Language Processing for MT: Need, Problems and Approaches

Ruchika Sinhal[1], Kapil Gupta[2]

[1]Dept of CSE, DMIETR, Sawangi(M), Wardha
ruchisinhal04@gmail.com
[2]Dept of CSE, DMIETR, Sawangi(M), Wardha
kaps04gupta@gmail.com

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. The demand for language translation has greatly increased in recent times due to increasing cross-regional communication and the need for information exchange. Most material needs to be translated, including scientific and technical documentation, instruction manuals, legal documents, textbooks, publicity leaflets, newspaper reports etc. Some of this work is challenging and difficult but mostly it is tedious and repetitive and requires consistency and accuracy. It is becoming difficult for professional translators to meet the increasing demands of translation. In such a situation the machine translation can be used as a substitute. 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, Need, Problems in Translation, Types of MT Systems, Approaches in MT

INTRODUCTION

The language, which human beings speak, is termed as natural language. The natural language is used by every common man. People use natural language for communication. Natural language processing includes refining, modifying and translating i.e. operating on the natural languages.

Divergence is one of the main problems in any NLP system, which proves the need of research in NLP domain. The divergence is defined as difference between language, or the form of text in which the language is present. In India itself, there are more than 438 languages spoken [1]. 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 language or languages they are familiar with. Therefore, for understanding and making communication easy, there is a basic need of translation tools. The translation can be done by humans; so why there is a need of machine translation? The need of MT is described in following points:

a) Too much to be translated

The first reason is that the "world of text" is huge. There are many large documents to be translated and it is not possible for a human to translate gigabytes of data in a short time. To reduce the human efforts and to give the results quickly the machine translators are used which can translate the text from one language to another by just one click.

b) Boring for human translators

A second reason is that the all technical materials are too boring for human translators to translate as humans do not like to translate them continuously. Hence they look for help from computers.

c) Major requirement that terminology used consistently

As far as large corporations are concerned, there is the major requirement that terminology is used consistently, the 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 not good for technical translation.

d) Increase speed and throughput

The use of computer-based translation tools can increase the volume and speed of translation throughput, and organizations like to have translations immediately.

e) Top quality translation not always needed

The fifth reason is that, top quality human translation is not always needed. Computers do not produce good translations. The fact is that there are many different circumstances in which top quality translation is not essential, and in this case, automatic translation can be used widely.

History of Machine Translation

W. John Hutchins, 1986 explained vast history of machine translation [2-4]. Many people are under the impression that MT is something quite new. MT has a long history – almost since before electronic digital computers existed. In 1947 when the first non-military computers have been developed, the idea of using a computer to translate has been proposed. In July 1949 Warren Weaver [5] (a director at the Rockefeller Foundation, New York) proposed method, which introduced Americans to the idea of using computers for translation. From this time on, the idea spread quickly, and in fact machine translation became the first non-numerical application of computers. The first conference on MT was held in 1952 [6]. Just two years later, there has been the first demonstration of a translation system in January 1954 [7]. Unfortunately this demonstration has been the wrong kind of attention as many readers thought that machine translation has been 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. However, it has been not too long before the first systems have been in operation, even though the quality of their output has been quite poor. In 1959 a system has been installed by IBM at the Foreign Technology Division of the US Air Force [8], 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. The committee found that the results being produced have been just too poor to justify the continuation of governmental support and recommended the end of MT research in the USA altogether. The committee 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 also started to do MT research in the mid 1950's. Russians concluded that if the Americans were not going to do MT any more than they would not either, because their computers have not been as powerful as the American ones. However, MT did continue in fact, and in 1970 the Systran system has been installed at the US Air Force (replacing the old IBM system). The Systran system for Russian to English translation continues in use to this day [9]. The year 1976 is one of the turning points for MT. In this year, the Météo system for translating weather forecasts has been installed in Canada and became the first general public use of a MT system [10]. The European Commission decided to purchase the Systran system. The Systran has been producing poor quality output, therefore committee decided to support the development of system better than systran, and began the Eurotra project– which did not produce a system in the end During the 1970's other systems began to be installed in large corporations [11]. In 1981, came the first translation software for the newly introduced personal computers, and gradually MT came into more widespread use [12]. In the 1980's there had been 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, the first translator workstations came to the market [13]. In the last decade MT has become an online service on the Internet [14-15].

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, dictionaries and grammars to support translation [16].

Machine Translation systems are needed to translate literary works from any language into native languages. The literary work is fed to the MT system and translation is done. Such MT systems can break the language barriers by making available work rich sources of literature available to people across the world.

MT also overcomes the technological barriers. Most of the information available is in English which is understood by only 3% of the population [17]. This has led to digital divide in which only small section of society can understand the content presented in digital format. MT can help in this regard to overcome the digital divide.

Problems in Machine Translation

There are several structural and stylistic differences among languages, which make automatic translation a difficult task. Some of these issues are as follows:

Word Order
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 [18]. Some languages have word orders as SOV. The target language may have a different word order. In such cases, word to word translation is difficult [19]. For example, English language has SVO and Hindi language has SOV sentence structure.

Word Sense
The same word may have different senses when being translated to another language. The selection of right word specific to the context is important [19].

Pronoun Resolution

The problem of not resolving the pronominal references is important for machine translation. Unresolved references can lead to incorrect translation [19].

Idioms

An idiomatic expression may convey a different meaning, that what is evident from the words in sentence. For example, an idiom in English language 'No brick in their walls', would not convey the intend meaning when translated into Hindi language [19].

Ambiguity

In computational linguistics, Word Sense disambiguation (WSD) is an open problem of natural language processing, which governs the process of identifying which sense of a word (i.e. meaning) is used in a sentence, when the word has multiple meanings [19].

Types of MT systems

The following are four types of Machine Translation (MT) systems:

MT for Watcher (MT-W)

MT for watchers is intended for readers who wanted to gain access to some information written in foreign language who are also prepared to accept possible bad 'rough' translation rather than nothing. This has been the type of MT envisaged by the pioneers. This came in with the need to translate military technological documents [20].

MT for revisers (MT-R)

MT for revisers aims at producing raw translation automatically with a quality comparable to that of the first drafts produced by human. The translation output can be considered only as brush-up so that the professional translator can be freed from that boring and time consuming task [20].

MT for translators (MT-T)

MT for translator's aims at helping human translators do their job by providing on-line dictionaries, thesaurus and translation memory. This type of machine translation system is usually incorporated into the translation work stations and the PC based translation tools [20].

MT for Authors (MT-A)

MT for authors aims at authors wanting to have their texts translated into one or several languages and accepting to write under control of the system or to help the system disambiguate the utterance so that satisfactory translation can be obtained without any revision [20].

MT Approaches

Machine Translation is an attempt to automate, all or part of the process of translating one human language to another. The translation 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 knowledge type, representation and interpretation of translation tools.

The categories of MT systems are described in the next three sections. Since our research focuses on EBMT, this model is described in more detail [16].

Knowledge Based MT

"The term knowledge based MT describe a 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 KBMT 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. Ontology 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 [15].

Statistical MT

The researchers in the field of speech recognition first outlined the idea of statistical approach in machine translation. SMT 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 been 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

$$\mathrm{Pr}(S/T) = \big(Pr(S) * Pr(T/S)\big)/Pr(T)$$

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 –

$$Pr(S/T) = agrmax(\mathrm{Pr}(S) * Pr(T/S))$$

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.

Rule Based MT

A rule based machine translation system consists of collection of rules called grammar rules, lexicon and software programs to process the grammar rules [21]. The collection of rules in RBMT is extensible and maintainable. Rule based approach is the first strategy ever developed in the field of machine translation. Rules are written with linguistic knowledge gathered from linguists. Rules play major role in various stages of translation as, syntactic processing, semantic interpretation, and contextual processing of language.

Tree structure is used to represent the structure of the sentence. A typical English sentence consists of two major parts: noun phrase (NP) and verb phrase (VP) [22]. These two parts can be further divided as per the structure of the sentence. ‘Rewrite rules’ are used to describe what tree structures are allowable for a given sentence. Only the sentences with right structure lead to correct translation. Following is the example of rules representing a simple grammar:

S → NP VP

VP → V NP

NP → Name

NP → ART N

where S stands for sentence, V for verb, N for noun and ART for article. A grammar can derive a sentence if there is a sequence of rules to rewrite the start symbol, S, into a sentence.

Logical form is commonly used in semantic interpretation. For example the sentence, Joe has been happy, can be written in logical form as:

$$\left(\langle PAST\ HAPPY\rangle\ (NAME\ j1\ "Joe")\right)$$

where PAST stands for past tense. Semantic interpretation is a compositional process in which interpretations can be built incrementally from the interpretations of subphrases. Lexicon plays a major role in semantic interpretation. Grammar rules are used to compute the logical form of the given sentence. Consider the grammar rule given below.

$$\left(S\ SEM\ (?\,semvp\ ?\,semnp)\right) \rightarrow\ (NP\ SEM?\,semnp)\ (VP\ SEM?\,semvp)$$

where SEM stands for semantic feature. The rule above states that a sentence consists of a noun phrase and verb phrase.

Example Based MT

EBMT is a corpus based machine translation, which requires parallel-aligned three machine-readable corpora. 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) has been the first to introduce the idea of translation by analogy and claimed that the linguistic data are more reliable than linguistic theories [23]. 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 2.1 shows the architecture of a pure EBMT [24].

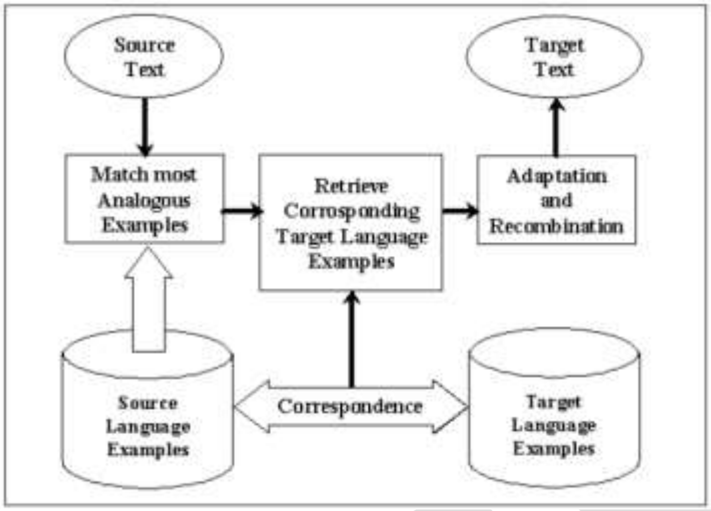
Source Text
Target Text
Match most Analogous Examples
Retrieve Corrosponding Target Language Examples
Adaptation and Recombination
Source Language Examples
Correspondence
Target Language Examples

Fig. 1 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* consist 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. The final phase identifies the discrepancy between the retrieved target segments with the input sentences' 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.

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.

REFERENCES:

[1] The Economist online, "Speaking in tongues-Language diversity around the world", 15 Feb, 2012, http://www.economist.com/blogs/graphicdetail/2012/02/daily-chart-9?fsrc=gn_ep

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

[3] D. Arnold, L. Balkan, S. Meijer, L.L. Humphreys, L. Sadler: *Machine Translation: an Introductory Guide*. Blackwells-NCC, London, Great Britain, 1994.

[4] Hutchins 95 J. Hutchins: Reflections on the history and present state of machine translation. In Proc. of *Machine Translation Summit V*, pp. 89–96, Luxembourg, July 1995.

[5] W.Weaver. Translation. In W.N. Locke, A.D. Booth, editors, *Machine Translation of Languages: fourteen essays*, pp. 15–23. MIT Press, Cambridge, MA, 1955.

[6] John Hutchins, *Milestones in machine translation No.4: The first machine translation conference*, June 1952 Language Today, no. 13, October 1998, pp.12-13 http://www.hutchinsweb.me.uk/Milestones-4.pdf

[7] The first public demonstration of machine translation: The Georgetown-IBM system, 7th January 1954, http://www.hutchinsweb.me.uk/GU-IBM-2005.pdf

[8] Masterman, Margaret and Kay, Martin, "Operational system (IBM-USAF Translator Mark I), at Foreign Technology Division, USAF, in 1959", www.hutchinsweb.me.uk/sources/Russian-IBM-1959.doc

[9] Systran, [Online]. Available: http://www.hutchinsweb.me.uk/IntroMT-10.pdf

[10] The EUROTRA project, http://www-sk.let.uu.nl/stt/eurotra.html

[11] J. Chandioux, A. Grimaila: Specialized machine translation. *In 2nd Conf. of the Association for Machine Translation in the Americas (AMTA 96)*, pp. 206–212, Montreal, Canada, Oct. 1996.

[12] "Machine Translation", http://en.wikipedia.org/wiki/Machine_translation

[13] John Hutchins, "The origins of the translator's workstation", Machine Translation, vol.13, no.4 (1998), p. 287-307

[14] Hutchins and Lovtsky, *in press.*

[15] Hutchins, J. 1986. Machine Translation: Past, Present, Future, Ellis Horwood/Wiley, Chichester/New York.

[16] Sergei Nirenburg and Yorick Wilks, Machine Translation

[17] D. D. Rao, "Machine Translation A Gentle Introduction", RESONANCE, July 1998.

[18] "Statistical machine translation", http://en.wikipedia.org/wiki/Statistical_machine_translation

[19] S.K. Dwivedi and P. P. Sukadeve, "Machine Translation System Indian Perspectives", *Proceeding of Journal of Computer Science Vol. 6* No. 10. pp 1082-1087, May 2010.

[20] "Machine Translation –A Rosetta stone for the 21th century?", http://www.ida.liu.se/~729G11/projekt/studentpapper-10/maria-hedblom.pdf

[21] "Rule-based machine translation", http://en.wikipedia.org/wiki/Rule_based_machine_translation

[22] Robin, "Machine Translation-Natural Language Processing-Rule based machine translation",http://language.worldofcomputing.net/category/machine-translation/page/2

[23] Makoto Nagao, A Framework of A Mechanical Translation between Japanese and English by Analogy Principle, *In Artificial and Human Intelligence* 1984, http://www.mt-archive.info/Nagao-1984.pdf

[24] Indranil Saha et.al. (2004). Example-Based Technique for Disambiguating Phrasal Verbs in English to Hindi Translation. Technical Report KBCS Division CDAC Mumbai.