ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

Design Aspects in Machine Translation

Ruchika A. Sinhal, Manoj B. Chandak

***Abstract* - This paper gives the brief introduction about machine translation. The field of machine translation is vast and useful. The paper describes its history. The process involved in translation, main basic approaches, database formation and the vital theory in the translation its need and drawbacks.**

Keywords: Machine Translation, Ambiguity, Process of Translation

I. INTRODUCTION

Why should we be interested in using computers for translation at all? The first and probably most important reason is that there is just too much that needs to be translated, and that human translators cannot cope. A second reason is that on the whole technical materials are too boring for human translators, they do not like translating them, and so 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. Lastly, companies want to reduce translation costs and on the whole with machine translation (MT) and translation tools they can achieve them. Any one of these reasons alone can be sufficient justification for using and installing either MT systems or computer translation aids.[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 in 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.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

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, computerised 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 used in the sense of translation of one language to another. The ideal aim of machine translation systems is to produce the best possible translation without human assistance. Basically every machine translation system requires programs for translation and automated dictionaries and grammars to support translation.

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.

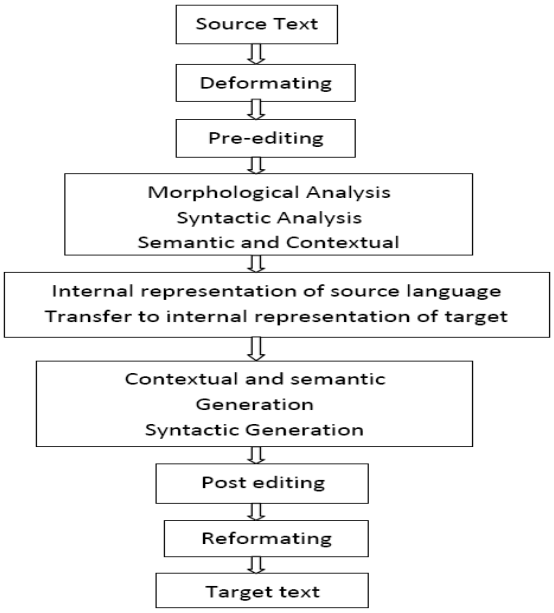

Figure 1: A Typical Machine Translation Process

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.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

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. Morphological analyzer 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 linguistic inputs. 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. TYPES OF MACHINE TRANSLATION

Machine translation system generally has two types of systems. They are:-

a. Bilingual Systems
b. Multilingual Systems

Machine translation systems that produce translations between only two particular languages are called *bilingual systems* and those that produce translations for any given pair of languages are called *multilingual systems*. Multilingual systems may be either uni-directional or bi-directional. Multilingual systems are preferred to be bi-directional and bi-lingual as they have ability to translate from any given language to any other given language and vice versa.

VI. 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.

a. 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:-

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

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

1 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.

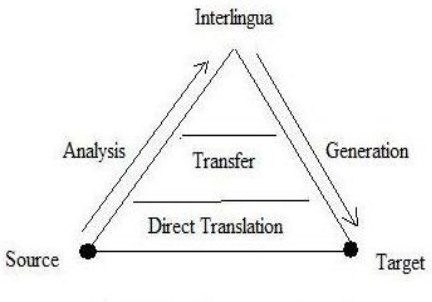

Figure 4 Machine Translation Pyramid

2 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.

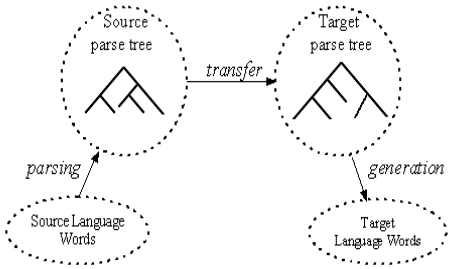

Figure 5. Transfer Model

3 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

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

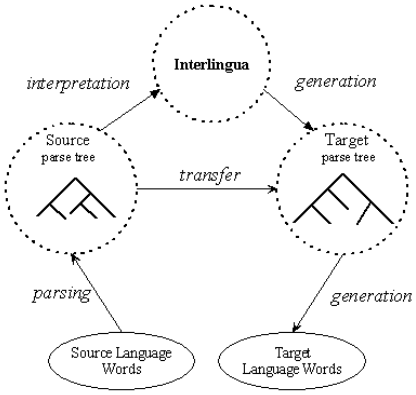

Figure 6. Interlingua Model

b. 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.

c. 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.

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.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

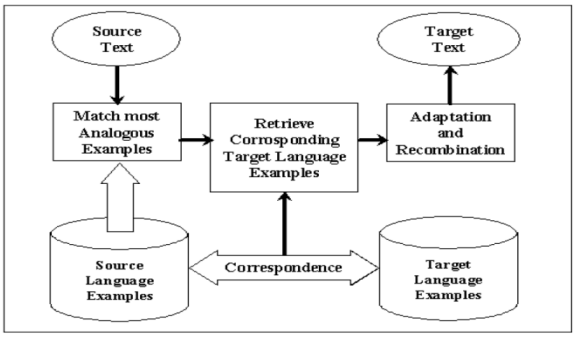

Figure 7 EBMT Architecture

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.

V. CHALLENGES IN MT

The translation task is not so simple as it appears; it has many challenges like – Polysemy, Homonymy, Synonyms, Metaphors and Symbols, new vocabulary developments, Lexical and Structural mismatch between the languages, Idioms and Collocations, complicated structures, referential ambiguity and ambiguities in the source and target languages.

Polysemy is the ambiguity of an individual word or phrase that can be used in different contexts to express two or more different meanings, such as *play*, *table*, *bank*, etc. The proper translation is difficult even for a human translator. Homonyms, on the other hand, are two or more words that share the same pronunciation and often the same spelling but differ in meanings such as *noun quail* and *verb quail*. Metaphors and Symbols depend on the underlying culture and history, which often cannot be translated. A Metaphor is a figure of speech in which a word or phrase that normally designates one thing is used to designate something else, such as *Sea of troubles*.

The Lexical and the structural mismatch are due to the difference in the way each language expresses ideas or feelings. For example, in Hindi language, the verb is inflected according to the gender of the subject in the sentence, whereas no such phenomenon is observed in English language.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

The multi-word constructs like Idioms and Collocations add more challenge in translation, as their meaning can't be derived from its constituents. In addition, multi-word constructs like phrasal verbs in English language exhibit different meaning in different context.

Perhaps the most important problem in language translation is ambiguity resolution. Since our project is concerned with a type of ambiguity resolution, we look at this problem in some detail in the next section.

a. Ambiguity - a Major Challenge

A major problem in MT is ambiguity in the source and target language. If a sentence or phrase is ambiguous, it has more than one interpretation. The ambiguity in a language can occur at various levels, Morphological, Syntactic, Semantic or Pragmatic.[4.]

Lexical ambiguity occurs when a word possesses more than one meaning. For example, the word *bank* has two meanings: *bank of a river* and *financial institution*. Consequently the sentence "*John went to the bank"* has two interpretations, due to the ambiguity of the word *bank.*

Structural ambiguity is concerned with the syntactic representation of sentences. It occurs when more than one valid syntactic structure can be associated with a given sentence. The ambiguity in prepositional phrase attachment is one source of structural ambiguity. For example, the sentence "*The detective found the man with a torch"* can be interpreted as follows –

1. The detective found a man with a torch in man's hand.
2. The detective with a torch found a man.

The corresponding parse trees are shown in Figure 8. Without resolving this ambiguity, the system cannot choose between these two candidate parse trees.

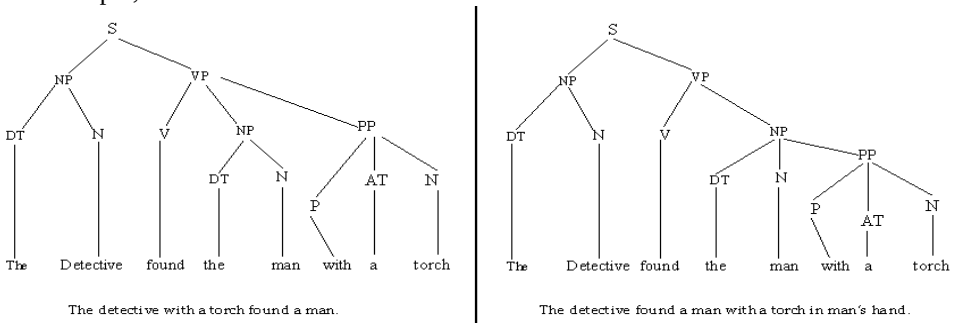

Figure 8 Ambiguous Parse Trees

Another kind of ambiguity is referential ambiguity, in which pronouns refer to certain words, but it is often difficult to find out, to which word it is referring. Sometimes, the references might even cross the sentence boundaries. For example, in the following two sentences – *she dropped the plate on the table and broke it.* Here, the ambiguity is that the pronoun *It* refers to which noun, the *table* or the *plate*.

The multi-word constructs – like Idioms and Phrasal verbs - are yet another cause of ambiguity. For example, the idiomatic phrase "*tip of the iceberg*" literally means some portion of the iceberg, whereas in the sentence "*The information you see on your computer screen after you do a web or database search is just the tip of the iceberg*", says that the information we see on the screen is just a small portion of the relevant knowledge. The actual meaning of the idiom can't be identified from its constituent words.

Phrasal verbs are highly context dependent in English language and are composed of a verb followed by particle, like – bring up, put on, bring down, etc. Phrasal verbs have different meanings in different contexts.

Let us consider the phrasal verb: *bring up,* which can be used in many ways are as follows –

1. John *brought up* an orphan child.
2. Child *brought up* the toy from the floor.
3. The labour minister *brought up* an issue for discussion in the parliament.
4. The students *brought up* the matter before principal.

ISSN 2249-6343
International Journal of Computer Technology and Electronics Engineering (IJCTEE)
National Conference on Emerging Trends in Computer Science and Information Technology (NCETSIT-2011)

In these sentences, the phrasal verb *bring up*, has four different meanings – *to rear/educate*, *compositional*, *to introduce* and *to call attention*. This project addresses its disambiguation of phrasal verbs in English.

REFERENCE

[1.] Hutchins W J (2003). *Machine translation and computer-based translation tools: what's available and how it's used.* Presentation at the University of Valladolid (Spain).

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

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

[4.] Cohen, J.M., "Translation", *Encyclopedia Americana*, 1986, vol. 27, pp. 12–15.

Ruchika A. Sinhal

Student of MTECH final year from CSE discipline from Shri Ramdeobaba Kamla Nehru Engg. College. Have completed engineering from Nagpur University. Presently working with my guide Mr. M. B. Chandak in field of Machine Translation. I have published my research work in International Conference published by Springer.

Prof. Manoj B. Chandak

Prof. M.B. Chandak heads the Department of Computer Science & Engineering at Shri Ramdeobaba Kamla Nehru College of Engg, Nagpur. He has obtained B.E. Computer Technology from Nagpur University and has done his M.E. (Computer Sciences) from Amravati University, Amravati.