Chapter Four
NATURAL LANGUAGE PROCESSING AND MACHINE TRANSLATION

This chapter provides a comprehensive introduction to the field of Natural Language Processing (NLP), covering the major techniques, levels and applications and an overview of machine translation.

4.0 Section A: Natural Language Processing (NLP)

4.1 Natural Language Processing (NLP)

Originally, Natural Language Processing was referred to Computational Linguistics. Compared to other information technology approaches, NLP is a relatively recent area of research and application. The sufficient success in the area of NLP encouraged research in information systems now and in future. Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications (Liddy, 2001).

There are three main contributors to the emerging of NLP discipline and practice. The first contributor is Linguistics – deals with the constructions, forms and structures of language. It concentrates on the universals of language. The second contributor is Computer Science – focuses on the development of the internal representations of data and the processing of these linguistic structures. The third discipline is Cognitive Psychology – deals with language as an indicator to the cognitive processes conducted by human when use language.

In fact, there are two distinct domains in Natural Language Processing - language processing and language generation. In the first domain the focus is on the language analysis in order to produce meaningful representations, while in the latter domain the focus is on the production of language forms the meaningful representations.

Natural language processing implies computer understanding, analysing & generating of natural languages. It aims at enabling computers to achieve human-like
comprehension and processing of languages. The task of Natural Language Processing is equivalent to the role of reader/listener. There are other different practical goals for NLP depending on the certain application of NLP. For example, in the case of the IR system as an application of NLP has the goal of representing the true meaning and intent of the user’s query, such query can be expressed as naturally in everyday language. As well as, the contents of the texts that are being retrieved will be introduced at all their levels of meaning so the right aligning between the query and response can be found.

4.1.1 Levels of Natural Language Processing (NLP)

According to the language synchronic model, the levels of human language processing follow one another in a strictly sequential manner. Levels of NLP explain the actual nature of NLP system and they represent active and dynamic interaction with each other. These levels carry the meaning of language and it is important to use these levels in order to understand language and later for language generation. Different NLP systems use different levels, or combinations of levels of linguistic analysis and such systems can said to be NLP-based systems. The follow description will represent the different levels of NLP in a sequential manner (phonological, morphological, lexical, syntactic, semantic, discourse and pragmatic).

4.1.1.1 Phonological Level

The focus of this level is the interpretation of spoken input or precisely the speech sounds within and cross the words. When NLP-based systems deal with spoken input, the phonological analysis process started by analyzing the sound waves then encoding them into digitized signals. These digitized signals enable the interpretation of the spoken input using different rules or by comparing them to the utilized language model.

There are, in fact, three types of rules used in phonological analysis: 1) phonetic rules – for sounds within words; 2) phonemic rules – for variations of pronunciation when words are spoken together, and; 3) prosodic rules – for fluctuation in stress and intonation across a sentence (Liddy, 2001).
4.1.1.2 Morphological level

This level concentrates on the componential nature of word. A word can be a single morpheme or a combination of more than one morpheme. Each morpheme carries the same meaning across words. In order to understand the meaning of any unknown word requires breaking down the word into its constituent morphemes. So, any NLP-based system can understand the meaning conveyed by each morpheme in order to generate the meaning.

4.1.1.3 Lexical level

This level deals with words. At this level, NLP-based system tries to interpret the meaning of the individual word. To understand the word meaning, there are many processes to be taken into consideration. The first process is the assignment of the POS (parts of speech) to each word. Sometimes to assign a part of speech tag to a certain word is challenged. For example, a word can function as more than one part-of-speech; in this case POS is assigned based on the context in which they occur (Qassim, M.M., 2015).

Sometime in order to perform the lexical level analysis and understanding, this requires a lexicon and its nature and extent of encoded information are determined by the utilized NLP system. Lexicons may be quite simple, with only the words and their part(s)-of-speech, or may be increasingly complex and contain information on the semantic class of the word, what arguments it takes, and the semantic limitations on these arguments, definitions of the sense(s) in the semantic representation utilized in the particular system, and even the semantic field in which each sense of a polysemous word is used (Liddy, 2001).

4.1.1.4 Syntactic level

Generally, syntax represents the meaning in most languages as words order and dependency relationship between words contributes to meaning of a language. For example, in the case of the two sentences: ‘The dog chased the cat.’ and ‘The cat chased the dog.’ They are different only in their syntax but convey quite different meanings.

At this level the concentration is at the analysis of words in a sentence in order to interpret the grammatical structure of the sentence. To cover this type of syntactic
analysis we need to represent both a grammar and a parser. The output of this analysis is a representation of the sentence which introduces the structural dependency relationship between the words. From various grammars the NLP-based system can utilized certain grammar and this will impact the choice of the parser.

4.1.1.5 Semantic level

As thought by many people that semantic level determines the meaning of language, but as mentioned above all the linguistic levels contribute to the meaning of the language. The semantic analysis contributes to the sentence meaning by focusing on the interaction among word-level meanings in their sentence. In the case of the polysemous words with multiple senses, the semantic disambiguation analysis permits only one sense of polysemous words to be represented in the semantic representation of the sentence. In fact, there are various methods to accomplish the disambiguation of the language. Each method requires certain type of information; some needs the frequency information with which each sense occur in the corpus, others require consideration of the context, and some require the pragmatic knowledge of the domain of the document.

4.1.1.6 Discourse Level

This level of analysis concentrates on the properties of the text as a whole as it conveys meaning by connections between component sentences. It deals with the whole text of the language. At the discourse level analysis there are different types of discourse processing. Discourse/text structure recognition and anaphora resolution and are the most popular types. Discourse/text structure recognition determines the functions of sentences in the text, which, in turn, adds to the meaningful representation of the text. For example, newspaper articles can be analyzed into discourse components such as: Lead, Main Story, Previous Events, Evaluation, Attributed Quotes, and Expectation. Anaphora resolution is the replacing of words such as pronouns with the suitable entity to which the anaphora refers.

4.1.1.7 Pragmatic Level

The pragmatic level focuses on the purposeful use of language in situations. It goes beyond the content of the texts for understanding. The main objective of this level processing is to explain how extra meaning is read into texts without being encoded in these texts. To perform such reading, this requires understanding of intentions, plans, and goals.
4.1.2 Applications of Natural Language Processing (NLP)

There are several applications of NLP. And, any application that utilized NLP is considered a candidate for NLP. The follow is a representation of the frequent applications of NLP: Machine Translation, Information Extraction, Information Retrieval, etc.

4.1.2.1 Machine Translation

Machine translation is the oldest application of NLP. Machine translation processes utilize different levels of NLP, ranging from low to higher levels of analysis.

4.1.2.2 Information Extraction (IE)

In contrast to Machine Translation, Information Extraction is considered as new application of NLP. IE concentrates on the recognition, tagging and extraction into structured representations. Other applications can utilize these extractions of information, such as visualization, question-answering, and data mining. The main elements of constructed information are persons, locations, companies, organizations, foundations etc.

4.1.2.3 Dialogue Systems

This application of NLP concentrates on a narrowly defined application, for example, home sound system. Recently, Dialogue System started using the phonetic and lexical levels of language processing. This empowers the habitable role of dialogue systems.

4.1.3 Summarization

This application require higher levels of NLP analysis, in particular, the discourse level as it reduces a larger text into a shorter, yet richly constituted abbreviated narrative representation of the original document.

4.2 Section-B: An Overview of Machine Translation

4.2.1 Introduction to Machine Translation

In late twentieth century, Human language technologies emerged because of the rapid development in the different technological domains. Machine translation is one of such technologies which have assumed great significance in modern time. It is the use of
computers to automate some or all of the process of translating from one language to
another (D. Jurafsky, J. M. Marten, 2008).

The field of machine translation is almost as old as the invention of computer itself
(Blekhman and Pevzner, 2000). Scientists began to take real steps to realize the dream
and vision of Descartes who wrote in 1629 about a mechanical process to convert one
human language to another.

In 1949 an American scientist Warren Weaver sent the memorandum to The
Rockefeller Foundation (American institution supporting the scientific research), in
which he demanded starting the research on the automation of translation between
natural languages (Arnold et al., 1994). Warren Weaver was inspired by cryptographic
techniques, which were developed very strongly during the years of the Second World
War, and he thought that there existed some fundamental similarities between these
cryptographic techniques and the process of translation between human languages
(Waibel et al., 2000). In fact, it seemed that the challenges and difficulties of machine
translation are more complicated than what Warren Weaver imagined.

The relationship between the computer and human languages started in the beginnin-
g of 1940s. Machine Translation was the first application and interest of this relationship.
First MT used for military and strategic objectives. The outputs of that early machine
translation were very poor.

The main reason behind those poor outputs was the deficiency of linguistic, non-
linguistic and technical knowledge when translating from the source to the target
language. Later, due to the evolution of researches in natural language processing and
machine translation; MT developed and several machine translation systems and
models were developed.

Historically, practical machine translation is very old and starts in 1952 after the Second
World War. In 1954 George Town developed an automatic translation system from
Russian to English. He involved more than sixty sentences, six linguistic rules, 250
items in dictionary list. The proposed system was specialized in the area of organic
chemistry.

Machine translation (MT) is aimed to enable a computer to transfer natural language
utterances in either text or speech from one language into another while preserving the
meaning and interpretation. Translation value can be measured according to the extent of equivalency between the input language and the output. MT technology has gone through several paradigms from its very beginning in the past half century, including word-to-word direct translation, rule-based transfer approach, inter-lingua approach, Example-based machine translation (EBMT), statistical Machine Translation (SMT) and knowledge-based machine translation (KBMT). The direct translation approach relies too much on dictionary look-up. The transfer approach incorporates language analysis and representation at various linguistic levels, but cannot find adequate knowledge to resolve ambiguities involved in the language analysis, transfer and generation. Thus, there are problems with other approaches too.

In the past decades, experts (e.g., syntacticians) wrote down expert knowledge (e.g., grammar rules) in some rule-based format. This achieved significance but only with limited success. The inadequacy of manually encoded knowledge remains a problem. There are always so many practical ambiguities that experts cannot foresee during the construction of knowledge base. The maintenance and scale-up problems also emerge when the knowledge base becomes larger and larger. For example, changing a single rule might cause unpredictable conflicts with other rules and, consequently, lead to a crash of the entire rule system. Thus, there is a necessity to look for an alternative approach to knowledge engineering for MT that can automatically acquire practical knowledge from, and also adapt itself towards real language data. Example-based machine translation model (EBMT) is considered to be one of the current attempts towards this goal.

In some approaches of machine translation; the pre-processing is necessary to be done and developed. For instance, the rule-based model needs processing the monolingual or bi-lingual dictionaries which needs developing some grammatical and morphological rules before designing an appropriate translation to another language. More analysis needs to be done for the source language.

The following figure 4.1 demonstrates that rule-based MT models have been classified to fulfill the language data representation with alignment and mapping, i.e. from direct MT that analyses and translates phrases in one step to Interlingua MT that analyses source language sentences to a representation level that in theory is language
independent. This is accompanied by more and more processing, both in parsing/analysis and in generation.

As illustrated in figure above, Vauquois proposed that the increasing levels of complexity move from the internal processing gradually away from words at the lower level and closer to concepts at the higher level. The levels of MT models can be classified according to abstraction level of the representation used for contrastive mappings. The classification is from bottom to up as illustrated in the figure 4.1 above.

At the bottom of the triangle is the lowest level: the direct transfer or the direct replacement systems. According to these systems each word or multi-word phrase in the source language is replaced with its equivalent in the target language.

The middle level: this level can be divided in to two parts the syntactic transfer and semantic transfer. In the case of the syntactic transfer, the best way to get the right target sentence grammar is to produce the syntax tree of source sentence and then map it into its equivalent in the target sentence. This requires a parser, a set of transfer (mapping) rules, and a bilingual lexicon.

The second part of this level is the semantic transfer. Indeed, there many phenomena which do not map across languages via syntactic analysis only, because most of the languages have their own semantic representation and way of representing meaning. To catch the meaning of any language you have to understand something of the meaning of what is being said, and of the idiomatic ways of expressing that meaning in the target language. This requires inventing some kind of meaning representation which reflects the meanings of what to say in the languages to be handled (shallow semantics).
It requires also a semantic analyzer, an extended lexicon including the semantic features, rules of demotion (a syntactico-semantic constituent is demoted from higher to lower in the syntax tree and promotion (the opposite). In fact most research and commercial machine translation systems use a combination of syntactic and semantic transfer, and at the same time an internal representation that contains features of both (syntactic transfer systems and semantic transfer systems). For examples: Eurotra (research); SYSTRAN, Logos (both commercial).

The highest level is Interlingua which analyses the source language phrases to reach the level of representation as language independent. The translation process starts, and here we use the source language data analysis. After that it reaches the generation of the target language phrases.

An Interlingua is a system of symbols and notation to represent the meaning(s) of language.

In fact, it is impossible to totally substitute human translator with machine translator. Machine Translation is an automation of the translation process this implies that MT is a partial substitute of human translator. Many challenges and problems are faced when we translate the source language into the target language and it is very difficult to for machine translation to tackle all these problems. For example, referential expressions, phrasal verbs, specific language collocation, etc. which require a lot of effort and human translator intervention. Furthermore, in order to achieve a very high degree of efficient and better translation in our output, the target language should be given the same efforts.

4.2.2 Machine Translation Architecture & Paradigms

It is crucial to distinguish between machine translation architecture and machine translation paradigms (or approaches). The former refers to the actual processing of translation, whereas the letter refers to the knowledge component that supports the processing field in translation. Recently, many researchers investigate different classes of machine translation paradigms. These paradigms cover most of the approaches of machine translation. A vast majority of which were reported in the last few years at a number of conferences including the Annual Meeting of the association for computational Linguistics (ACL), the International Conference on Theoretical and
Methodological Issue in Machine Translation (TMI), The Conference of the Association for Machine Translation in the Americas (AMTA), the International Conference on Computational Linguistics (COLING), and the Machine Translation Summit (MT-Summit) (J. Dorr et al, 2002).

The following figure represents the actual and basic architecture of machine translation (machine translation processes).

**Figure 4.2: Machine Translation Architecture**

### 4.2.3 Arabic Machine Translation Paradigms

Machine Translation has three main processes: analysis, transfer and generation. Indeed, through these processes MT generates the required translation from the source language into the target language. There are many machine translation approaches (or paradigms). Depending on the linguistic principles these approaches can be divided into three main paradigms: linguistic- based paradigms, non-linguistic- based paradigms and hybrid paradigms. The systems which are based on linguistic theories strive to utilized of the different linguistic properties relates to the constraints of syntax, lexicon, and semantic to generate an appropriate target-language realization of the source-language sentence. It is assumed that, Machine translation approaches are about building a probabilistic model to combine faithfulness and fluency for better translation. When such set of consecutive words occur frequently and statistics about their translations are
calculated, it might lead to better translation. The machine translation paradigms are illustrated in the following figure.

**Figure 4.3: Machine Translation Paradigms (approaches)**

Up till now, many totally different approaches to machine translations have been developed.

These are among others syntactic transfer, interlingua-based machine translation, knowledge-based machine translation, systems based on statistics or neural nets, etc. (Ney et al., 2000; Canals et al., 2000; Loukachevitch and Dobrov, 2000). Here is a detailed description of the common paradigms of MT in the following lines.

**4.2.3.1 Linguistic based - MT Paradigms**

Linguistic-based MT paradigms are MT methods which are well-grounded in linguistic theory and focus on the linguistic information in their analysis. They utilized the different levels of linguistic data- at lexical level, morphological level, syntactic level, semantic level, etc. Systems based on linguistic theory strive to use the constraints of the syntax, lexicon and semantics to produce an appropriate target language realizations of the source language texts. There are many MT paradigms that are based on the
linguistic information such Constraint- Based MT, Knowledge - based MT. Lexical- Based MT, Rule- based MT, Principle-Based MT and Dictionary-based MT.

4.2.3.1.1 Direct Approach (Dictionary based Machine Translation) (DMT)

Direct approach is also known as “Dictionary-based machine translation” or “binary translation”. The Direct method was popular in the first generation of machine translation systems. In Direct Machine Translation, dictionary is considered as an essential requirement in translation. Where the word’s equivalent is used to develop the translated verse. Dictionary based machine translation was the first generation of automated language translation. The first generation of machine translation (late 1940s to mid-1960s) was entirely based on machine readable or electronic dictionaries (Tripathi and Sarkhel, 2010). Despite the noted fast progress and development in machine translation approaches, most of them use dictionaries with other different techniques. This method is based on use of word’s equivalent and this makes it an effective method in phrases translations but not sentences.

![Source input](Morphological analysis) → (Bilingual Dictionary) → Local reordering → Target output

**Figure 4.4: Direct Machine Translation Approach**

4.2.3.1.2 Rule-based Machine Translation (RBMT)

Rule-based approach is the first method used by researchers in the field of machine translation. It is a complex translation method that requires a large amount of inputs and processing. RBMT is also called Knowledge Based Machine Translation that retrieves rules from bilingual dictionaries and grammars based on linguistic information about source and target languages (Bijimol, Abraham, 2014).

This method requires a large set of linguistic rules and diverse dictionaries. The software utilized these rules in three stages: analysis, transfer, and generation. Some of used dictionaries include main entries and other dictionaries include specialized vocabularies. The rules in this method are written by humans based on their linguistic information. The linguistic analysis in this method implies both the syntactic and semantic levels. The weakness of a rule-based approach is that it is impossible to write
rules that cover all languages, as this requires great linguistic knowledge (Charoenpornsawat et al. 2002).

According to Hutchins and Harold (1992) rule-based machine translation approaches can be classified based on their architectures into three main categories: Direct approach, Transfer based approach, and Interlingua approach.

Rule-Based Machine Translation (RBMT) systems utilize bilingual dictionary, collections of rules. These systems are manually developed over time by human experts, then latter generate the target structures. The dictionary specifies a set of rules for translating each word in the source language. After the words are translated, simple reordering rules are applied.

4.2.3.1.3 Knowledge-based Machine Translation (KBMT)

Knowledge-based (KBMT) Knowledge-Based Machine Translation (KBMT) systems are based on the point that “high quality translation requires in-depth understanding of the text” (Arnold et al., 1994). This approach requires mentioning real-world knowledge, as well as knowledge of the “differences in cultural backgrounds and differences in conceptual divisions” (Hutchins and Harold 1992) between diverse languages. “KBMT systems rely on an augmenter” (Trujillo, 1999).

The KBMT approach needs huge amounts of knowledge, a parser to align the source language into semantic representations and a generator to align those representations into the target language as shows in figure 4.5 below. By utilizing data mining and text mining techniques, knowledge-based machine translation systems can overcome several problems, such as lexical ambiguity, polysemy, and different shades of meanings, syntactic and structural ambiguity, discourses, and anaphoric ambiguity.

![Figure 4.5: Knowledge-based MT Approach](image)

63
4.2.3.2 Non-linguistic Based Paradigms (Corpus-based MT)

Non-linguistic-based paradigms are MT methods which use corpus as data-base and considered as the main component in developing any MT systems. There many MT paradigms that are based on corpus such as statistical machine translation, example-based machine translation, dialogue-based machine translation and neutral network-based machine translation.

In the early days of MT (1950s and 1960s) there were two contrastive approaches (Hutchins, 2006:380), usually dubbed the ‘empiricists’ and the ‘rationalists’. The former methodologies contained elements of statistical techniques in analyzing texts to derive dictionary rules, whilst the latter methodology followed a strictly linguistic approach, perhaps due to the inadequate computer facilities at the time.

4.2.3.2.1 Statistical Machine Translation (SMT)

The basic theory behind statistical machine translation is information theory; according to this theory text is translated based on a probability process. In SMT every source language segment (S) has any number of possible translations (T), and the most appropriate translation of that segment is the translation that is assigned the highest probability by the proposed system. As illustrated in figure 4.6 the statistical approach does not need any linguistic knowledge, but it does require a large sized bilingual corpus. SMT utilizes a bilingual corpus for each language pair (source and target), a monolingual corpus for the target language (TL), decoder and a language model whose parameters stem from the analysis of monolingual and bilingual corpora. The main advantage of SMT is its ability to generate suitable translations, even if a given sentence has not any equivalent or similar translation in the training corpus.

![Figure 4.6: Statistical Machine Translation Approach Model](image-url)
So when a new translation is presented, the SL sentences are segmented into smaller phrases. Then they are matched with source language equivalents in the bilingual corpus and their translations harvested by the utilized decoder. The decoder utilized a heuristic search algorithm to harvest and choose the most appropriate translations. Finally, the most probable translation is generated and then verified by the language model as a valid TL sentence otherwise; the translation process must be repeated.

SMT utilized a bilingual corpus which is aligned and tagged to equivalent at word level. This alignment is measured in fertility i.e. the ratio of how many TL words a source-language (SL) word can give rise to. Crossed links are possible in SMT systems. These links allows for caused by first, word order differences (known as distortion) due to the differences between languages at different levels. And second, varying degrees of fertility when mapping between words and their equivalence takes place, such as one-to-one equivalence (fertility=1), one-to-many equivalence (fertility=2) and one-to-zero equivalence (fertility=0).

As shown in the example above in figure 4.7 he is aligned to huwa (هو), here the degree of fertility is 1(one-to-one equivalent). Does not are aligned to la (no), here the degree of fertility is 2(two-to-one equivalent). The definite article (the) Al (ال) has no equivalent and the degree of fertility here is 0. The statistical translation method decomposes sentences in the bilingual corpus into smaller chunks and then calculates the probabilities of each TL phrase being a translation of its parallel SL phrase. These calculated probabilities are stored as ‘phrase tables’, in which multiple TL phrases are listed as possible translations of a SL phrases, each with a varying probability (Callison-Burch and Koehn). The probabilities are calculated during the preparation stage and stored. IBM’s work on French-English translation is considered as a classic example of SMT using the Canadian Hansards.
Recently, statistical machine translation has greatly contributed to the development of machine translation. It is now the most widely used method of machine translation, but statistical machine translation has not yet met the required quality.

4.2.3.2.2 Example-based Machine Translation (EBMT)

Example-based machine translation is a form of automatic translation that utilized a large corpus of previously translated sentences to generate translations of new sentences. The system does not have the whole sentence to be translated in its trained corpus. The system matches

Instead, the system matches the fragments (words and small phrases) and stitches them with the help of the target language model stored in the corpus. This method of translation was first proposed by Makato Nagao in 1984. He proposed that the automated translation is done by analogy.

“Man does not translate a simple sentence by doing deep linguistic analysis, rather, man does translation, first, by properly decomposing an input sentence into certain fragmental phrases, and finally by properly composing these fragmental translations into one long sentence. The translation of each fragmental phrase will be done by the analogy translation principle with proper examples as its reference.” – Nagao (1984)

Example based machine translation (also called as, memory-based, case- based, analogy- based and experience guided). EBMT and SMT both are seen as variants of corpus-based approaches to MT systems, because both require a large sized trained parallel corpus and they do not require linguistic knowledge. During the 1990s both of them became familiar at MT conferences. Typically, EBMT stands between RBMT and SMT, as it merges both data driven and rule based techniques.

In fact the main difference between rule based systems and case-based systems is indicated in quote by Riesbeck and Schanktak en as following;

“A rule based system will be flexible and produce nearly optimal answers but it will be slow and prone to error. A case based system will be restricted to variations on known situations and produce approximate answers but it will be quick and its answers will be grounded in actual experience. In very limited domains the trade-offs favour the rule
based reasoner but the balance changes as domains become more realistically complex.” (Riesbeck & Schank, 1989).

4.2.3.3 Hybrid Method (HM)

Hybrid method means merging different machine translation approaches or methods to produce better translation. Practically, it is difficult to merge fundamentally different approaches. Hybrid methods have been explored in the research community without any real success because they are utilized to incorporate abstract syntactic rules to reach the last translation. Here we require knowledge of how syntactic-based translation rules should be applied, knowledge of how words, phrases and patterns should be translated and knowledge of how syntactically based target structures were developed.

4.2.4 Why Machine Translation is Important?

According to WIKIPEDIA (2013), the significance of machine translation has increased since the middle of the twentieth century. Translation seems to be of great importance to the linguists. It is the basic means of transferring ideas, thoughts, and so on across cultures. More recently, globalization has affected the way of people lives all over the world and this huge world becomes a small village and the different cultures and languages are brought together. Accordingly, these rapid advances and developments in people lives require developing intercultural communication. The importance of translation comes from the importance of languages in human life (Matsuura, 2008). Machine translation has been a major focus of attention because of globalization, expanded media and technology, increased international trade, recognition of national minorities, increased migration, etc. There are several advances which increase the necessity for MT which made machine translation of a great value to meet the needs of modern life. Among those advances are the social media and networks such as Google Talk, WhatsApp, Skype, Yahoo Messenger, MSN Messenger, etc. Many people of different cultures and languages communicate with each other through these social media and MT plays a crucial rule as an urgent device to narrow the gap between those people.

Director-General of UNESCO, states that “Languages are indeed essential to the identity of groups and individuals and to their peaceful coexistence. They constitute a strategic factor of progress towards sustainable development and a harmony between the global
and the local context (Mattuura, 2008: 11). The demand for developing machine translation systems is very high especially there is no machine translation that fully meet people’s requirements, in terms of speed, quality, less cost and retrieval time of translation.

Finally, machine translation is the main administrative activity in addressing the natural language of various fields. Thus, efficient methods with special rules should be available to introduce a useful machine translation system to generate better translation of natural language texts into other natural languages, face challenges and remove anomalies.

4.2.5 Why Machine Translation is Difficult?

In general, there are several problems and challenges that face any machine translation system. As presented in Yates’s paper, the difficulties of MT were reviewed from the perspectives of the complexity of human language and translation. Mostly, the several problems and errors faced by Machine Translation result from the non-equivalence between the source and target languages in both linguistic and non-linguistic domains. Human languages are highly ambiguous. Actually the difficulty of MT is due to that language is full extraordinary features, exceptions and ambiguities at all linguistic levels. MT is not able to recognize all these extraordinary features and ambiguities without the human intervention. Most of the current MT systems deal only with morphological, lexical and limited syntactical analysis, far from sufficient analysis needed to support complicated translation and not to discuss any kind of the non-linguistic features such as cultural understanding, mood…etc.

In the case of Arabic machine translation, the foremost challenge is overcoming ambiguity and its rich and highly complex morphology. Arabic is notorious for its morphological ambiguity (Attia, 2006). According to Daimi (2001), Fehri (1993), Chalabi (2000), there are many complexities in Arabic. The following list is the major issues involving Arabic:

- Arabic writing direction is from right to left in a horizontal form.
- There are no capital letters in Arabic.
- Punctuation in Arabic is similar to English, except for commas, which sit ‘on’ the line instead of ‘under’ the line.
• Arabic uses gender for all known nouns (none are neutral).
• Space is left between words in sentences.

4.2.6 Challenges of Arabic Machine Translation

Within a particular context, the ambiguity in the main aspect is the property of sings, symbols, words, concepts, terms as indeterminate, indefinable or without a clear definition, and leading to an unclear meaning. Translation problems and ambiguities are the results of linguistic properties differences. The problems are mainly due to the differences in linguistic systems and languages.

Arabic words can often be ambiguous due to the three-letter root system. These consonant roots interlock with patterns of vowels or consonants to words or word stems. This root system allows the language to evolve to cover a wide range of meanings (Salem, 2009).

Solving problems in the analyzing and selecting of meanings will help to generate more efficient translation. A lot of studies have been conducted to narrow the wide gap between English and Arabic languages in the field of machine translation. (Soudi et al., 2012) focus on the specific problems of Arabic machine translation when translating from and into Arabic. (Habash, 2010) discusses several issues about Arabic as a modern standard language (MSA) and machine translation, such as morphology and Arabic script. (Abu Shquier, 2009) introduces a ruled based English-Arabic machine translation and focuses on specific syntactic features such as word agreement and ordering. (Attia, 2008) introduces the problem of morphological and syntactic ambiguities in Arabic and examines different methodologies to deal with these problems with a view to machine translation. Translation problems can be stated as linguistic elements that cause problems for translation process when the translator has to decide between more than one way of rendering it.

(Elming et al., 2009) improve the quality of translation by applying pre-translation syntactic reordering approach developed on a close language pair (English-Danish) to English-Arabic in a statistical machine translation.

Divergence is a language dependent phenomenon. It leads to translation ambiguity. It appears when a word in one language can be translated in more than one way into
another language. Mostly this phenomenon is different from language to another, in other words it is not that the same set of divergences will exists across all the languages.

In this chapter we are going to discuss the commonly challenges, problems and ambiguities in English to Arabic MT. Challenges of Arabic machine translation can be classifies into two main classes; linguistic and non-linguistic challenges as following:

**4.2.6.1 Linguistic Challenges**

In this section we will discuss the different types of linguistic challenges which are faced by MT and how they affect the quality of output translation. Here we will discuss the lexical ambiguities, morphological ambiguities, syntactic ambiguities, semantic ambiguities, pragmatic ambiguities.

**4.2.6.1.1 Lexical Ambiguity**

The diversity between languages causes many challenges in machine translation. Among these challenges are the lexical challenges. "languages are differently equipped to express different real world relations, and they certainly do not express all aspects of life with the same equal ease; finding a notional category which is regularly expressed in all languages is difficult" (Ivir, 1981: 56). Lexical ambiguity occurs when a word can be interpreted in more than one. (Hutchins et l., 1992) introduced three sorts of lexical ambiguity: category ambiguity, homographs ambiguity and transfer ambiguity.

A word that has more than one meaning is said to be **lexically ambiguous.** On the other hand, a phrase or sentence that has more than one structure is said to be **structurally ambiguous.**

Lexical differences between languages at the lexical level result in translation problems, some of which will be discussed in this section. Here we will discuss the three main types of the lexical ambiguity and their impact on the generated translation. And their control techniques will be discussed practically in the following chapter.

**4.2.6.1.1.1 Category Ambiguity**

Category ambiguity is the simplest type of lexical ambiguity, here a word can be assigned to two or more syntactic category based on the contextual environment. In English, the word *use* can be a verb or noun based on the context it is used in. In Arabic
the word مهمة (mohemah) can be assigned to two categories, it can be a noun means an assignment or an adjective means important.

4.2.6.1.2 Homograph Ambiguity

Homograph ambiguity is a type of lexical ambiguity arises when a word can have two or more different meanings. In other words, two or more words which share the same spelling but differ in the meanings. They may share the same pronunciation. In English, the word bank can mean “an establishment for keeping money, valuables” or “a land sloping up along side of a river or canal”. In Arabic, the word حمل can mean “give birth to” or “afford or endure something”.

4.2.6.1.3 Transfer Ambiguity

Transfer ambiguity is the third type of lexical ambiguity. Transfer ambiguity arises when one word in the source language can be translated into several words or expressions in the target language expressions. Here the source of translation ambiguity is the ambiguous perspective of the target language. Most of the English words have more than one translation in Arabic. The English adjective old has two meaning depending on the context. It can be translated either into (someone old or elderly) or into (something old or outdated). The following example will explain the transfer ambiguity clearly.

I saw the old woman in the shop.

When a machine translation translates the previous example, it may select the inappropriate Arabic equivalent translation of the adjective old that means (something old or outdated).

4.2.6.1.2 Morphological Challenges

Translating between two morphologically rich languages poses challenges in analysis, transfer and generation. The complex morphology induces an inherent data sparsity problem, and the limitation imposed by the dearth of available parallel corpora is magnified (Habash and Sadat, 2006).

Arabic has richer and more complex inflectional morphology than English inflectional morphology. English inflectional morphology is simpler when compared to Arabic for
example, Nouns in English are inflected for number only (singular, plural) and verbs are only inflected for both number (singular, plural) and tense (present, past, past-participle). But, English Adjectives are not inflected.

On the other hand, Arabic nouns and adjectives are inflected for gender (masculine, feminine), number (singular, dual, plural), state (definite, indefinite). Arabic verbs are inflected also for gender, number and tense (perfective, imperfective, command,), voice (active, passive) and person (first, second, third). The number of unique Arabic words is over 50% more than the number of unique English words [KH10a].

4.2.6.1.3 Syntactic Challenges

Arabic has a different word order sequence that makes it a significant challenge to MT, due to the possibility of expressing a sentence in Arabic in various subject-verb-object combinations with the same meaning. In Arabic, three elements make up a sentence, namely subject, verb and object. Using all of these elements, Arabic can be classified into four types of sentences, according to different word orders i.e., SVO, VSO, VOS, and SOV. Thus, it is a difficult task to find a machine translation that meets human requirements. It is not yet clear whether machine translation can satisfy peoples’ requirements in terms of translation quality and retrieval time. We assume that many kinds of phenomena exist, some of which are suitable for MT.

4.2.6.1.3.1 Word Order

As the verbal sentence (VSO) is the default order in Arabic or more widely used, not taking care of it can produce unnatural nominal sentences. On the other hand, in newspapers headings, book and article titles, the nominal order is the norm and thus Syntactic ambiguity arises because of the different ways of analysing the underlying structure of a given sentence based on the system grammar.

The classical example, *He saw the girl with the telescope*, is ambiguous. It could be the girl who has a telescope or the man has the telescope. Here it is the syntax not the meaning of the words which is unclear. The different readings of this sentence are represented in the following tree in figure 4.8.
The two trees in Figure 4.8a and Figure 4.8b represent the two different analyses in the sense of recording two different ‘parse histories’. In linguistic terms, they correspond to the two readings of the sentence: one in which the PP is the same level as the subject (i.e. the man has the telescope). and the other where the PP is part of the NP (i.e. the girl has the telescope).
4.2.6.2 Non-linguistic Challenges

Currently, the available MT systems concentrate only on linguistic analysis at lexical and morphological levels, and provide insufficient syntactical and semantic analysis to face the challenges of machine translation, and there is no any indication to the cultural aspect of language representation or mention any cultural understanding.

Other non-linguistic aspects affect badly the translation process and generate poor translation. These aspects lie in the historical, cultural, social issues, and sometimes mood. Sometimes the translation conducted by the same person may not satisfy him in other situations or in another time because his mood is changed depending on the time and circumstances.

In the case of cultural issues, some phrases in one language have no equivalent in the other language. For example the religious word تيمم (tayammum) means an Islamic act of using clean sand if there is no clean water, has no equivalent concept in English. This problem is difficult for human translator to solve if he does not have any idea of this term and the problem for machine translation becomes more and more.

The gap between English and Arabic language is very wide and the variations in their cultures socially and religiously increase the difficulty of translation from English to Arabic. Here Example-based machine translation approach with reference to the social and cultural context can deal with these problems and determine the right words to choose.

All the above types of MT challenges will be mostly solved by using the integration of both Head-driven Phrase Structure Grammar as a framework and Example-based Machine translation as a model.

4.2.7 Summary

This chapter provides a comprehensive introduction to the field of natural language processing (NLP), covering the major techniques and theories and an overview of machine translation. Natural Language Processing is a theoretically motivated range of computational techniques for analyzing and representing naturally occurring texts at one or more levels of linguistic analysis for the purpose of achieving human-like language processing for a range of tasks or applications. In this chapter we represented
the different levels of NLP in a sequential manner (phonological, morphological, lexical, syntactic, semantic, discourse and pragmatic). And, the frequent applications of NLP are machine translation, information extraction, information retrieval, etc.

In the second section we introduced an overview of Machine Translation as one of such technologies which has assumed great significance in modern time. It is the use of computers to automate some or all of the process of translating from one language to another. We presented the different processes of MT: analysis, transfer and generation. In this chapter we also presented the architecture and different paradigms of MT. The former refers to the actual processing of translation, whereas the latter refers to the knowledge component that supports the processing field in translation. We got the different approaches of MT including those based on the linguistic constraints, those non-linguistic based approaches and the hybrid approaches. The most important thing presented here is the several challenges faced by MT both the linguistic and non-linguistic challenges as results of languages divergence and ambiguities. The linguistic challenges lie in the lexical, morphological, syntactic and semantic challenges. Other non-linguistic aspects affect badly the translation process and generate poor translation. These aspects lie in the historical, cultural, social issues, and sometimes mood. Sometimes the translation conducted by the same person may not satisfy him in other situations or in another time because his mood is changed depending on the time and circumstances.