A Study of Machine Translation Methods

Bijimol T.K
Asst. Professor, Dept. of Computer Science
Santhigiri College of Computer Sciences, Vazhithala
Idukki, Kerala.
annabi2003@gmail com

Dr. John T. Abraham
Asst. Professor in Computer Science
Bharata Mata College, Thrikkakara
Kochi, Kerala
johntabraham@yahoo.com

***Abstract*—Effort to access other language document leads to the development of machine translation system which involves lots of heterogeneous features and its implementations. Information professionals are widely used the advantages of machine translation for satisfying their user's needs. 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 [13]. Machine translation is widely employed in numerous applications and a few translation agencies including government agencies are supporting implementation of tools [6]. 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 as in [2]. However linguistic quality and ease of integration with the existing tools are the indicators for evaluation as in [3]. Linguistic quality means that translated output can take less time to post-edit and ease of integration supports better communication with translation management system through API. In order to calculate the linguistic quality it is possible to find the time spends and track the edit distance and review distance. The amount of editing required to make the machine translation output to publishing standard is referred to as edit distance. Review distance is the rework of translation from vendor by in-house linguists. [4] WER is used for calculating the performance of the system because it is more intuitive than BLEU as in [3].

IV. RULE BASED MACHINE TRANSLATION (RBMT)

RBMT is also called Knowledge Based Machine Translation that retrieves rules from bilingual dictionaries and

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 1)

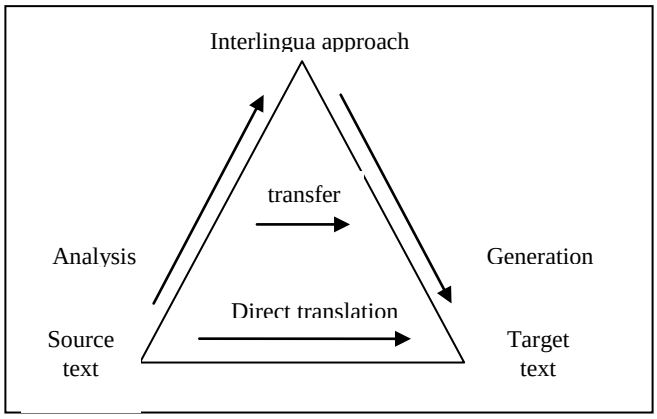

Figure 1- Different Methods of Rule Based Machine Translation

A. *Direct method (Dictionary Based Machine Translation)*

Source language text are translated without passing through an intermediary representation. Anusaarka is the example of system that uses direct approach. Indian Institute of Information Technology, Hyderabad, develops it.

B. *Transfer RBMT 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. Mantra is a transfer based tool which is a funded project of India Government.

C. *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 [13]. The main property of this model is single representation for different languages and much easier to multilingual machine translation. UNITRAN (UNIversal TRANslator) system is an example of Interlingua model.

The edit distance and review distance of Rule Based Machine Translation system are given here.

Table 1- Edit/Review distance for RBMT languages

Language	Edit distance	Review distance	Volume (words)
French	46.33%	9.1%	38900
Italian	49.05%	16.94%	40149
Spanish	33.67%	6.30%	56269
Simplified Chinese	54.43%	2.69%	80367

The benefits of RBMT are easy customization and predictability. Easy customization means user dictionaries is adjusted to fix errors and predictability is the quality to understand the output you can expect with basic understanding of the tool.

RBMT has some disadvantages as in [9] and first one is unavailability of good dictionaries. New dictionary building is truly high-priced task. Another limitation is, it's necessary to set some linguistic information manually. In addition, it's very difficult to manage rule interactions and ambiguity in the large system. RBMT allows building new rules and extends it but these changes are very expensive.

V. 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 as in [8]. Corpus based model learn from bilingual corpora that were POS-tagged and parsed, word-aligned and Phrase-aligned and builds translation rules between source and target language [14]. Following is the different types of Corpus Based Machine Translation models.

A. *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 [11]. Also, this tool is cheaper than that of Rule based tools. Better use of resources is another advantage of SMT model that leads to more natural transactions. n-gram based SMT is one of the examples of SMT system. Statistical word-based translation model, Statistical phrase-based model and Statistical syntax-based model are different kinds of statistical models [1].

The limitations of SMT are:

- Corpus creation is costly with limited resources
- Unexpected result
- Do not work well between languages that have considerably different word orders.

B. 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 [7]. 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[1]. 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 [10].

An important model that uses the corpus-based machine translation is Statistical Machine translation and following table shows the edit and review distance of this method.

Table 2- Edit/Review distance for SMT languages

Language	Edit distance	Review distance	Volume (words)
Danish	36.18%	0.32%	89747
Norwegian	37.19%	N/A	105674
Swedish	43.56%	2.41%	115544

V. 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 [13]:

(a) 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.

(b) 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 [11]. Edit distance and review distance of HMT is given in the following table

Table 3- Edit/Review distance for HMT languages

Language	Edit distance	Review distance	Volume (words)
German	77.88%	13.77%	17269
Russian	69.11%	5.85%	33764

The edit distance of data in the above table are high it is because of rich morphology of these two languages. Hybrid for French and Italian has taken place recently, but no figures are available now [3]. Hybrid Machine translation method shows better results than other methods for both spoken and written input [12]. Hybrid approach is used in Telugu to English MT [13] that points out some advantages which are (a) fast application development (b) customizable machine translation (c) high accuracy and simple and easily understandable design. In addition, Direct MT require only less human post-editing so translators and buyers demanding this. It gives cheap, fast and quality translation output. It has some limitations also that are part-of –speech agreement mistakes, extra punctuation and wrong capitalization as in [3].

VI. 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.

VII. 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] S. Tripathi, J. K. Sarjgek, “Approaches to machine translation”, Annals of Library and Information Studies, Vol. 57, Dec 2010, pp. 388-393

[2] L.R. Nair, D.S. Peter, P.R. Renjith, “Design and Development of a Malayalam to English Translator-A Transfer Based Approach”, IJCL, Vol. 3, Issue. 1, 2012, pp.

[3] 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

[4] F.J. Och, “Minimum Error Rate Training in Statistical Machine Translation”, Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics, July 2003, pp. 160-167

[5] M. Nagao, "A Framework Of A Mechanical Translation Between Japanese And English By Analogy Principle". In A. Elithorn and R. Banerji. Artificial and Human Intelligence. Elsevier Science Publishers, 1984

[6] M. N. Al-Kabi, T. M. Hailat, E.M Al-Shawakfa, I. M Alsmadi, “Evaluating English to Arabic Machine Translation Using BLEU”, IJACSA, Vol. 4,2013, pp. 66-73

[7] E. Sumita, H. Iida, “Experiments and Prospects of Example-based Machine Translation”, available at: http://acl.ldc.upenn.edu/P/P91/P91-1024.pdf

[8] M. Guidère, “Toward Corpus-Based Machine Translation for Standard Arabic”, Translation Journal, vol. 6, no. 1, January 2002,

[9] A.-L Lagarda, V. Alabau, Casacuberta, R. Silva, E. Díaz-de-Liaño, "Statistical Post-Editing of a Rule-Based Machine TranslationSystem". Proceedings of NAACL HLT 2009, pages 217–220,

[10]R.D Brown, “Example Based Machine Translation in Pangloss System”, available at: http://www.scism.lsbu.ac.uk/inmandw/ir/example-based-machine-translation.pdf

[11]P. Kohen, “Statistical Machine Translation”, Cambridge University Press, New York, 2010, pp-53

[12] S. Nirenburg, H. Somers, Y. Wilks, “Readings in Machine Translation”, Asco Typesetters, Hong Kong, pp.157, 233

[13] T. V. Prasad, G.M. Muthukumaran, “Telugu to English Translation using Direct Machine Translation Approach, International Journal of Science and Engineering Investigations, Vol. 2, Issue. 2, 2013, pp. 25-35

[14]D. Dinh, N. L. Ngan, D. X. Quang, V. . Nam, “A Hybrid Approach to Word Order Transfer in the English-to-Vietnamese Machine Translation”, available at: http://www.amtaweb.org/summit/MTSummit/FinalPapers/58-Dien-final.pdf