IOSR Journal of Computer Engineering (IOSR-JCE)
e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 16, Issue 1, Ver. II (Jan. 2014), PP 22-25
www.iosrjournals.org

Machine Translation Approaches and Design Aspects

Ruchika A. Sinhal, Kapil O. Gupta
Dept. of CSE Shri Ramdeobaba College of Engineering and Management Nagpur
Dept. of IT Datta Meghe Institute of Engineering Technology and Research Wardha

Abstract: *Machine translation is a sub-field of computational linguistics that investigates the use of software to translate text or speech from one natural language to another. On a basic level, MT performs simple substitution of words in one natural language for words in another, but that alone usually cannot produce a good translation of a text, because recognition of whole phrases and their closest counterparts in the target language is needed. The paper focuses on Example Based Machine Translation (EBMT) system that translates sentences from English to Hindi. Development of a machine translation (MT) system typically demands a large volume of computational resources. For example, rule based MT systems require extraction of syntactic and semantic knowledge in the form of rules, statistics-based MT systems require huge parallel corpus containing sentences in the source languages and their translations in target language. Requirement of such computational resources is much less in respect of EBMT. This makes development of EBMT systems for English to Hindi translation feasible, where availability of large-scale computational resources is still scarce. Example based machine translation relies on the database for its translation. The frequency of word occurrence is important for translation.*

Keywords: *Machine Translation, Problems, Process of MT, Approaches*

I. INTRODUCTION

The natural language is used by every common man. The language which we speak is termed as natural language. We use natural language for communication. Natural language processing is processing, refining, modifying and translating i.e. operating on one of these natural languages.

1.1 Need of Machine Translation

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]. In 1990, 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. [2, 3]

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 [4]. MT also overcomes the technological barriers. Most of the information available is in English which is understood by only 3% of the population [5]. 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.

1.2 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 [6]. 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 [7]. 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 [7].

Pronoun Resolution

The problem of not resolving the pronominal references is important for machine translation. Unresolved references can lead to incorrect translation [7].

Idioms

An idiomatic expression may convey a different meaning, that what is evident from its words. For example, an idiom in English language „No brick in their walls", would not convey the intend meaning when translated into Hindi language [7].

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 [7].

II. PROCESS OF MACHINE TRANSLATION

Machine translation is the process of translating from source language text into the target language. The following diagram shows all the phases involved.

Fig. 1 A Typical Machine Translation Process

Text Input

This is the first phase in the machine translation process and is the first module in any MT system. The sentence categories can be classified based on the degree of difficulty of translation. Sentences that have relations, expectations, assumptions, and conditions make the MT system understand very difficult. Speaker"s intentions and mental status expressed in the sentences require discourse analysis for interpretation. This is due to the inter-relationship among adjacent sentences. World knowledge and commonsense knowledge could be required for interpreting some sentences.

Deformating and reformatting

This is to make the machine translation process easier and qualitative. The source language text may contain figures, flowcharts, etc that do not require any translation. So only translation portions should be identified. Once the text is translated the target text is to be reformatted after post-editing. Reformatting is to see that the target text also contains the non-translation portion.

Pre-editing and Post editing

The level of pre-editing and post-editing depend on the efficiency of the particular MT system. For some systems segmenting the long sentences into short sentences may be required. Fixing up punctuation marks and blocking material that does not require trarnslation are also done during pre-editing. Post editing is done to make sure that the quality of the translation is upto the mark. Post-editing is unavoidable

especially for translation of crucial information such as one for health. Post-editing should continue till the MT systems reach the human-like.

Analysis, Transfer and Generation

Morphological analysis determines the word form such as inflections, tense, number, part of speech, etc. Syntactic analysis determines whether the word is subject or object. Semantic and contextual analysis determine a proper interpretation of a sentence from the results produced by the syntactic analysis. Syntactic and semantic analysis are often executed simultaneously and produce syntactic tree structure and semantic network respectively. This results in internal structure of a sentence. The sentence generation phase is just reverse of the process of analysis.

Morphological analysis and generation

Computational morphology deals with recognition, analysis and generation of words. Some of the morphological process are inflection, derivation, affixes and combining forms. Inflection is the most regular and productive morphological process across languages. Inflection alters the form of the word in number, gender, mood, tense, aspect, person, and case. Morphologicalanalyzer gives information concerning morphological properties of the words it analyses.

Syntactic analysis and generation

As words are the foundation of speech and language processing, syntax can considered as the skeleton. Syntactic analysis concerns with how words are grouped into classes called parts-of-speech, how they group their neighbors into phrases, and the way in which words depends on other words in a sentence.

Grammar formalism

Grammar formalism is a framework to explain the basic structure of a language. Researchers propose the following grammar formalisms:

Phrase Structure Grammar (PSG) Dependency Grammar Case Grammar Systematic Grammar Montague Grammar

The variants of PSG are

Context Free PSG Context Sensitive PSG Augmented Transition Network Grammar (ATN) Definite Clasue (DC) Grammar Categorical Grammar Lexical Functional Grammar (LFG) Generalised PSG Head Driven PSG Tree Adjoining (TAG)

Not all the grammars suit a particular language. PSG, for example, does suit Japanese while dependency grammar does suite. Case grammar is popular as sentence in differet languages that express the same contents may have the same case frames.

Parsing and Tagging

Tagging means the identification of linguistic properties of the individual words and parsing is the assessment of the functions of the words in relation to eah other.

Semantic and Contextual analysis and Generation

Semantic analysis composes the meaning representations and assigns them the linguisticinputs. The semantic analyzers use lexicon and grammar to create context independent meanings. The source of knowledge consists of meaning of words, meanings associated with grammatical structures, knowledge about the discourse context and commonsense knowledge.

III. APPROACHES IN MACHINE TRANSALTION

The MT systems can broadly be categorized on the basis of its knowledge type, its representation and interpretation.

3.1 Knowledge Based MT

"The term knowledge based MT has come to 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 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].

3.2 Statistical MT

The model works with the intuition that the translated sentence has been passed through a noisy channel, which distorted the source sentence S to the translated sentence T. 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 [6]. 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 Pr (S/T) = probability that the decoder will produce S when presented with T, Pr(T/S) = probability that the translator will produce T when presented with S

3.3 Example Based MT

EBMT is a corpus based machine translation, which requires parallel-aligned 3 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.

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 [7].

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

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

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

The adapted chunks are –
- [vaha] [jana]
- [kharidaraha 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.

REFERRENCE

[1.] Hutchins W. John and Harold L. Somers (1992). An Introduction to Machine Translation. London: Academic Press.
[2.] Hutchins and Lovtsky, in press.
[3.] Hutchins, J. 1986. Machine Translation:Past, Present, Future, Ellis Horwood/Wiley, Chichester/New York.
[4.] Sergei Nirenburg and YorickWilks, Machine Translation
[5.] D. D. Rao, "Machine Translation A Gentle Introduction", RESONANCE, July 1998.
[6.] "Statistical machine translation", [Online]. Available, http://en.wikipedia.org/wiki/Statistical_machine_translation
[7.] IndranilSaha et.al. (2004). Example-Based Technique for Disambiguating Phrasal Verbs in English to Hindi Translation. Technical Report KBCS Division CDAC Mumbai.