A STUDY OF MACHINE TRANSLATION METHODS AND THEIR CHALLENGES

Anand Ballabh[1], Dr. Umesh Chandra Jaiswal[2]

[1]M.tech (IT), Dept of Computer Science & Engg, M.M.M U.T, Gorakhpur, (India)

[2]Associate. Professor, Dept of Computer Science & Engg, M.M.M U.T, Gorakhpur, (India)

ABSTRACT

Machine Translation methods are different and each has its own benefits and drawback. No translation tools can generate an exact version of source language but gives gist of information which can utilize to find the type of information contained n the source text. Sometimes, it is necessary to perform post-editing by in-house linguistic after generating translation output with translation engine. This work explains various approaches used in machine translation process such as Dictionary based, Rule based, Corpus Based and Hybrid Translation methods. This paper concludes with the assumption that no perfect translation systems exist, even though Hybrid method is better than that of all available methods because it combines the advantages of various translation methods.

Keywords: Machine Translation, Hybrid Machine Translation, Rule Based, Corpus Based, Statistical, Computational Linguistics, Language Translation

I. INTRODUCTION

The idea of language translation is developing currently that solves the issues of linguistic diversity. It is not possible to know and grasp all the languages within the world by human beings. Around 5000 languages present in the world that shows the need of language translation methods and its developments Researches within the field of language translation are exploring the possibilities of message transferring from one language to different. Government agencies and research institutes are providing initiatives to develop tools for machine-controlled text translation, which might be effective for international business communications into information professionals to improve their information services. Machine translation is the part of computational linguistics that studies the use of software tools to translate text or speech from one language (source language) to another (target language). Most recently, machine translation tools achieved translation excellence. Dictionary based machine translation was the first generation of automated language translation and it was purely based on electronic dictionaries. It translates the phrases but not sentences. Next, Rule Based Machine Translation (RBMT) systems, Corpus Based systems and Hybrid Machine Translation systems were developed. RBMT builds linguistic rules based on morphological, syntactic and semantic information related to source and target language. At the same time, Corpus Based systems generate translations from bilingual text corpora. Hybrid method is advanced method that combines the benefits of individual techniques to attain an overall better language translation.

II. WHERE WE ARE USING MACHINE TRANSLATION?

Language translation systems facilitate the individuals to communicate each other from different places so they can utilize the advantages of information and communication technology. Machine translation is widely employed in numerous applications and a few translation agencies including government agencies are supporting implementation of tools . Translation tools will primarily used for conducting research by reviewing foreign websites and articles. In addition, marketing, legal purposes, software localization, email translation for customer enquiries, website translation, manuals and documents translation, customer support, personal communication like travel reservations, managing assets abroad etc are possible with MT software.

III. MEASURES FOR SELECTING MACHINE TRANSLATION TOOLS

Accuracy and speed of translation are two main measures to evaluate the performance of MT tools. However linguistic quality and ease of integration with the existing tools are the indicators for evaluation. Linguistic quality means that translated output can take less time to post-edit and ease of integration supports better communication with translation management system .

IV. MACHINE TRANSLATION APPROACHES

Many machine translation approaches have already been developed for the natural languages as Sanskrit ,English Hindi, Spanish and other languages etc.

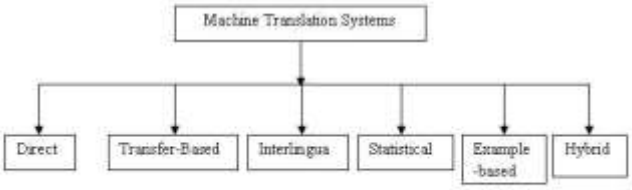

Figure 1- Different Approaches of Machine Translation System

V. RULE BASED MACHINE TRANSLATION (RBMT)

RBMT is called Knowledge Based Machine Translation that retrieves rules from bilingual dictionaries a grammars based on linguistic information about source and target languages. RBMT generates target sentences on the basis of syntactic, morphological and semantic regularities of each language. It converts source language structures to target language structures and it is extensible and maintainable as in [1]. There are three types of RBMT systems (Figure 2)

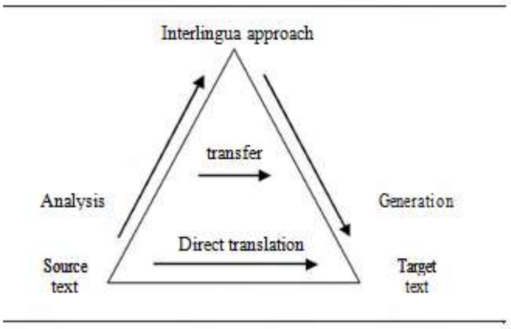

Figure 2- Different Methods of Rule Based Machine Translation

5.1 Direct method (Dictionary Based Machine Translation)

Source language text are translated without passing through an intermediary representation. The words will be translated as a dictionary does word by word, usually without much correlation of meaning between them. Dictionary lookups may be done with or without morphological analysis. Anusaarka is the example of system that uses direct approach. Indian Institute of Information Technology, Hyderabad, develops it.

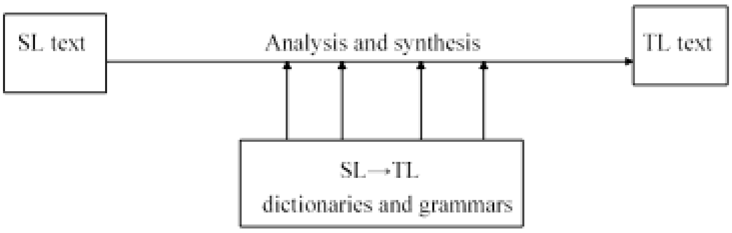

Figure 3- Dictionary Based Machine Translation System

5.2 Challenge with Dictionary Based Machine Translation

5.2.1 Language Divergence

Divergence is a common problem in translation between two natural languages. Language divergence occurs, when lexically and syntactically similar sentences of the source language are not translated into sentences that are similar in lexical and syntactic structure in the target language.

For example, consider the following English sentence (ES) and their Sanskrit translation(SS)

ES: She is in fear.

SS: Saa vibheti

. (She) (is in fear)

example has a structural variation The prepositional phrase is in fear” is translated by the verb vibhati This is an instance of a translation divergence.

5.3 Transfer Rules Based Machine Translation Systems

Morphological and syntactical analysis is the fundamental approaches in Transfer based systems. Here source language text is converted into less language specific representation and same level of abstraction is generated with the help of grammar rules and bilingual dictionaries. In the transfer approach of translation divergence, there is transfer rule for transforming a source language (SL) sentence into target language (TL), by performing lexical and structural manipulations Mantra is a transfer based tool which is a funded project of India Government.

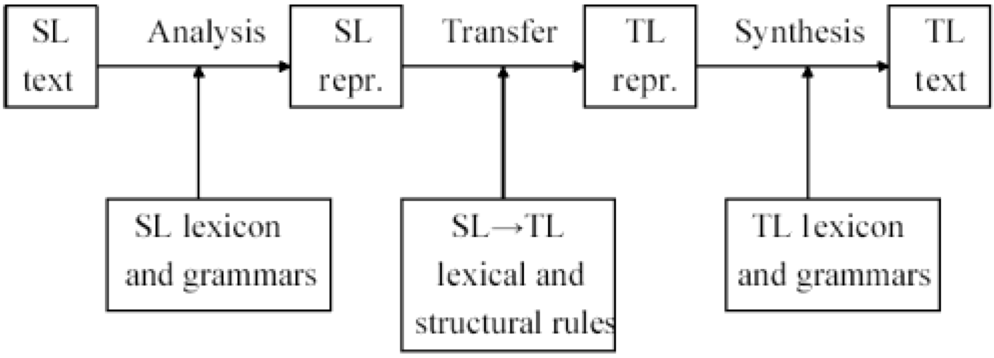

Figure 4- Different Methods of Rule Based Machine Translation

5.4 Transformation Process

5.4.1 Morphological Analysis

Surface forms of the input text are classified as

(a) Part-of-speech (e.g. noun, verb, etc.) and

(b) Sub-category (number, gender, tense, etc.)

5.4.2 Lexical Categorization

In any given text some of the words may have more than one meaning, causing ambiguity in analysis. Lexical categorization looks at the context of a word to try and determine the correct meaning in the context of the input

5.4.3 Lexical Transfer

This is basically dictionary translation the source language lemma (perhaps with sense information) is looked up in a bilingual dictionary and the translation is chosen.

5.4.4 Structural Transfer

While the previous stages deal with words, this stage deals with larger constituents

5.4.5 Morphological Generation

From the output of the structural transfer stage, the target language surface forms are generated.

5.4.6 Challenge with Transfer Rules Based Machine Translation

- Managerial phenomena: Managerial phenomena should be used only as a foundation and must be coupled with considerable work to apply a model-building approach.
- number of rules : The number of rules will grow drastically in case of general translation systems.

5.5 Interlingual RBMT Systems (Interlingua)

This model is indented to make linguistic homogeneity across the world. In this method, source language is translated into an intermediary representation which does not depends on any languages. Target language is derived from this auxiliary form of representation.

5.5.1 Challenge with Interlingual Rules Based Machine Translation

- handle exceptions:Hard to handle exceptions to rule for interlingual.
- number of rules : The number of rules will grow drastically in case of general translation systems. chanllenge is that the definition of an interlingua is difficult and maybe even impossible for a wider domain.

VI. CORPUS BASED MACHINE TRANSLATION

One of the main methods of machine translation is Corpus Based Machine Translation because high level of accuracy is achieved at the time of translation by this method. Large volumes of translations are presented after the development of corpus based system that is used in various computer-aided translation applications . Following is the different types of Corpus Based Machine Translation models.

6.1 Statistical Machine Translation (SMT)

Statistical models are applied in this method to create translated output with the assistance of bilingual corpora. The concept of Statistical Machine Translation comes from information theory. The important feature of this method is no customization work is required by linguists because the tool learns translation methods through statistical analysis of bilingual corpora.

6.2 Challenge with Statistical Based Machine Translation

- Sentence alignment: In parallel corpora single sentences in one language can be found translated into several sentences in the other and vice versa. Sentences. aligning can be performed through the Gale-Church alignment algorithm.
- □ Statistical anomalies: Real-world training sets may override translations of, say, proper nouns. An example would be that "I took the train to Berlin" get smis-translated as "I took the train to Paris" due to an abundance of "train to Paris" in the training set.
- Data dilution: Data dilution is a statistical anomaly unique to a subset of natural language and has shown a negative impact on Machine Translation adoption for commercial use.
- □ Idioms: Depending on the corpora used, idioms maynot translate "idiomatically".
- □ Different word orders: Word order in languages differs. Some classification can be done by naming the typical order of subject (S), verb (V) and object (O) in a sentence and one can talk, for instance, of SVO or VSO languages.

VII. EXAMPLE BASED MACHINE TRANSLATION

This method is also called as Memory based translation in which set of sentences from source language is given and generates corresponding translations of target language with point to point mapping. Here examples are used to convert similar types of sentences and previously translated sentence repeated, the same translation is likely to be correct again . The main advantage of this model is it work well with small set of data and possible to generate output more quickly by train the translation program. Example based method is mainly used to translate two totally different languages like Japanese and English as in. It is not possible to apply deep linguistic analysis that is one of the main drawbacks of Example based engine. PanEBMT is an example of EBMT tool.

7.1 Challenge with Example Based Machine Translation

7.1.1 Parallel Corpora

EBMT is a corpus based MT, so this requires a parallel aligned corpus. The sources of machine readable parallel corpora are own parallel corpus of researchers, public domain parallel corpora. The EBMT system is generally to be best suited to a sublanguage approach and an existing corpus of translations can serve to define implicitly the sublanguage which the system can handle.

7.1.2 Size of Example Database

There is a question: How many examples are needed in the example database to achieve the best translation result? The quality of translation is improved as more examples are added to the database. There is some limit after which further examples do not improve the quality of translation..

7.1.3 Generalized Examples

In some systems, similar examples are combined and stored as a single “generalized” example. Brown for instance, tokenizes the examples to show equivalence classes such as “person’s name”, “date”, “city name” and also linguistic information such as gender and number. In Generalized Examples approach, phrases in the examples are replaced by these tokens, thereby making the examples more general.

Computational Problem: All the approaches of EBMT systems have to be implemented as software and significant computational factors influence many of them. One problem of such approaches, which stores the examples as complex annotated structures, is the huge computational cost in terms of creation, storage and matching or retrieval algorithms

VIII. HYBRID MACHINE TRANSLATION (HMT)

HMT takes the advantages of RBMT and Statistical Machine Translation. It uses RBMT as baseline and refines the rules through statistical models. Rules are used to pre-process data in an attempt to better guide the statistical engine. Hybrid model differ in various ways.

8.1 Rules Post-Processed by Statistics

Rule based tool is used for translation at first. Statistical model is applied to adjust the translated output of rule based tool.

8.2 Statistics Guided by Rules

In this method, rules are applied to pre-process input that gives better guidance to statistical tool. Rules are also used to post-process the statistical output that caused to normalized output. This method has more flexibility, power and control at the translation time. DFKI-LT is an example of Hybrid Machine Translation Engine.

8.2.1 Challenge with HYBRID Based Machine Translation

- speech agreement mistakes.
- extra punctuation and
- wrong capitalization.

IX. DISCUSSION OF RESEARCH FINDINGS

Machine translation uses the method based on linguistic rules which convert source language to target language. Natural language understanding is the most important thing for the success of machine translation. As explained above different methods are available for automated machine translation. Type of technology chosen for machine translation is primarily depends on the source and target language pair. If customization is performed in regular basis, RBMT is better and it gives good result. But comparing with Corpus based and Hybrid method it is less efficient. Target language does not have rich morphology features it is good to use Corpus Based MT especially Statistical MT. When source and target languages are more complex, Hybrid MT is better to use because this combines the advantages of different approaches.

X. CONCLUSION

Machine Translation is an automated process within which computer software is used to convert text from one natural language to another. Translator ought to interpret the contents within the source text and build sentence structure of target language for translation. This process demands wide knowledge in grammar, structure of sentence and its meanings in the source and target languages. Machine Translation has an important role today in various applications such as customer management, documents translation, communications, software localization website translation etc. Dictionary Based, Rule Based, Corpus Based and Hybrid approaches are the main methods for machine translation. Each of these has its own advantages and limitations as explained above.

It's a proven fact that no two translation system can produce identical translations of same text in the same language pair. Also it is necessary to perform post-editing for quality translations

REFERENCES

[1] Dr. Pragya Shukla1, Akanksha Shukla" A Framework of Translator From English Speech To Sanskrit Text" IJCL, Vol. 3, Issue. 11, 2013,

[2] C. Dove, O. Loskutova, and R. Fuente, "What's Your Pick: RbMT, SMT or Hybrid?" , 2012, available at: http://amta2012.amtaweb.org/AMTA2012Files/papers/Doveetal.pdf

[3] Rick Briggs, "Knowledge Representation in Sanskrit and Artificial Intelligence", AI Magazine Volume 6 Number 1,1985

[4] ANGLABHARTI: a multilingual machine aided translation project on translation from English to Indian languages.

[5] http://www.academia.edu/1102901/Importance_of_Sanskrit_Langua ge.

[6] en.wikipedia.org/wiki/Machine_translation.