Annals of Library and Information Studies
Vol. 57, December 2010, pp. 388-393

Approaches to machine translation

Sneha Tripathi[1] and Juran Krishna Sarkhel[2]

[1]Assistant Librarian, Central Library, Banaras Hindu University, Email: sneha.tripathi@gmail.com

[2]Professor, Department of Library and Information Science, University of Kalyani, Email: jksarkhel@hotmail.com

Accessibility to other language web documents has always been a concern for information professionals. Many translation tools such as Babelfish and Google Translator are extensively used by librarians and information professionals to meet the varying demand of their users. These tools though do not produce exact translated verse but provides gist of the information that could be used by the librarians to understand the type of information contained in the document. Thus, it is important for librarians and information professionals to know about the various possibilities available for translation as well as the technology being used for it. The papers follow ups the different approaches used for mechanization of translation process along with a discussion over their features and limitations. The study concludes with the remark that dependency of library and information professionals over available translation tools should be up to certain extent, i.e., only for the initial sorting of the document. The translation tools available cannot be used as regular content analysis tool.

Introduction

Language is an effective medium of communication. It explicitly represents the ideas and expressions of human mind. More than 5000 languages exist in the world which reflects the linguistic diversity. It is difficult for an individual to know and understand all the languages of the world. Hence, the methodology of translation was adopted to communicate the messages from one language to another. Developments in Information Communication and Technology (ICT) have brought revolution in the process of machine translation. Research efforts have been on to explore the possibility of automatic translation of one language (source text) to another language (target text). Several tools, free as well as proprietary, are now available which support translation of text into one or more languages. Over internet, online translation is offered by Yahoo and Altavista through Babelfish. *Bing Translator* of Microsoft and *Google Translate* from Google are tools widely used for the translation by librarians and other members of web community. Firefox uses Greasemonkey application to translate the text in other languages. Google Chrome Beta offers translation if the accessed web page is in a language other than default language (mostly English). There have been major initiatives from various research organizations and government agencies to develop tools for automatic translation of text. This is to achieve wider outreach and bridge the gap of language diversity. As a major amount of web literature is available in languages other than English, these translation tools would be effective for librarians and information professionals to improve upon their information services. The present study deals with various approaches that have been adopted to achieve the automated translation of the text.

Machine translation

Machine translation is one of the research areas under "computational linguistics"[1]. Various methodologies have been devised to automate the translation process. However, the objective has been "to restore the meaning of original text in the translated verse". In general, the process of translation has two levels:

Metaphrase

Metaphrase means "word-to-word" translation. It relates to "formal equivalence", i.e., the translated version will have "literal" translation for each word in the text. However, the translated text may not necessarily convey the meaning of the original text. That means sometimes the semantics may differ from the original text.

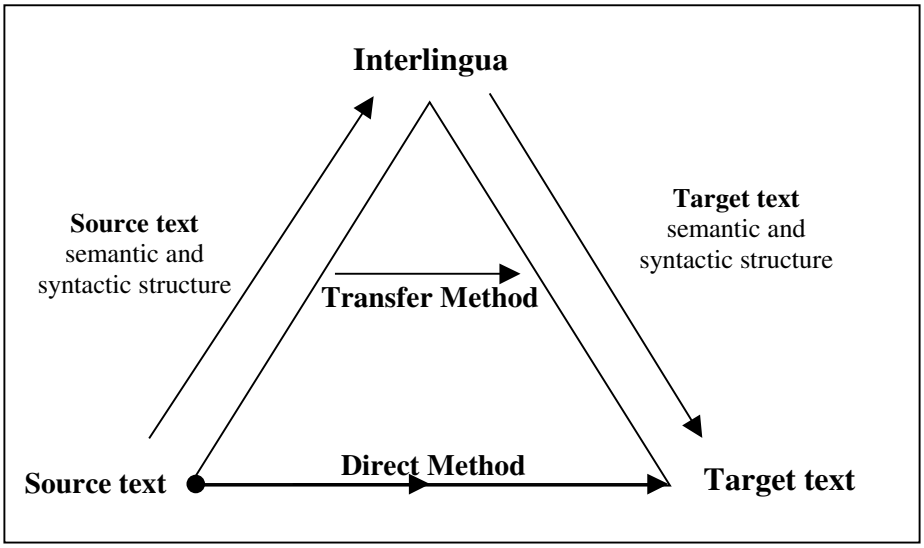

Fig. 1—Different Methods of Rule based Machine Translation
(Source:http://www.axistranslations.com/translation-article/machine-translation-definition.html)

Paraphrase

It relates to "dynamic equivalence", i.e., the translated text would contain the gist of the original text but may not necessarily contain the word-to-word translation.

Different methods of machine translation are explained in the following sections.

Dictionary based machine translation

This method of translation is based on entries of a language dictionary. The word's equivalent is used to develop the translated verse. The first generation of machine translation (late 1940s to mid 1960s) was entirely based on machine-readable or electronic dictionaries. To some extent this method is still helpful in translation of phrases but not sentences. Most of the translation approaches developed later-on more or less utilizes bilingual dictionaries with grammatical rules[2].

Rule based machine translation

Rule Based Machine Translation (RBMT) has much to do with the morphological, syntactic and semantic information about the source and target language. Linguistic rules are built over this information. Also millions of bilingual dictionaries for the language pair are used. RBMT is able to deal with the needs of wide variety of linguistic phenomena and is extensible and maintainable[3]. However, exceptions in grammar add difficulty to the system. Also, the research process requires high investment. For Indian Languages, *Anglabharati* (and *Anubharati*) is a rule based machine translation system from English to Hindi and other Indian Languages.

The objective of RBMT is to convert source language structures to target language structures. The methodology could have several approaches (Figure 1).

Direct approach

Words of Source Language are translated without passing through an additional/intermediary representation. *Anusaarka* is a machine translation system based on direct approach. It has been developed at Indian Institute of Information Technology, Hyderabad and covers all major Indian languages.

Transfer based

Transfer model belongs to the second generation of machine translation (mid 60s to 1980s). In this, source language is transformed into an abstract, less language-specific representation. An equivalent representation (with same level of abstraction) is then generated for the target language using bilingual dictionaries and grammar rules. These systems have three major components:

Analysis

Analysis of the source text is done based on linguistic information such as morphology, part-of-speech, syntax, semantics, etc. Heuristics as well as algorithms are applied to parse the source language and derive

- the syntactic structure (for language pair of the same family, for example Tamil and Telugu are siblings of same family i.e. Dravidian Languages etc.) of the text to be translated; *Or*
- the semantic structure (for language pair of different families, Hindi from Devnagari Family and Telugu from Dravidian Family)

Transfer

The syntactic/semantic structure of source language is then transferred into the syntactic/semantic structure of the target language.

Synthesis

This module replaces the constituents in the source language to the target language equivalents.
This approach, however, has dependency on the language pair involved. Thus, two independent monolingual dictionaries were suggested in Eurotra project[4]. Also, there are different representations for different languages. PaTrans (Translation for Patents) is based on transfer based approach and is one of the outcomes of Eurotra Research. *Mantra* is also a translation model for Indian Languages based on transfer approach. It is Government of India funded project and the parser used for language processing is known as *Vyakarta*.

Interlingua

This is considered to belong to third generation of machine translation. It is an inherent part of a branch called Interlinguistics. Interlingua aims to create linguistic homogeneity across the globe. Interlingua is a combination of two Latin words *Inter* and *Lingua* which means between/intermediary and language respectively. In Interlingua, source language is transformed into an auxiliary/intermediary language (representation) which is independent of any of the languages involved in the translation. The translated verse for the target language is then derived through this auxiliary representation. Hence, only two modules i.e., analysis and synthesis are required in this type of system. Also, because of its independency on the language pair for translation, this system has much relevance in multilingual machine translation. This emphasizes on single representation for different languages. The parameterization model proposed by Ali[5] is one of the enhancements over inter-lingua model with only one analysis component (multi-lingual parser) and one synthesis which work multi-lineally. The UNITRAN[6] system is one implementation of this model. It uses parameterization in both the syntactic and lexical distinctions. Indian Institute of Technology, Powai (http://www.cfilt.iitb.ac.in/machine-translation/) is working on developing translation systems for Indian languages based on Interlingua.

Knowledge based machine translation

This kind of system is concerted around "Concept" lexicon representing a domain. KANT[7] is an example of Knowledge Based Machine Translation (KBMT) system for multilingual translation, developed on a large scale knowledge base and controlled language system.

Corpus based machine translation

Since 1989, corpus based approach for machine translation has emerged as one of the widely explored area in machine translation. Because of high level of accuracy achieved during the translation, this method has dominated over other approaches. Some of the corpus based approaches are explained below:

Statistical Machine Translation (SMT)

Warren Weaver, in 1949, had introduced the idea of Statistical Machine Translation (SMT). In this, statistical methods are applied to generate translated version using bilingual corpora. Example: n-gram based SMT[8]; Occurrence based SMT[9], etc. Macherey[10] has experimented statistical methods for spoken language understanding for SMT.

Statistical word-based translation model

- Fundamental unit – Word
- Reordering; Algorithms related to alignment of words are required to achieve utmost accuracy in sentence translation
- Compound words, idioms, homonyms create complexity for simple word based translation

***Statistical phrase-based model*[11,12]**

- Fundamental unit – a phrase or sequence of words
- A sequence of words in the source and the target language is developed. Decoding is done based on the vector of features with matching values for the language sequence pair.

Statistical syntax-based model

- Fundamental unit is the translation rule.
- Translation rule consists of sequence of words and variables in the source language, a syntax tree in the target language (having words or variables at leaves), and a vector of feature values which describes the language pair's likelihood[13,14].

Liu and Gildea[15] in one of their studies have explored the semantic roles to improve syntax based machine translation.

Example-based machine translation

Example-based translation (also known as Memory based translation) is based on recalling/finding analogous examples (of the language pairs). This concept of "Translation by Analogy" was first proposed by Makoto Nagao in 1981[16]. An Example-Based Machine Translation (EBMT) system is given a set of sentences in the source language (from which one is translating) and corresponding translations of each sentence in the target language with point to point mapping. These examples are used to translate similar type of sentences of source-language to the target language. The basic premise is that, if a previously translated sentence occurs again, the same translation is likely to be correct again[17].

Advantages of an EBMT system over SMT system as put forth by Frederking[18]:

- This can work with small set of data (even with one sentence pair)
- Trains translation program and decodes more quickly
- Less principled (at least in theory)

However some studies corroborate SMT as one of the paradigms of EBMT[19].

Context based machine translation

CBMT is being developed as a corpus-based method that requires neither rules nor parallel corpora. Instead, CBMT requires an extensive monolingual target text corpus, a full-form bilingual dictionary, and optionally (to further improve translation quality) a smaller monolingual source-text corpus to run its algorithm[20].

Strengths of a CBMT system

- Accurate corpus-based MT that learns from monolingual text means that CBMT is extensible to virtually any language pair
- Preserves context in translated verse
- Able to handle longer strings compared to other approaches
- Can handle word ambiguities
- Phrasal synonym and association has the capability to generate alternative phrases in case a suitable match is not found in the target language.
- CBMT can segregate translated segments as high or low on the basis of level of confidence. This saves time and expense if post editing is needed as one needs to concentrate only on segments with low confidence.

CONTRAST[21] and REFTEX[22] are examples of Context Based Machine Translation System.

Conclusion

Many of the translation software available are able to execute literal translation of the text. But none of the solutions is perfect to create dynamic equivalence between the translated and original text. Every course

of machine translation has its own advantages and loopholes.

The different approaches for machine translation explained above suggest that the translation at metaphrase level is attainable through the currently available translation software but achieving paraphrase level or dynamic equivalence between the source and target language still appears to be a far fetched dream for computer linguists.

Language is evolutionary in nature; hence, it is difficult to say that one approach would be sufficient to handle the translation process. Knight[23] stated that machine translation as one of the elusive goals for computer science research. A decade has passed since the statement was made but it still appears to be relevant looking at the level of accuracy being achieved in machine translation. Linguistic irregularities, ambiguities, lack in universality of grammar and lexicon, are some of the reasons behind the failure of systems to achieve 100 percent accuracy in the machine translation.

Some of recommendations as made by Ali[24] to achieve the utmost accuracy in machine translation are developing universal meta-language; standardization of lexical organization and content and development of translation oriented multilingual textual corpus. However, the suggested measures appear to be unachievable looking at vast number of variants of existing languages.

The process of translation involves a thorough cognitive analysis of the source text which require various course of action such as study of grammar; semantic and syntax analysis, etc. Also, a thorough knowledge as well as understanding of the target language is also must for a translator (whether human or machine supported).

Thus, librarians and other information professionals may depend upon these tools only to certain extent i.e. just to gather the gist about the document. Then they can take help of language experts about the content of document before classifying or making it available for users. Thus, these tools may be helpful for initial sorting of the documents not as regular tools for analysis of the content of the other language document.

References

1. Uszkoreit H, What is computational linguistics? (2000) Available at: http://www.coli.uni-saarland.de/~hansu /what_is_cl.html (Accessed on 25th April 2010)
2. Yang V S-C, Electronic dictionaries in machine translation. *Encyclopaedia of Library and Information Science*, 48 (1991) 74-92.
3. Kaji H, An efficient execution method for rule-based machine translation (1988). Available at: http://www.aclweb.org/anthology/C/C88/C88-2167.pdf (Accessed on 25th April 2010)
4. Krauwer S, The Eurotra Project (2008). Available at: http://www-sk.let.uu.nl/stt/eurotra.html (Accessed on 25th April 2010)
5. Ali N, Machine translation: a contrastive linguistic perspective. Available at: http://www.unesco.org/comnat/france/ali.htm (Accessed on 25th April 2010)
6. Dorr B J, A parameterized approach to integrating aspect with lexical-semantics for machine translation" *Proceedings of the 30th annual meeting on Association for Computational Linguistics*, (1992) pp. 257-264. Available at: http://www.aaai.org/Papers/AAAI/1987/AAAI87-095.pdf (Accessed on 25th April 2010)
7. Mitamura T, Nyberg E H and Carbonell J G, Automated corpus analysis and the acquisition of large, multi-lingual knowledge bases for MT" (1993) Available at: http://www.lti.cs.cmu.edu/Research/Kant/PDF/tmi-93-final.pdf (Accessed on 25th April 2010)
8. Costa-jussă M R, Crego J M, Lambert P, Khalilov M, Fonollosa J A R, Mariño J B and Banchs R E, Ngram-based statistical machine translation enhanced with multiple weighted reordering hypotheses Proceedings of the Second Workshop on Statistical Machine Translation, Prague, June 2007, pp. 167–170.
9. Ambati V and Lavie A, Occurrence based statistics in machine translation. Available at: http://www.cs.cmu.edu/~vamshi/publications/obs.pdf (Accessed on 25th April 2010)
10. Macherey K, Bender O and Ney H, Applications of statistical machine translation approaches to spoken language understanding, *IEEE Transactions On Audio Speech And Language Processing*, 17(4) (2009) pp. 803-818
11. Koehn P, Och F J and Marcu D, Statistical phrase based machine translation (2003). Available at: http://www.aclweb.org/anthology/N/N03/N03-1017.pdf (Accessed on 25th April 2010)
12. Zens R, Och F J and Ney H, Phrase based machine translation. *Lecture Notes in Computer Science*; Springer (2002) pp. 35-56.
13. DeNeefe S, Knight K, Wang W and Marcu, D, What can syntax-based MT learn from phrase-based MT? Available at: http://www.isi.edu/natural-language/mt/ats-vs-ghkm.pdf (Accessed on 25th April 2010)
14. Yamada K and Knight K, A syntax-based statistical translation model. Available at: http://www.aclweb.org/anthology/P/P01/P01-1067.pdf (Accessed on 25th April 2010)
15. Liu D and Gildea D, Can semantic roles improve syntax-based machine translation? (2008). Available at:

http://www.cs.rochester.edu/~dliu/pub/liu-gildea-nowhere08.pdf (Accessed on 25th April 2010)

16. Hutchins J, Example based machine translation – a review and commentary. *In* Recent advances in example-based machine translation. (Ed. Michael Carl and Andy Way) (2003). Available at: http://www.hutchinsweb.me.uk/MTJ-2005.pdf (Accessed on 25th April 2010)
17. Carbonell J G and Brown R D, Example based machine translation (2004) Available at: http://www.cs.cmu.edu/~ralf/ebmt/ebmt.html (Accessed on 25th April 2010)
18. Frederking Bob, Example-based MT (EBMT) (2007). Available at: http://www.cs.cmu.edu/afs/cs/user/alavie/11-731/731-cmt/www/ebmt2007.pdf (Accessed on 25th April 2010)
19. Somers H L, Example-based machine translation, *Handbook of natural language processing* (Ed. R. Dale, H. Moisl and H. Somers), (2000) pp. 611-627. Available at: http://personal.cityu.edu.hk/~ctckit/papers/EBMT-review-CUHK.pdf (Accessed on 25th April 2010)
20. Carbonell J G, Klein S, Miller D, Steinbaum M, Grassiany T and Frey J, Context-based machine translation Proceedings of the 7th Conference of the Association for Machine Translation in the Americas (2006) pages 19-28.
21. Isahara H and Uchida Y, Analysis, generation and semantic representation in CONTRAST - a context-based machine translation system *Systems and Computers in Japan*, 26 (14) (2007) pp. 37-53.
22. Kjærsgaard P S, REFTEX – A context-based translation aid *Proceedings of the 3rd Conference of the European Chapter of the Association for Computational Linguistics,* (1987) pp. 109-112. Available at: http://acl.ldc.upenn.edu/E/E87/E87-1020.pdf (Accessed on 25th April 2010)
23. Knight K, Automating Knowledge Acquisition for Machine Translation *Artificial Intelligence Magazine*, Vol. 18 (4) (1997). Available at: http://www.aaai.org/ojs/index.php/aimagazine/article/view/1323/1224 (Accessed on 25th April 2010)
24. Ali N, *op. cit*.