A Computational Approach for Translation of Texts

Ravindra Kumar R., Sulochana K.G., Jayan V., Sunil R.

Abstract:

Translation is the communication of meaning of a source-language text by means of an equivalent target-language text. Translators always risk inappropriate introduction of source-language idiom and usage into the target-language translation. On the other hand, such introductions have imported useful source-language calques and loanwords that have enriched the target languages.

Machine translation is the process of transferring the sense of a sentence from one language to another language with computational aids. Consider the sentence "Child picked up the lesson fast". We need a sense disambiguation for the verb phrase "picked up" as it is having multiple meanings depending on the context. Similarly for noun phrases like "Chinese fishing net" have a single meaning in Malayalam "cInavala". We need an extensive dictionary including the noun phrases in a specific domain with all syntactic and semantic information to get a better translation and to reduce ambiguity. The terminologies used in some domain in a language are not readable with proper pronunciation in another language. Thus a terminology dictionary has also a prominent role in the translation process. Some simple words in one (source) language may not have a single word equivalent in the target language. An example is the word "switch" in English. Its Malayalam translation is വൈദ്യുതി ഗമനാഗമനനിയന്ത്രണയന്ത്രം

vaiduti gamana:gmana niyantraNayantram. Back translation of this particular term will be "the electricity flow control device". End user of the translated text may not understand what is actually meant by that particular term. Another concern is the transliteration of the proper nouns. We know that the style of usage of language is different in different domains.

A machine assisted translation system may not be able to generate a perfect translation. But can generate translated output with considerable efficiency and a provision for post editing. No expertise is needed for the post editing purpose. This will reduce the human effort in the field of translation. Shortage of experienced and expert human translators can be overcome by the service of a Machine Assisted Translation system.

1. Introduction

The translation process may be stated as decoding the meaning of the source text; and re- encoding this meaning in the target language. Behind this ostensibly simple procedure lies a complex cognitive operation. To decode the meaning of the source text in its entirety, the translator must interpret and analyze all the features of the text, a process that requires in- depth knowledge of the grammar, semantics, syntax, idioms, etc., of the source language, as well as the culture of its speakers. The translator needs the same in-depth knowledge to re- encode the meaning in the target language. [1] An effective translation requires the translation of the sense in one culture to another culture. In Kerala we know that the meaning of commonly used words will differ. For example, five different words are used in Malayalam to represent Tapioca. In extreme north it is 'koLLi', then in the south Malabar it is 'pu:Lakkizhangu' and coming to the south it is 'kappa'. Sometimes the short form 'ci:ni' of 'maraccIni' is also used in some parts of Kerala. But generally we either use 'kappa' or 'maracci:ni'. It is a tough task to choose the best meaning by considering all the regional variation of translation of a word. This paper gives a brief description about human and machine translation, problems of translation and the dictionary structure.

2. Human Translation

Many knowledge resources are not explored by users just because of the fact that they are not proficient in the language in which it is written. If such resources have to be made available to all,

translation of these resources on the local languages is a must. Many world classics were translated in to Malayalam and they are having wide popularity in Kerala. A translator should have the following qualities:

- a *very good* knowledge of the language, written and spoken, *from which* he is translating (the source language);
- an *excellent* command of the language *into which* he is translating (the target language);
- familiarity with the subject matter of the text being trans lated;
- a profound understanding of the etymological and idio matic correlates between the two languages; and
- a finely tuned sense of when to *metaphrase* ("translate literally") and when to *paraphrase*, so as to assure true rather than spurious *equivalents* between the source- and target-language texts.[2]

The translation of a huge volume of knowledge resource or any other text material may take years to complete. Timely completion of the translation process requires some aids that support the translator to speed up the work. A bilingual dictionary is the main aid, but what should be the content of a dictionary? How it can be searched? These are the questions that we ask when it come to a Bilingual dictionary. The dictionary should be designed in such a manner that it should contain all the syntactic and semantic information. If possible that should depict the meaning using some pictures. So that one translator can visualize the object and get its equivalent word in target language. The word like "backwater" cannot be seen in an Oxford dictionary. But this is the most commonly used term in the Kerala tourism domain. Each word should be easily tracable.

If we are using digital dictionary lot of search algorithms and data structures will take care of this aspect. We have a shortage of a good SL-Malayalam bilingual dictionaries, even the most commonly used E-IL bilingual dictionaries. The online dictionaries and other digital dictionaries available are not sufficient enough to fulfill the requirement of a translator.

3. Machine Translation

Machine translation (MT) is a process by a computer program which analyzes a source text and produces a target text without human intervention. In reality, a machine translation typically requires a human intervention in the form of pre-editing and post-editing. The Machine Translation starts with the text analysis work like removal of unwanted characters, sentence boundary detection, identification of proper nouns, abbreviations and acronyms, etc. of the source text for machine translation (pre-editing). After Machine translation, reworking of the machine translated output is performed by a human translator (post-editing). Commercial machine translation tools can produce useful results, especially if the machine translation system is integrated with a translation memory.

Unedited machine translation is available to a large public through tools on the Internet such as Babel Fish, Babylon, and StarDict. Under favorable conditions these will give an abstract idea of the source text in target language. There are also companies like Ectaco which produce pocket translation devices that utilize MT. [3]

We cannot rely on unedited machine translation because the machine may not have transferred the actual sense of the source text. The sense of the text will be context dependent. Therefore human intervention is needed in the MT output to ensure the quality of the translation.

4. Problems of Knowledge Text Translation

Claude Piron writes that machine translation, at its best, automates the easier part of a translator's job; the harder and more time-consuming part usually involves doing extensive research to resolve ambiguities in the source text, which the grammatical and lexical exigencies of the target language require to be resolved. Such research is a necessary prelude to the pre-editing in order to provide input for machine-translation software, such that the output will not be meaningless. [1]

4.1 Multiword Expressions

In general English language multi-words, such as "ice cream", "data base" or "follow-up", composed of several words referring together to one concept or item, are quite rare, but in medical domain they may be common. The number of words in a string may vary; words may be separated by a space or a hyphen. Independently they may have one meaning but if they are combined they will have another meaning. Consider the phrase "flat bottomed barge"; the word by word translation of this phrase in a sentence will not give a good translation. While taking together as a phrase and giving the meaning *patte:ma:ri* the MT system will give perfect translation. The phrase like 'coconut tree' has single word meaning in Malayalam, i.e. *tengu*. If we translate word by word the output will be *te:ngayuTe maram*. But when we go through the dictionary we will not find the phrase 'coconut tree'. If we are using an electronic dictionary, the independent word meanings will be taken for translation, which will not give proper translation. We can have a lot of such examples to be added in the dictionary. When we consider the organization names, university degrees, etc in a sentence they should be treated as a single entity. If each word in the sentence is translated separately the translation will become worse. Same will occur with the case of human translation. If a translator does not know the exact equivalent word in the target language he will get confused and will give bad translation. In all the domains we

have such multiwords. While making a dictionary for translation purpose, we have to identify such phrases and include in the dictionary. The difference in machine translated output when the dictionary is with and without the term "Chinese fishing net" as a multi word can be seen in the example below:

The Chinese fishing nets have become a very popular tourist attraction.
Correct Translation:

ചീനവല ഒരു വളരെ ജനപ്രിയമായ വിനോദസഞ്ചാര ആകർഷണം ആയിട്ടുണ്ട്.

ci:navala *oru valare janapriyama:ya vino:dasanca:ra a:karshaNam a:yiTTundu*

Wrong Translation:

വലകൾ മീൻ പിടിയ്ക്കുന്ന ചൈനീസ് ഒരു വളരെ ജനപ്രിയമായ വിനോദ സഞ്ചാര ആകർഷണം ആയിട്ടുണ്ട്

valakaL mi:n piTikkunna cini:s *oru vaLare janapriyama:ya vino:dasanca:ra a:karshaNam a:yiTTunTu.*

4.2 Semantic Disambiguation

In MT systems, semantic disambiguation is another factor. The word 'go' has 63 senses and 'fall' has 35[4], if we use a rule based MT system the disambiguation will become a tough task. Normally the most commonly used target language meaning is assigned for these words in the dictionary. So only post editing can overcome this problem.

There are three types of lexical ambiguity: Polysemy, Homonymy and Categorical Ambiguity. Polysemy is the most problematic because adding more number of target language meanings will make the disambiguation complex. This needs more accurate and fine segregation of semantic category. In homonymy we don't find much difficulty. The meanings will be less in number compared to Polysemy. Categorical ambiguity can be resolved by using a good POS tagger. How one can translate the sentence like

"Peter shoots the minister."? Here one should know whether Peter is a photographer or a criminal. Even a human translator also fails to disambiguate the sense of the word 'shoot'. How a MT system or a human translator deal with the negatively primed sentences such as "The astronomer married the star." Here one should take the meaning of 'star' as 'celebrity' instead of astronomical object. The phrase like "that accounts for the milk in the coconut" is having the meaning "now everything is clear". How can a machine translation system or even a human translator translate the sentence like this? When it occurs independently as a part of the sentence, the translation will be more difficult. If it occurs in a paragraph, a human translator can infer the real meaning from the overall context.

Another problem is with the sentences having the words with dominant sense and the secondary sense. In the sentence "A good sprinter uses his arms" we should use the secondary sense "weapon" for arms instead of the primary sense "human limbs". Disambiguation of such sense is also a difficult task with a MT system.

4.3 Translation of foreign words

The Malayalam vocabulary consists of a number of words borrowed from Sanskrit and Tamil. The arrival of the Europeans further enriched the Malayalam vocabulary, with the language absorbing numerous words and idioms from English, Portuguese, Dutch, etc. Infact English stands next to Sanskrit in lending words to Malayalam. Likewise, many Malayalam words found their way into other languages (e.g. Coir, Copra, Catamaran etc.). Some foreign words may have the Malayalam equivalent word meaning. But in daily life they may not be commonly used in conversation or writings. Telephone, mobile, etc. are such daily used examples that do not use Malayalam equivalent words. We cannot think of using Malayalam equivalent word *vaiduti gamana:gmana niyantraNayantram* for switch in daily usage. So we should be wise enough to accept loan words which do not affect language negatively. Now a days lot of

technical terms are used in daily life. Generating target language is difficult in this context. If we create the complex language equivalent as that of switch discussed above will be of no use.

4.4 Transliteration of Terminologies

In MT systems even the perfect translation can be made unnatural if we have a poor transliteration module. This is the area where the MT system finds a pitfall. Human translators find difficulty in this area. Mere application of rules will not be sufficient for MT systems in transliteration. Most often, names of medical equipments, medicine names, scientific names of the plants and animals comes in combination of two or three words. We cannot find an equivalent target language meaning. That has to be transliterated anyway. A person proficient in that particular domain only can transliterate the words correctly. If we use a Transliterator of a MT system the output will be unacceptable. The transliteration will be different for the words from different region of the world. Consider the name "Cronje", the actual pronunciation is 'krONye'. In normal case 'nj' combination is transliterated as 'Fc' as per the transliteration scheme developed for Malayalam. But in many foreign languages 'j' is pronounced as 'y' and in some cases 'Y'.

4.5 Acronyms and Abbreviations

Translation of acronyms and abbreviations also posses having some problems. Consider the case of UNESCO and CEDTI. Former can be read and written as 'yunesko' not 'yu en i es si O' but the later can be read and written as 'si I di ti Q'. Some acronyms are read continuously as they are written. Some should be read letter by letter. So we need a list of commonly used Acronyms in the dictionary. When we translate the Abbreviations like Mr. and Ms. ('SrI' and 'SrImawi' in Malayalam) we should know the equivalent target language word for the Abbreviations.

4.6 Idioms and Proverbs

The words develop a specialized meaning as an entity, as an *idiom*. Moreover, an idiom is an expression, word, or phrase whose sense means something different from what the words literally imply. When a speaker uses an idiom, the listener might mistake its actual meaning, if he or she has not heard this figure of speech before. [9] Idioms usually do not translate well; in some cases, when an idiom is translated into another language, either its meaning is changed or it is meaningless. More than 25000 idioms are estimated in English language. [10] Take an example of a sentence with an idiom 'kicked the bucket'.

The old man finally kicked the bucket.

മുതിർന്ന മനുഷ്യൻ അവസാനം തൊട്ടി തട്ടി (MT output)

mutirnna manushyan avasa:nam toTTi taTTi

വൃദ്ധൻ അവസാനം മരിച്ചു.(Actual translation)

vrudhan avasa:nam mariccu

Proverbs are often borrowed from similar languages and cultures, and sometimes come down to the present through more than one language. It is a simple and concrete saying popularly known and repeated, which expresses a truth, based on common sense or the practical experience of humanity. "No flies enter a mouth that is shut", is a proverb that can be traced back to an ancient Babylonian proverb (Pritchard 1958:146). A translator must know actually what is meant by the phrase and then that should get translated in to the target language.

4.7 Pictures

Pictorial representation is helpful for a human translator to get the target language equivalent of the scientific terms of the animals and plants by observation. *"Loxodonta africana"* is the

scientific name for the African elephant. A person with adequate knowledge in the biological term can only translate this term. So the pictorial representation is a requirement in a dictionary.

4.8 Dictionary Structure

A dictionary developed for a human translation or MT system should contain above discussed contents. This will not only improve the translation quality but also the translation time. A dictionary structure as per the discussion above will be as shown below:

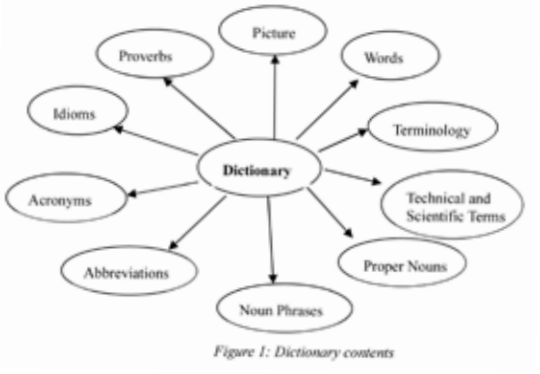

Figure 1: Dictionary contents

5. Conclusion

The field of knowledge translation has a great role in this globalized market. Now a days we can see many advertisements of the requirement of freelance translators. Many MNCs are hiring people with high payments for translation work. There is always a shortage of expert translators in the country. As we have discussed the translator should be proficient in both languages. For a good translation the target language should always be his mother tongue. In order to overcome the shortage of an expert translator we should design some translation supporting system. A human assisted MT system equipped with a good dictionary is a good alternative for

an expert translator. An online dictionary will reduce the manual searching speed compared with a paper dictionary. An effective MT system can be developed if we overcome the problems discussed above. But an MT system will give more accurate results when it is restricted to some specific domain.The issues like divergence patterns can also be handled well using a rule based MT system. More research should be carried out on handling the diverse English sentence patterns.The issues like identification of phrasal verbs, assigning proper meanings for the prepositions in English are some of the major issues concerned with MT systems.This can be overcome in post-editing module of MT system. The post editing facility incorporated in the MT system will be sufficient for further refinement of the translation. A computational approach will reduce the human effort and the effective translation time and thus the cost of translation.

REFERENCES

Graeme Hirst, "Semantic Interpretation and the resolution of ambiguity", Cambridge University Press, Page 5

http://en.wikipedia.org/wiki/Machine_translation

http://www.prokerala.com/malayalam/language.htm

http://wapedia.mobi/en/Translation?p=2

http://en.wikipedia.org/wiki/Translation#Machine_translation

Jackendoff, R. (1997). The architecture of the language faculty. Cambridge, MA: MIT Press.

John Pestian, Lukasz Itert and W lodzis law Duch, "Devel opment of a Pediatric Text-Corpus for Part-of-Speech Tagging"

Kasparek, Christopher, "The Translator's Endless Toil," *The Polish Review,* vol. XXVIII, no. 2, 1983, pp. 83-87. Includes a discussion of European-language cognates of the term, "translation."

Satoshi Kinoshita, Miwako Shimazu, Hideki Hirakawa, R & D Center, Toshiba Corporation, "Better Translation with Knowledge Extracted from Source Text"

Saeed, John I. (2003), *Semantics.* 2nd edition. Oxford: Blackwell.

(Paper presented in the seminar Growth of Malayalam Language and the Role of Knowledge Text Translation on January 29, 2011.)