A Simple Solution for Divergence in Pure EBMT

Ruchika Sinhal
Deptt of CSE
Shri Ramdeobaba College
of Engg and Mang., Nagpur

M.B.Chandak
Head, Deptt. of CSE
S.R.C.O.E.M
NAGPUR

Dr.R.V.Dharaskar
Director
MPGI, CAMPUS
NANDED

ABSTRACT

Over the past years there is continuous involvement in field of Machine Learning. There are different applications which help common man to tackle with different languages all over the world. Then also there are many challenges faced in this filed and building the applications. The paper gives the brief description about the concept in machine translation, the challenges involved and our way of solving the problem.

Keywords

Machine translation, Divergence, Process of translation.

1. INTRODUCTION

The natural language is used by every common man. The language which we all speak is nothing but the common or even termed as natural language. We need this language for communication. Natural language processing is nothing but the processing, refining, modifying and translating i.e. operating on one type of language to get another.

When we all humans are capable of speaking, then what is need for processing this natural language in our life? The need can be explained by divergence. The divergence is difference in language and also in the form of text in which the language is present. In India itself there are more than 20 languages spoken. The most ancient of all languages is Sanskrit. Many people do not understand Sanskrit but they can if the text is translated into their national or even the languages they are familiar with. Therefore for the understanding and making communication easy there is basic need of translator. This translation can be done by humans also; so why there is need to involve machine in the translation? The first reason is that text in "world of text" is huge. There are many big databases to be translated and it is not possible for a human to translate gigabytes of data in less time. To reduce the human efforts and to give the results quickly the translators are developed which can translate the text from one language to another by just one click. A second reason is that the whole technical materials are too boring to translate for human translators as they don't like to translate them continuously and consistently. Hence they look for help from computers. Thirdly, as far as large corporations are concerned, there is the major requirement that terminology is used consistently; they want terms to be translated in the same way every time. Computers are consistent, but human translators tend to seek variety; they do not like to repeat the same translation and this is no good for technical translation. A fourth reason is that the use of computer-based translation tools can increase the volume and speed of translation throughput, and companies and organizations like to have translations immediately, the next day, even the same day. The fifth reason is that top quality human translation is not always needed. Because computers do not produce good translations, some people do not think that they are any use at all. The fact is that there are many different circumstances in which top quality is not essential, and in these cases, automatic translation can and is being used widely.[1]

The need for machine translation can be briefly stated into following points briefly:

- Too much to be translated
- Boring for human translators
- Major requirement that terminology used consistently
- Increase speed and throughput
- Top quality translation not always needed
- Reduced cost

The history of machine translation is very vast, as explained by W.John Hutchins [1]. Many are under the impression that MT is something quite new. In fact, it has a long history (Hutchins, 1986, 2001) – almost since before electronic digital computers existed. In 1947 when the first non-military computers were being developed, the idea of using a computer to translate was proposed. In July 1949 Warren Weaver (a director at the Rockefeller Foundation, New York) wrote an influential paper which introduced Americans to the idea of using computers for translation. From this time on, the idea spread quickly, and in fact machine translation was to become the first non-numerical application of computers. The first conference on MT came in 1952. Just two years later, there was the first demonstration of a translation system in January 1954, and it attracted a great deal of attention of the press. Unfortunately it was the wrong kind of attention as many readers thought that machine translation was just around the corner and that not only would translators be out of a job but everybody would be able to translate everything and anything at the touch of a button. It gave quite a false impression. However, it was not too long before the first systems were in operation, even though the quality of their output was quite poor. In 1959 a system was installed by IBM at the Foreign Technology Division of the US Air Force, and in 1963 and 1964 Georgetown University, one of the largest research

projects at the time, installed systems at Euratom and at the US Atomic Energy Agency. But in 1966 there appeared a rather damning report for MT from a committee set up by most of the major sponsors of MT research in the United States. It found that the results being produced were just too poor to justify the continuation of governmental support and it recommended that the end of MT research in the USA altogether. Instead it advocated the development of computer aids for translators. Consequently, most of the US projects – the main ones in the world at that time – came to an end. The Russians, who had also started to do MT research in the mid 1950s, concluded that if the Americans were not going to do it any more than they would not either, because their computers were not as powerful as the American ones. However, MT did in fact continue, and in 1970 the Systran system was installed at the US Air Force (replacing the old IBM system), and that system for Russian to English translation continues in use to this day. The year 1976 is one of the turning points for MT. In this year, the Météo system for translating weather forecasts was installed in Canada and became the first general public use of a MT system. In the same year, the European Commission decided to purchase the Systran system and from that date its translation service has developed and installed versions for a large number of language pairs for use within the Commission. Subsequently, the Commission decided to support the development of a system designed to be 'better' than Systran, which at that time was producing poor quality output, and began support for the Eurotra project – which, however, did not produce a system in the end... During the 1970s other systems began to be installed in large corporations. Then, in 1981 came the first translation software for the newly introduced personal computers, and gradually MT came into more widespread use. In the 1980s there was a revival of research, Japanese companies began the production of commercial systems, and computerized translation aids became more familiar to professional translators. Then in 1990, relatively recently, the first translator workstations came to the market. Finally, in the last five years or so, MT has become an online service on the Internet.The term machine translation (MT) is translation of one language to another. The ideal aim of machine translation system is to produce the best possible translation without human assistance. Basically every machine translation system requires automated programs for translation and also dictionaries and grammars to support translation.

2. APPROACHES IN MACHINE TRANSLATION

Machine Translation is an attempt to automate, all or part of the process of translating one human language to another. It requires some knowledge of source and target languages and its way of interpretation to carry out the translation work. The MT systems can broadly be categorized on the basis of its knowledge type, its representation and interpretation.

We briefly discuss the categories of MT systems in the next three sections. Since our project focuses on EBMT, this model is described in more detail. Then we discuss the specific problems caused by phrasal verbs in translation.

1. Knowledge Based MT

"The term knowledge based MT has come to describe a rule – based system displaying extensive semantic and pragmatic knowledge of domain, including an ability to reason to some limited extent, about concepts in the domain."

The basic aim of KBMT is to obtain high quality output in a specific domain with no post-editing work. The KBMT systems are generally domain specific, especially a domain that is less ambiguous, like technical documents. The reason for it to be domain specific is that representing complete knowledge of the whole world is very difficult. The domain model is used to represent the meaning of the source language text. The basic components of a KBMT system are:-

1. Ontology of the domain, which serves as an intermediate representation during translation. It usually includes the set of distinct objects resulting from an analysis of a domain.
2. Source language lexicon and grammar for the analysis.
3. Target language lexicon and grammar for the generation.
4. The mapping rules between the intermediate and source/target language.

For example, the KANT system developed by CMT at Carnegie Mellon University is a practical translation system for technical documentation from English to Japanese, French and German.[3.]

We can further classify the KBMT systems based on their approach to translation as follows–
1. Direct Translation Model
2. Transfer Model
3. Interlingua

a. Direct Translation Model

Direct MT systems are built with one language pair in mind, and the only processing needed is to convert one specific source language to another specific target language. The stages in direct model may be morphological analysis, lexical transfer, preposition handling and SVO 2 rearrangement. The main characteristic of this model is that it does lexical transfer before syntactic transfer. This means that the words would be first replaced by corresponding target language words and then it is modified for grammatical correctness.

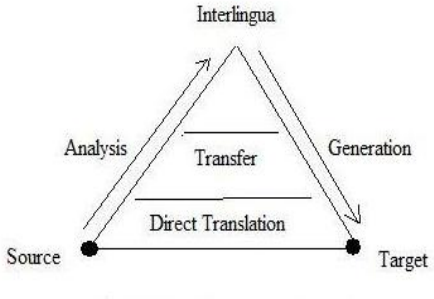

Figure 4 Machine Translation Pyramid

b. Transfer Model

Every language pair we use has some similarity and dissimilarity between them. It may be in its topology or morphology. The language pair can have lexical gap. The core idea of transfer model is to reduce this difference between them by applying the knowledge of difference, also known as *contrastive knowledge*.

This model has three basic phases: Analysis, Transfer, and Generation. In Analysis phase the source text is parsed to generate the parse tree using grammar rules. In Transfer phase, syntactic and lexical transfer reduces the syntactic and lexical differences. The Transfer phase bridges the gap between the output of the source language parser and the input to the target language generator. The Generation phase is the reverse of the parse tree generation, in other words it suitably linearizes the transformed tree. The main disadvantage of this model is that it needs to go through the entire life cycle for every language pair. Figure 5 shows the transfer model.

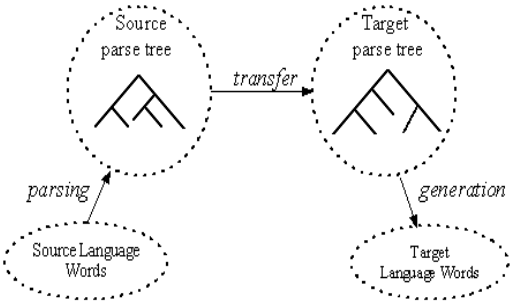

Figure 5. Transfer Model

c. Interlingua Model

Interlingua based models also take the source language text and constructs a parse tree. It moves one step further, and transforms the source language parse tree into a standard language- independent format, known as *Interlingua*. The idea of Interlingua is to represent all sentences that mean the *same* thing in the same way independent of language. It is to avoid explicit descriptions of the relationship between source and target language; rather it uses abstract elements, like AGENT, EVENT, ASPECT, TENSE, etc. The main advantage of this model is that it can be used with any language pair. The generator component for each target language takes the Interlingua as input and generates the translation in the target language. Figure 6 shows the general model of Interlingua

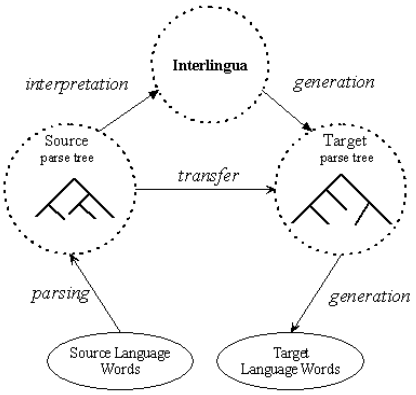

Figure 6. Interlingua Model

2. Statistical MT

The researchers in the field of speech recognition first outlined the idea of statistical approach in machine translation. It is based on statistics derived from corpora of naturally occurring language, not with pre-fabricated examples. The view of the statistical approach is that every sentence in one language is a possible translation of any sentence of other language. The statistical model tries to find the sentence *S* in the source language for which the machine translator has produced a sentence *T* in the target language. This is based on the *Bayesian* or *Noisy channel* model used in speech recognition.

The model works with the intuition that the translated sentence has passed through a noisy channel, which distorted the source sentence to the translated sentence. To recover the original source sentence we need to calculate the following –

1. The probability of getting the original sentence *S* in the source language.
2. The probability of getting the translated sentence *T* in the target language.

These are known as *Language model* and *Translation model* respectively. We assign to every pair of sentence (S, T) a joint probability, which is the product of the probability Pr(S) computed by the language model and the conditional probability Pr(T/S) computed by translation model. We choose that sentence in the source language for which the probability Pr(S/T) is maximum. Using Bayes theorem, we can write

$$Pr\ (S/T) = (\ Pr\ (S) * Pr\ (T/S)\ ) / Pr(T)$$

where S = Source Text, T = Target Text, Pr (S/T) = probability that the decoder will produce S when presented with T, Pr(S) = probability that S would be produced in the source language, Pr(T/S) = probability that the translator will produce T when presented with S, and Pr(T) = probability that T would be Target language, but, here Pr(T) does not change for each S as we are looking for most-likely S for the same translation T.

In order to get the most-likely translation, we need to maximize *Pr(S)*P(T/S).* Thus, the formula to find the most likely translation *T* for a given sentence *S* is as follows –

$$Pr\ (S/T) = agrmax\ (Pr\ (S) * Pr\ (T/S)).$$

The statistical system computes the language model probabilities (the probability of a word given all the words preceding it in a sentence), the translation probabilities (the probability of the translation being produced) and uses a search method to find the greatest value (agrmax) for the product of these two probabilities thus giving the most probable translation.

3. Example Based MT

BMT is a corpus based machine translation, which requires parallel-aligned 3 machine-readable corpora[2]. Here, the already translated example serves as knowledge to the system. This approach derives the information from the corpora for analysis, transfer and generation of translation. These systems take the source text and find the most analogous examples from the source examples in the corpora. The next step is to

retrieve corresponding translations. And the final step is to recombine the retrieved translations into the final translation.

EBMT is best suited for sub-language phenomena like – phrasal verbs; weather forecasting, technical manuals, air travel queries, appointment scheduling, etc. Since, building a generalized corpus is a difficult task. The translation work requires annotated corpus, and annotating the corpus in general is a very complicated task.

Nagao (1984) was the first to introduce the idea of translation by analogy and claimed that the linguistic data are more reliable than linguistic theories. In EBMT, instead of using explicit mapping rules for translating sentences from one language to another, the translation process is basically a procedure for matching the input sentence against the stored translated examples. Figure 7 shows the architecture of a pure EBMT.

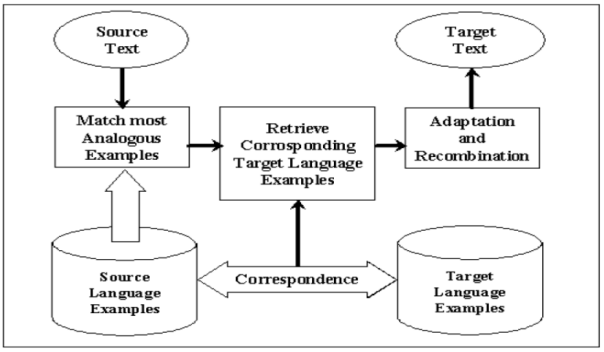

Figure 7 EBMT Architecture

The basic tasks of an EBMT system are –
- Building Parallel Corpora
- Matching and Retrieval
- Adaptation and Recombination

The knowledge base *parallel aligned corpora* consists of two sections, one for the source language examples and the other for the target language examples. Each example in the source section has one to one mapping in the target language section. The corpus may be annotated in accordance with the domain. The annotation may be semantic (like name, place and organization) or syntactic (like noun, verb, preposition) or both. For example, in the case of phrasal verb as the sub-language the annotations could be subject, object, preposition and indirect object governed by the preposition.

In the matching and retrieving phase, the input text is parsed into segments of certain granularity. Each segment of the input text is matched with the segments from the source section of the corpora at the same level of granularity. The matching process may be syntactic or semantic level or both, depending upon the domain. On syntactic level, matching can be done by the structural matching of the phrase or the sentence. In semantic matching, the semantic distance is found out between the phrases and the words. The semantic distance can be calculated by using a hierarchy of terms and concepts, as in WordNet. The corresponding translated segments of the target language are retrieved from the second section of the corpora.

In the final phase of translation, the retrieved target segments are adapted and recombined to obtain the translation. It identifies the discrepancy between the retrieved target segments with the input sentence's tense, voice, gender, etc. The divergence is removed from the retrieved segments by adapting the segments according to the input sentence's features.

Let us consider the following sentences –
- [Input sentence] John brought a watch.
- [Retrieved - English] He is buying a book.
- [Retrieved - Hindi] vaHa eka kitaba kharida raha he

The aligned chunks are –
- [He] → [vaha]
- [is buying] → [kharida raha he]
- [a] → [eka]
- [book] → [kitaba]

The adapted chunks are –
- [vaha] → [jana]
- [kharida raha he] → [kharida]
- [kitaba] → [gaghi]

The adapted segments are recombined according to sentence structure of the source and target language. For example, in the case of English to Hindi, structural transfer can be done on the basis of Subject-Verb-Object to Subject-Object-Verb rule.

3. PROPOSED WORK

The structure of the two languages if same then it is not necessary that they will always produce proper output for the given input. The two languages may differ in the structure; they may differ in the pattern generation also. Divergence is the generation of structurally different output for an entered input. The EBMT is basically adapting of the source text to the target text. The adaptation is reached if the degree of divergence is low. This divergence may create a structurally wrong output for the input thus giving wrong interpretation of the sentence. The output to be generated properly the divergences are to be studied and the way is to be proposed to resolve these divergences.

Adaptation is the major part of EBMT. The input text should translate and then adapt itself to the structure of the output text. One major difficulty in adaptation is called "*divergence*". Dorr [3] describes divergence in the following way: "*translation divergence arises when the natural translation of one language into other results in a very different form than that of the original."* In general, divergence occurs due to some inherent incompatibility between the source and target languages. Study of adaptation therefore needs a careful study of divergence too.

The existence of translation divergences makes the straightforward transfer from source structures into target structures difficult [4]. Divergence can be of two broad categories:

a) Syntactic divergence
b) Lexical – Semantic divergence

The difference between these two types of divergences is that the former category is characterized by syntactic properties associated with each languages (i.e., properties that are independent of the actual lexical items that are used) whereas the later category is characterized by properties that are entirely lexically determined. In this work we concentrated on the second type of divergence. In following part we will focus on divergence of lexical-semantic type.

Divergence is basically language to language phenomenon. Dorr proposed his work in English, Spanish and German languages. The divergence can also be studied by English and Hindi examples[5]. There are basically seven types of divergences. These are Lexical-Semantic Divergence.

1. *Thematic divergence:* the verbal object in one language becomes as the subject of the main verb in other language. For example: "*Deepa pleases Nitu*." will be translated into Hindi as " n*itu deepa ko pasand kartii hei*" ("*Nitu likes Deepa*."). The verbal object in English "*Nitu*" becomes the subject of the main verb in Hindi.

2. *Promotional divergence*: the modifier is realized as an adverbial phrase in one language but as the main verb in other language. For example: "*Fan is* on" in English, will be translated as "*pankhaa chal rahaa hai*" This means that English modifier "*on*"(an adverb) is realized as the main verb in Hindi.

3. *Structural divergence*: the verbal object is realized as a noun phrase in one language and as a prepositional phrase in other language. For example, the English sentence "*Ram attended the meeting*" will be translated as " *ram sabha mai upashtit tha*". In English " the meeting" is the noun phrase but in Hindi it becomes prepositional phrase *" shaba mein*" (*in the meeting*)

4. *Conflational divergence***:** the sense conveyed by a single word in one language requires at least two words of the other language. For example, "*He stabbed me*" will be translated as "u*sne mujhe chaaku se maaraa*". The English word "*stab*" has no one-word equivalent in Hindi, and therefore the introduction of the word "chaaku" was necessitated. Similarly for "*love*", " *swear*"etc.

5. *Categorial divergence***:** changes in category. For example, the predicate is adjectival in one language but nominal in other language. The English sentence "*I am feeling hungry*." will be translated into Hindi as " *mujhe bhukh lag rahii hai*." In English "*hungry*" is adjective and but in Hindi "*bhukh*" (*hunger*) becomes the noun.

6. *Lexical divergence*: the event is lexically realized as the main verb in one language but as a different verb in other language. Consider the sentence "*They run into the room.*" Its Hindi translation is "w*oye daurte huye kamre mein ghus gaye*" The event is lexically realized as the main verb "*run*" in English but as a different verb "*ghus"* (literally (*to enter*)) in Hindi, and "*run*" is used as participle.

7. *Demotional divergence***:** A main verb in one language is realized as an adverbial modifier in the other. As shown in [5] the example "*I like eating.*" will be translated into German "*Ich esse gern* (literal meaning *I eat likingly*)", the word "*like*" is realized as a main verb in English but as an adverbial modifier in German "*gern*". But we have not come across any divergence of this type between English and Hindi.

The translator to be made is made using the database. The database will contain parallel corpus. The corpus containing examples of both hindi and English. Translation will be worked upon different sentences which are not present in database also present in database. The translation will be then checked and aligned according to the format. Adaptation will be performed and corresponding output will be generated.

4. CONCLUSION

The paper thus discusses the concept of machine translation. The history of machine translation with its boons. Machine translation is a vast field. The different types of approaches are present in which research and work is being performed. Example Based Machine Translation is the main approach in project so it is explained briefly. The main challenge of divergence is phased in translation. A solution is proposed according to my view.

5. REFERENCE

[1] Hutchins W. John and Harold L. Somers (1992). *An Introduction to Machine Translation*. London: Academic Press.

[2] Somers Harold. – Review Article: Example-Based Machine Translation. Centre for Computational Linguistics, UMIST, 1999.

[3] Bonnie Jean Dorr, (1993). "*Machine Translation: A View from the Lexico*n", The MIT press, USA.

[4] Bonnie J. Dorr. (1994), "Machine Translation Divergences: A Formal Description and Proposed Solution", *ACL Vol. 20, No. 4, pp. 597-631.*

[5] *Deepa Gupta Niladri Chatterjee*Study of Divergence for Example Based English-Hindi Machine Translation