Chapter Three
LITERATURE REVIEW

3.0 Arabic Machine Translation

Machine translation (MT), a subfield under Artificial Intelligence, is the application of computers to the task of translating texts from one natural (human) language to another (Okpor, 2014). MT has been in existence since the 1940s and it is as old as computer. Using computers to translate human languages is the first application of computer and code breaking was the other application. MT was the first computer-based application in natural language processing (NLP), and its history is old (Hutchins and Somers, 1992). There is a big progress in the development of MT. It is fired up by the globalization and the need to understand and share ideas with different languages and communities. The prime mover of machine translation advancement and research is the new developments in the field of technology, the increase of multinational companies and governments for translation as a result of globalization. Today, all over the world, soft-ware designers, programmers, hardware engineers, neural-network experts, AI specialists, linguists, and cognitive scientists are enlisted in the effort to teach computers how to port words and ideas from language to language (Zughoul et al, 2005).

Though the development and research in MT has been going on since the 1950s, translation and interpretation still require a real and effective research and development of MT.

As English is a universal language. Most of the researches in Arabic Machine Translation and its implemented systems are mainly concentrated on the translation between English and Arabic. The first English to Arabic machine translation system was proposed in the late seventies by Weidner Communications Inc in Provo, Utah. This system was based on the Direct Method and used a large bilingual dictionary as well as a dictionary for idiomatic expressions. The system consisted of two stages: analysis of the syntax of the source language (English) and generation of the target language (Arabic). The analysis of the syntax of English was not deep enough, it was to the extent required to generate Arabic equivalents.
The aim of this proposed system is to produce fully automated Arabic translations of unlimited English documents and domains. The system was designed for English language as a source language and Arabic language as a target language. It included a module for post-editing but it did not include pre-editing module. The system was commercially utilized by Omnitrans of California Inc for the purpose of translating the Encyclopedia Britannica into Arabic. But, unfortunately, the project was stopped for lack of funding.

Arabic is notorious for its complex morphology (McCarthy 1979; Azmi 1988; Beesley 1998; Ratcliffe 1998; Ibrahim 2002). Hoda M.O. Mokhtar et al. have proposed MT system, which is an automatic system for English-to-Arabic translation of scientific text. Debili (1992) introduced the problem of the automatic alignment of sentences in the bilingual text pairs. Their experiments were applied to French-English and French-Arabic text pairs.

Al-Anzi et al. (1997) have proposed MT system to translate English web pages to Arabic. The system partitions the English sentence into different parts according to HTML tag occurs. Then it translates the part of the English sentence independently of others and inserts the translation between the HTML tags that were present in the source. Its result showed that the system had faced difficulties when an HTML tag appeared inside a sentence AL-Tanni (2005).

There are commercial MT systems. “Al-Mutarjim Al-Arabey” which translates English text into Arabic, “golden Al-Wafi translator” which also translates English text into Arabic [http://www.atasoft.com/products/mutarjim_v2.htm.] and “Sakhr CAT” translator is a computer-aided translation system supporting bidirectional bilingual translation between English and Arabic Mokhtar et al (2000).

Rafea et al (1992) developed an English-Arabic MT system which translates a sentence from the domain of the political news of the Middle East. Lonsdale et al (1994) introduced MT model based on the Interlingua approach to translate technical text from English to French. As they presented in their study, machine translation for scientific text is possible if we investigate the problem in the right way. Chafia and Ali (1995) introduced machine translation system from Arabic to English and French. As they mentioned analyzing and
reordering Arabic constructions is necessary in order to get good results according to Arabic rules. Maalej (1994) introduces English nominal compounds into Arabic MT system. The motivation for developing this system is the frequent occurrence of compound nouns in referring and naming in all text-types.

Barhamtoshy (1995) proposes a translation method for compound verbs. Really, translating Arabic verbal sentence of scientific text to English is exceptional because there is no other similar study so far. Pease et al (1996) developed English to Arabic MT system that translates medical texts. Stalls and Knight (1998) proposed MT model based on statistical approach to translate named and technical terms from Arabic text to English. As they mentioned that the translation of named and technical terms is problematic when the languages involved in translation have different alphabets, such as Japanese/English and Arabic/English.

Mokhtar et al. (2000) developed MT system based on transfer approach using unification based grammar to translate English scientific text to Arabic.

As mentioned earlier, the translation in Arabic language still limited and there are many areas in Arabic machine translation need much research. On the other hand, its results are still not totally satisfactory. Little work has been done in developing Arabic-to-English MT systems. Shaalan (2000) described machine translation system. The system was based on the transfer based machine translation approach for translating the Arabic interrogative sentence into English in agriculture domain. They use the imperative form of the verbal sentence of the integrative sentences.

Chalabi (2001) presented an Arabic-English MT engine that allows any Arabic user to search and navigate through the Internet using the Arabic language.

Larkey et al. (2002) measured the string kernel distance between Romanized Arabic and English word using the spelling feature. As they mentioned According to the above, the machine translation from Arabic to English is problematic and challengeable task and needs special grammar and efficient techniques to generate an effective system. The best and suitable model of machine translation for such challengeable translation is rule-based. Arabic language has a rich and highly complex morphology and syntax. And this means for better translation many processes and a lot of linguistic analysis is required, such as word segmentation, word analysis, etc.
Othman et al. (2003) developed an efficient chart parser that will be used for translating Arabic sentence. The issue of enhancing the quality of machine translations has been gaining interest amongst researchers in recent years.

Attia (2003) introduced English into Arabic machine translation system using a transfer approach. In his study he concentrated on three main points; the analysis of the source language (English), problems related to the transfer of English into Arabic, and the generation of the target language (Arabic). The study had a limit to only electronic texts i.e., texts which are written in a machine readable format.

Shaalan et al. (2004) developed machine translation system using transfer approach to translate English Noun Phrase (NP) into Arabic. They chose Arabic NP in their study because NPs form the majority textual content of the scientific and technical documents.

Arabic has always been a challenge in computational morphology and a difficult testing ground for morphological analysis technologies. When translating from a morphologically rich language, the translation process is passed into multiple steps, which are called tokenization.

Habash and Sadat (2006), Lee (2004) stated that tokenization is helpful when translating Arabic, as Arabic is segmented by simple punctuation tokenization. This tokenization rank is not enough for syntactic analysis (Hatem and Omar, 2010).

Attia (2007) implemented a rule-based tokenizer that handles tokenization as a pre-processing stage in MT. The advantage of this implementation is that it can become more manageable and deterministic in debugging. Its lack of robustness makes it inapplicable, as no single morphological transducer can claim to comprise all language words. Different models of tokenization are applied at different levels of linguistic depth, while the tokenizer interacts with other components. According to Beesley and Karttunen (2003), based on the level of analysis, there are three strategies to develop Arabic morphologies:

1. One-level rules: analyzing Arabic at the stem level and using regular concatenation.
2. Two-level rules: analyzing Arabic words as being composed of roots and patterns in addition to concatenations.
3. Three-level rules: analyzing Arabic words as being composed of roots, templates, and vocalization, besides concatenations.

Lee et al. (2003) introduced a tough word segmentation algorithm, which segments a word into a prefix and suffix stem sequence. They followed a method which is classified by a small manually segmented Arabic corpus and uses it to present an unsupervised algorithm to build a segmenter for Arabic words based on a large unsegmented Arabic corpus. The proposed algorithm can be used latter to identify any number of suffixes and prefixes of a given token. And generally, it can be applied to different language families. The algorithm achieves about 97% segmentation accuracy on a development test corpus containing 28,449 word tokens.

Alsharaf et al. (2004) presented that it is not easy to translate French into French Arabic using existing approaches (i.e., direct, transfer, pivot, statistical, and example). That is because both languages are from different language families and are linguistically different. And to facilitate the translation process between these languages this requires that certain linguistic phenomena that are specific to the pair must be analyzed.

Attia (2005) developed a morphological analyzer, which uses a one level rules-approach analyzer that considers stems as the base forms of Arabic words, and handles spelling differences through alteration rules. Abraham and Salim (2005) analyzed Arabic constructions to English using algorithm based on supervised alignment data. The resulted performance of their algorithm was contrasted with that performance conducted by human annotation.

Hasan et al. (2006) first introduced Arabic to French machine translation system using statistical approach. They applied the proposed system to the medical domain. The main steps of their system were pre-processing (such as Arabic tokenizer), training the models, and generating translation. Which are important to create a system for corpus acquisition.

Abu Shugier and Sembok (2007) asserted that “Arabic differs tremendously in terms of its characters, morphology, and diacritic, from other languages; and to claim otherwise would be a mistake.”
Toutanova et al. (2008) introduced a novel method to improve the quality of Statistical Machine Translation (SMT) by applying models that predicted word forms from their stems using extensive linguistic features (especially, morphological and syntactic information) from both the source and the target languages. They introduced translation system that translated English into two morphologically complex languages, namely Russian and Arabic using an independent model of morphology generation. The model was effective and achieved an accuracy of over 91.

Salem et al. (2008) developed Arabic to English machine translation system called UniArab. The system was based on the Role and Reference Grammar model. They introduce in details the system’s design and how it accommodated the particulars of Arabic to generate English. The system has many limits at the level of its lexicon and the translation process. In other words, the system failed to translate many words, because their structure did not exist in database of the developed system. Another serious problem of the system is its inability to deal with ambiguities.

Attia (2008) mentioned that the traditional classification of Arabic parts-of-speech into nouns, verbs and particles, is not sufficient for a complete computational grammar. This was confirmed by Farghaly and Senellart (2003), Chalabi (2001), Alsalman (2004), and they described and evaluated the Arabic machine translation and pointed out the rules that must be followed in Arabic translation.

Arabic caught attention in the Natural Language Processing (NLP) community, because of its importance at the level of linguistic differences and similarities between it and other languages and its political importance. Arabic linguistic features (specially, its highly rich and complex morphology), make motivating challenges for Arabic language researchers. Valuable work has been done in Arabic natural language processing in different applications such as machine translation (Farghaly and Senellart (2003); Shaalan et al. (2004), entity extraction (Shaalan and Raza (2009), and sentiment analysis Almas and Ahmed (2007). Nowadays Arabic machine translation became the area of interest for many machine translation researchers.

Hatem and Nassar (2008) presented a modified Dijkstra’s shortest path algorithm, and developed a machine translation system that merges between rule-based machine translation system and example-based machine translation system. The proposed system was used to identify the target language phrases by listing the indexes of the
source sentence’s words. Those phrases were found in the target language corpus and constructed a directed graph to identify the phrases that form a shortest path walk in the graph. Their modified method was applied in a hybrid English to Arabic MT system.

Nguyen and Vogel (2008) introduced a context-dependent morphology preprocessing technique for Arabic-English translation. In their study, they used the morphological alignment between Arabic and English to teach a model removing those nonaligned Arabic morphemes. They presented the relation between the size of the reordering window and morphology processing. Their proposed model was only applied on two main systems; a news domain system and a travel-domain system.

Abu Shugier (2009) developed English to Arabic machine translation using a rule-based approach, he concentrated on word agreement and ordering. The main objective of designing this system was that it would be used as a tool and integrated with a general machine translation system. The system achieved a high score that was 96.1.

Badr et al. (2009) developed this English to Arabic machine translation using phrase-based statistical model. They applied syntactic phrase and rules reordering in English-to-Arabic statistical machine translation. They presented the effect of combining reordering with Arabic morphological segmentation; a pre-processing technique to improve Arabic to English and English to Arabic translation. Despite that this study introduced effective results; it had a limited capacity to deal with long distance phenomena, because they relied on local alignments.

Farghaly and Shaalan (2009) highlighted the significance of Arabic and presented many solutions to help and guide current and future practitioners in the field of Arabic natural language processing. Furthermore, the paper introduced solutions that have already been proposed by some pioneering researchers in the field of Arabic natural language processing.

Elming and Habash (2009) introduced both syntactic reordering process and the effect of the alignment method on learning reordering rules within English to Arabic translation tasks. By using such techniques their results were effective as there are important improvements in the quality of machine translation.

The Weidner English to Arabic machine translation system was released by The Sultanate of Oman for the purpose of translating official English documents into
Arabic. However, development of the English to Arabic system stopped shortly after the company was acquired by foreign investors in 1984 (Fargaly, 2010).

Habash (2010) discussed modern standard Arabic. The author focused on Arabic script, phonology, orthography, morphology, syntax and semantics, and machine translation issues about Arabic, such as morphology and Arabic script.

Shirko et al. (2010) developed MT system that translates Arabic noun phrases to English using transfer based machine translation approach. Bisazza and Federico (2010) developed a chunk-based reordering automatically technique to identify and move clause-initial verbs in the Arabic side of a word-aligned parallel corpus. The technique was used to reprocess the training data, and to collect statistics about verb movements. According to the proposed technique before decoding clause-initial verbs their reordering patterns are identified and built on the test sentences. In fact, the proposed technique focused on the syntactic problem of VSO word order.

Carpuatet al. (2010) introduced a new technique for improving overall statistical machine translation quality using a certain syntactic parser to reorder VS sentences into SV for Arabic-to-English word alignment. Though the proposed system surpassed a strong baseline in terms of BLEU systems and presented more globally readable translations, this technique has limits and many problems still unsolved.

Yassine et al. (2010) developed a high performance Arabic NER (automated Named Entity Recognition) system based on the different linguistic features, such as lexical, syntactic, syntagmatic and morphological features. They used an Arabic-English parallel corpus. The system achieved valuable results and noted performance for almost all data-sets that were obtained from broadcast news. Hatem and Omar (2010) proposed a transfer-based approach in Arabic to English MT, in order to solve the word ordering problem. Their approach was tested on 100 titles from the Aljazeera news website.

There are many differences at all levels between Arabic and English, such differences must be taken into consideration by machine translation developers.

Besançon et al. (2009) introduced the In File evaluation paradigm (INformation, FILtering,and Evaluation) in general and focused on a study of the Arabic part of the corpus in par-ticular. They tried to cover mismatch between both profiles and Arabic
documents. In their thesis they also introduced the problems that may arise when transferring information from English and French to Arabic.

Guidere (2002) tried to apply a corpus-based machine translation form that depended on a bilingual corpus of French and Arabic texts and translation part alignment. The author made use of mapping for combining both statistical and linguistic information. And, he proposed specific procedures to build a machine translation system based on bilingual corpora.

Moghrabi (1998) proposed a knowledge-based family of Interlingua machine translation system between French and Arabic in the sub-world of cooking recipes. The author described the design of the generation component and how this design generates a variety of outputs; all these output give the same conceptual meaning. It focused on the importance of the meaning of the text being processed and articulated all of its available knowledge-bases in order to achieve a flexible meaningful wording.

The first work in Arabic-Chinese MT was by Habash and JunHu (2009). They made a comparison between two approaches for Arabic-Chinese machine translation using English as a pivot language. The proposed system introduced many complex Arabic-Chinese syntactic variations. Their outputs proved that using English language as a pivot was a useful procedure than a direct translation from Arabic to Chinese.

Bouillon et al. (2008) proposed an interlingua-based medical speech translation system between Japanese and Arabic and vice versa. They also proposed a simple generic tool for debugging Interlingua translation rules, and a method for developing speech understanding performance by re-scoring N-best speech hypothesis lists. And for increasing their system efficiency they used statistical tuning method.

Doaa et al. (2006) developed a multilingual parallel corpus includes Arabic to Spanish, English to Arabic, and English to Spanish aligned at the sentence level and tagged at the POS level. The results of this method were evaluated against a gold standard system. These results were over 90%, even though the percentages were different from one language pair to another.

The Language Weaver Arabic to English MT system Professors Kevin Knight and Daniel Marcu of the University of Southern California (USC) founded The Language Weaver Inc. in January 2002. The goal of this foundation was to apply their
pioneering research in statistical natural language processing to the commercial objective of producing useful automated machine translation systems. Agiza et al. (2012) introduced an approach for English-to-Arabic translation based on NLP which fills the gap in the field of scientific translation. The transformer approach employed in the proposed system combines some English grammar rules, structure transformational rules and Arabic morphological synthesis rules. The system analyzed, designed and implemented using c#.Net and SQL server. The results are successful Arabic translations from scientific English sentences. The developed system translation has high syntactic and semantic quality. Simple and modular translation is generated that enables future modification and extendibility easily.

Nazlia Omar et al., (2010,2012) developed an Arabic to English Machine Translation system for both noun and verb phrases using transfer-based approaches. In the case of the noun phrases, this system has managed to perform the syntactic reordering for this language pair. The results of this system displays reasonable improvements in translation quality over related approach, the proposed method was tested on 88 thesis titles and journals from the computer science domain. The accuracy of result was 94.6%. On the other hand, in the case of the verb phrase the aim of their study was to introduce Verbal Sentence rule based Machine Translation, The system trained on 45 verbal sentences from different Arabic scientific text and tested on 30 new verbal sentences from different domains. They evaluated their system against two other machine translation systems namely Systran and Google. The accuracy of the result was 93%.

When we compare the existing Arabic Machine Translation Systems to their American, European, and Asian counterparts, Arabic Machine Translation is very little touched and still in its infancy stage. There are some endeavors to produce translation commercial software such as those have been produced by Apptek, ArabTrans, CIMOS and ATA companies.

In 1990, Apptek (Virginia, U.S.A.) proposed an AMT system which led to the English-to-Arabic machine translation system. In this system Transphere used for translating both several, general and specialized domain texts. The used lexicon of this system contains more than 100,000 entries. it runs under UNIX and Windows This system with word processing can be utilized as a tool for computer-assisted learning of Arabic. Moreover,
this system has been applied as a translation engine in large complex products related to air traffic, transportation, communications and control.

CIMOS (Paris, France) developed a software package *Al-Nakeel* to assist human translators several domains and areas. It is intended to translate between different languages. Recently, it translates from English to both Arabic and French and from French to both Arabic and English. One advantage of this proposed system is that each area has its own customized dictionary and translation-memory database. Another advantage of this system is that the system can by using sentence connections and analysis. An important characteristic feature of this developed system is its ability to produce translations at the speed of 20,000 words per hour. *This software* costs U.S. $1000 and runs under MS Windows.

ATA (London, U.K.) proposed *Al-Mutarjim Al-Araby* to assist human translators in a wide range of areas such as Science, Technology, Commerce, Finance, law, Oil industry, Agriculture, Medicine, Military, etc. This system is considered as the first PC-based professional English-Arabic translation system. The proposed system was first introduced in 1995 at the minimum speed of 1000 words per minute. The lexicon of this software contains a 300,000 English word and phrase dictionary. The system also gives alternative meanings for the user to choose the better meaning. It also has a transliteration option for non-dictionary names and proper nouns and performs translation of abbreviations.

This company also produces *Al-Wafy* as a scaled-down version of *Al-Mutarjim Al-Araby*. The price of *Al-Mutarjim Al-Araby* is U.S. $500 and that of *Al-Wafy* is U.S. $50. *Al-Alamiyah* (Riyadh, Saudi Arabia) is Machine Translation software between English and Arabic to generate MT software that would translate general texts in different areas.

In October 1996, it had completed around 70% of the software by building a 10,000 root English lexicon, a Morphological Analyzer and Grammar rules. But its Morphological Analyzer still has many problems that prevented it from introducing in the market.

Despite the existing efforts in the area of AMT but these efforts still very little work when compared to their American, European, and Asian counterparts. Unfortunately,
Arab governmental institutions and Pan Arab institutions are absent in encouraging such a vital area of technology. In fact, Most of the commercial products available now have been proposed by enthusiastic researchers because of their strong will and their realization of the importance of the area in this age of technology.

In conclusion, developing machine translation system from Arabic to other languages (and vice versa) is not an easy work. We note that there are a lot of obstacles and many challenges because of the different features of each language, such as morphosyntactic features, agreement features, as well as, each language has its own challenges and ambiguities. Figure 3.1 summarizes the current methods from the literature review are utilized to develop machine translations from English to Arabic.

![Figure 3.1: Summary of MT (adopted from Alqudsi et al., 2012)]