CHAPTER FIVE

SUMMARY AND CONCLUSION

The processing of natural language by the computer to simulate human linguistic behavior is called NLP or Natural language processing. This is a very important element in translation and in today’s translation systems NLP is able to paraphrase a text that is input, to translate a text into another language and to answer questions about the contents. NLP is unable to draw inferences from a given text like human beings, but is able to do all the other things mentioned above perfectly. Linguistics and Computer Science have helped in developing NLP to a large extent to help in translation. Research in this field began in the early 1940s and has helped in developing efficient translation systems.

The research was supported by two groups; the language processing community who supported the theoretical perspectives of generative grammar and the speech community supporting the statistical information theory. Though practical work did not progress much in the beginning, theoretical work proceeded smoothly. From the 1980s translation began to gain popularity as there were many areas in which computers were used and all the users did not know English. The research in NLP in the last twenty years has advanced very well and this is due to the connectivity supported by the internet and the higher need for translation at a global level. A number of companies have found different technologies for translation based on the need of the user.

Language processing (NLP) dealing with analysis of the language to derive a meaning and language generation (NLG) to produce a language expected by the user are two important aspects of NLP. NLG is a little different from NLP as it needs to
plan and strategize a method to generate the language required and so it is more difficult than NLP. NLG can either be a ‘One dimensional’ or a ‘Multi dimensional’ model depending on the user’s requirement. NLP covers the areas such as phonology, phonetics, lexicon, morphology, syntax, semantics and pragmatics in languages so that translation system can give the correct output for the user.

Although MT was only a mere concept in the 17th century, it became a reality in the twentieth century. In the twenty first century with the use of modern technology and communication systems MT has become a very significant factor globally. From the memorandum stage by Warren Weaver in 1949 to the stage after the 1980s with bilingual corpora being stored in the system, MT has changed according to the changes in the computer systems. Computer-Aided Translation uses both machine-aided human translation system and human aided machine translation and there is no clear boundary between these two systems.

MT uses both bilingual and multilingual methods for translation and these systems can either be unidirectional or bidirectional depending on the product. These systems initially followed the direct translational approach where as today the technology is called the indirect approach that uses the interlingua approach which can convert the SL text into more than one language. The indirect system also uses a transfer method which is a system in-between the two. MT has to closely work with languages based on the linguistics pertaining to each language it deals with. Any MT system must have the basic grammatical rules, phonology and morphology of the different languages it deals with. There are languages that have non-roman alphabet script like Arabic which need to be standardized in order to have a perfect output. The lexical ambiguity of the words in each language must be dealt with carefully and this can be done only if the developer works along with experts in the languages the system supports.
Over the years several models of MT have been developed namely the rule-based MT (RBMT), constraint-based MT, Principles-based MT, Example-based MT also known as Memory-Based MT and Hybrid systems that use the strengths available in both statistical and rule-based translation. The text that is input can be represented based on the logical structure of the content and this concept is the basis of Role Reference Grammar, a functional theory of language used in modern MT systems. This system allows a direct mapping between the syntactic and semantic representations of a sentence with the help of a linking algorithm. Due to the remarkable changes in the telecommunications, the world is witnessing a networking of translational services. The demand for this service is very high as translation is required in many situations. Apart from the traditional type of translation there is need for online or real time translation especially for news and other such important events that are to be disseminated to users globally. This type of translation is also called network-based translation and a number of companies provide services for the same.

In this study it is observed that Arabic is a language that is typologically and structurally very different from other European languages and so the schema used in MT systems for translation of English to any other European languages has to be completely changed in order to have a perfect output when English is translated to Arabic. While most languages are written from left to right, Arabic is written from right to left and this language makes use of diacritics for clear pronunciation of consonants which a developer has to take into consideration while translating English to Arabic. A number of other nuances that are specific to Arabic should also be considered during translation. The basic building block of Arabic is stem generation; in a number of examples given in the thesis it is clear that stems are generated from the same root and so the words are semantically related and these words that have the same stems exhibit similarities in style. Words in Arabic are generally very complex.
and they are made up of proclitics, prefixes, stems, suffixes and enclitics. The spoken language has many different forms though all the different forms have descended from Classical Arabic used in Quran. Arabic verbs are similar to other Semitic languages with a root; the root gives the basic meaning of the verb. Changes made in the vowels between consonants either with prefixes or suffixes can influence and bring about changes in other components like person, gender, number, tense, mood and voice.

Rule-Based Machine Translation is machine translation that is based on the rules abstracted from the parallel corpora of the two languages and elsewhere, further supplemented by the database of bilingual dictionaries.

These systems help in giving near perfect outputs based on the morphological, syntactic, and semantic analysis of both the source and the target languages. In the late seventies the first English to Arabic translation system was developed based on direct method using the unidirectional system which did not analyse the semantic or syntactic structure. This was a simpler system that allowed one to directly translate the Encyclopedia Britannica to Arabic. Later a few more companies like Systran, Japanese MT systems and EUROTRA came up with translation system that did similar types of translation which had its own limitations. Apart from rule based machine translation and knowledge based machine translation there are also corpus based systems and hybrid systems which are different approaches used for machine translation. The rules used in this approach are prepared based on the parts of speech and there are two types of rules used here, one is transfer link rule and the other is morphological rule.

The three types of RBMT are Direct RBMT, transfer based RBMT and interlingual RBMT. Direct systems use the basic concept used in Dictionary, transfer based system is based on the idea of Interlingua and the interlingual RBMT systems
use an abstract meaning where the text to be translated is transformed into an interlingua which is an abstract language-independent representation. The translation systems available presently work on language agreement property displaying consistency in the morphology where the agreement between verb and subject is a very important factor to be considered by the system apart from gender and person agreement as well. Arabic shows word-order asymmetries for agreement which makes it all the more difficult to translate from or to English.

In RBMT 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). Based on the structure of the sentence these two parts can be further divided. Only a sentence with the right structure can be converted into a correct translated sentence. Translation in RBMT is done based on the pattern that matches the rules and also knowledge and reasoning are used for understanding the language as well for translating contents that are ambiguous in nature. One can also use context specific knowledge for inferring the meaning based on the rules to derive the correct output.

The initial phase in machine translation is text input where all types of sentences are fed into the system. In machine translating system it is essential that the figures, flowcharts, and other contents that not get translated are deformatted initially and after the translation is done for the text a process called reformatting is conducted wherein the system puts back the contents that were not translated back in their locations. The target text which is the final phase is the output that contains all the information that was in the original text. The source text gets pre-edited before translation and post editing is a significant step after translation as they help in getting the correct output.
In Machine Translation some specific problems are faced due to the differences in the grammar of the source language and the target language. Agreement and word reordering problems are very common between languages. Noun-formation which is a sub-process of word-formation is a fairly difficult area of research in both Standard English and MS Arabic. Adjective-noun agreement is not available in all languages and in translation this poses a problem at times when the languages are not similar in their grammatical structures. Verb subject agreement is an important feature in English, but that is not very uncommon in languages derived from Persian but when one tries to translate English into Arabic, problems arise as there are subtle differences that also have to be taken into account while translating the English text to Arabic.

It is possible for a rule-based machine translation system to be made very effective if the system functions with rigorous rules. This is possible if the developer works closely with experts who have the knowledge of both the languages and the developer also has a very good knowledge of the grammar involved in the translation. If a fully automated, high quality machine translation (FAHQMT) is developed, then it can improve the quality of MT output and increase its usefulness.

English-Arabic translation was based on transfer models till the early 1990s but later by early 2000s the translation software began to use artificial intelligence. The Rule Based Machine Translation system is being used extensively since the time SYSTRAN was developed in the early 1970s. In this system there are three basic modules available that is essential for the system to give the perfect output the user expects. The modules are a) a parallel aligned corpora, b) digitized SL resources and c) architectural design that is specific and precise. Various types and subtypes of sentences are dealt with in the parallel aligned corpora and this being a huge task is also a difficult task as the corpora must have different English sentences translated.
into Arabic. The digitization of SL resources involves the different types of
dictionaries translated into Arabic which deals with complex concepts like translation
of idioms and phrases. The architecture should specifically help in matching the
patterns, tokenization, dictionary lookup, morphologising, parsing, identifying
sentence types, disambiguation, generation of Arabic sentences and reorganizing the
sentences to output the Arabic equivalent.

The workflow in this architecture explicitly shows how a sentence, once input,
can be sent through the parser and the tokenizer, get morphologised and finally after
getting identified gets translated into an Arab sentence, which gets rearranged
properly by the software to yield the expected meaning.

This has been implemented using Java and SQL. The software was coded
using dot net. The software developed includes the types and subtypes of imperative
sentences in English and converts them into Arabic sentences based on the user’s
requirement. The system developed here has been able to avoid certain discrepancies
found in other software like Systran and Google though the scope to improve it
further is always there.

Rule-based MT will certainly go a long way in making the machine simulate
human linguistic behaviour. It is arguably the only consummation of the marriage
between language and engineering issuing in a legal progeny, in the area of Machine
Translation.