak Bhattacharyya and Om P.Damani, "Hindi and Marathi to English Cross Language Information Retrieval at CLEF 2007", in the working notes of CLEF 2007.

[11] Pattabhi R. K. Rao and Sobha L, "AU-KBC FIRE2008 Submission - Cross LingualInformation Retrieval Track: Tamil-English", First Workshop of the Forum forInformation Retrieval Evaluation (FIRE), Kolkata. pp 1-5, 2008.

[12] Prasad Pingali and Vasudeva Varma, "IIIT Hyderabad at CLEF 2007 - Adhoc Indian Language CLIR task", in the working notes of CLEF 2007.

[13] Prasenjit Majumder, Mandar Mitra Swapan parui and Pushpak Bhattacharyya, "Initiative for Indian Language IR Evaluation", Invited paper in EVIA 2007 Online Proceedings.