GRAMMAR-BASED AND EXAMPLE-BASED TRANSLATION TECHNIQUES FOR HANDLING WORD AGREEMENT AND ORDERING FROM ENGLISH TO ARABIC IN MACHINE TRANSLATION

MOUIAD FADEIL MOH`D ALAWNEH

THESIS SUBMITTED IN FULFILMENT OF THE DEGREE OF
DOCTOR OF PHILOSOPHY

FACULTY OF INFORMATION SCIENCE AND TECHNOLOGY
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

PENGENDALIAN PERSETUJUAN DAN SUSUNAN PERKATAAN DARI BAHASA INGGERIS KE BAHASA ARAB MENGGUNAKAN TEKNIK BERASASKAN TATABAHASA DAN BERASASKAN CONTOH DI DALAM TERJEMAHAN MESIN

MOUIAD FADEIL MOH`D ALAWNEH

TESIS YANG DIKEMUKAKAN UNTUK MEMPEROLEH IJAZAH
DOKTOR FALSAFAH

FAKULTI TEKNOLOGI DAN SAINS MAKLUMAT
UNIVERSITI KEBANGSAAN MALAYSIA
BANGI

2013

DECLARATION

I hereby declare that the work in this thesis is my own except for quotations and summaries which have duly acknowledged.

23 Sep2013

MOUIAD FADEIL MOH`D ALAWNEH
P35233

ACKNOWLEDGEMENTS

بسم الله الرحمن الرحيم

Alhamdulillah, all praises to Allah for the strengths and His blessing in completing this thesisForemost, I would like to express my sincere gratitude to my supervisor, Prof. Dr. Tengku Mohammad Tengku Sembok,and Dr. Masnizah Mohamed, for the continuous support of my Ph.D study and research, for his patience, motivation, enthusiasm, and immense knowledge. His guidance helped me in all the time of research and writing of this thesis. My thanks also to Assoc. Prof. Dr. Nazlia Omer.Who have always been approachable and have offered advice whenever I needed it, support and encouragement throughout my Ph.D. study.

Thanks to everyone I've had the pleasure of working with in UKM. Thanks to everyone who has helped me, enabling me to carry out the work contained in this thesis. There are numerous people outside of UKM who have been there to remind methere is life beyond this thesis. To all my friends, who have been around forlonger than I can remember.

I want to thank all my family; especially my parents,my brothers and sisters for giving me the ambitionto take my education this far, for their unconditional emotional support, and for their constant encouragement.

Finally, thanks to my beloved wife and my precious baby daughter Noor. You have been unbelievably patient and have offered me moresupport that I was willing to ask for. Thanks for remindingme to relax, for understanding and for making me laugh. I could not have done thiswithout you.

ABSTRACT

Machine translation (MT) has been defined as the process that utilizes computer software to translate text from one natural language to another. This definition involves accounting for the grammatical structure of each language and using rules, examples and grammars to transfer the grammatical structure of the source language (SL) into the target language (TL). There is very few MT system that satisfies user needs. While the specific purpose MT systems satisfied a high levels of accuracy, but these systems are restricted in specific domains. Poor performance of MT from English to Arabic has motivated this research to build a MT system where word agreement and ordering are important to ensure the generation of sentences in the TL. Mistakes in MT output can either be the result of analysis at the SL level, or due to generation problems at the target TL level. Some of this agreement should be handled between the subject and verb where the number, gender, person and features of the subject are important factors in verb derivation, as well as verb features. Some other agreements are required between the adjective and the noun where the Arabic adjective depends on the number, gender and person of the noun; other agreements are those between the numbers and the countable noun. This study presents English to Arabic approach for translating well-structured English sentences into well-structured Arabic sentences, using a grammar-based and example-translation techniques to handle the problems of ordering and agreement. This technique combines rule-based MT (RBMT) and example-based MT (EBMT) which is called hybrid-based MT (*HERBMT*). The proposed methodology is flexible and scalable. The main advantages of HERBMT are that it combines the advantages of RBMT and EBMT, and it can be applied to other languages with minor modifications. EBMT extracts an example of target language sentences that are analogous to input source language sentences. The extraction of appropriate translated sentences is preceded by an analysis stage for the decomposition of input sentences into appropriate fragments. RBMT is used when examples of the source language to be translated into the target language are not found in the machine database. In addition, RBMT can be applied to other languages with some modifications. The database has been designed to be flexible where most of the rules are defined in the database tables in order to generalize the code. The OAK Parser is used to analyze the input English text to get the part of speech (POS) for each word in the text as a pre-translation process. Validation rules have been applied in both the database design and the programming code in order to ensure the integrity of data. A major design goal of this system is that it will be used as a stand-alone tool, and can be integrated with a general machine translation system for English sentences. The evaluation is carried out on 250 independent test suites, and the analysis indicates that HERBMT achieved good performance with an average of 97.2% precision.

ABSTRAK

Terjemahan mesin merupakan satu proses yang menggunakan perisian komputer untuk menterjemah teks ke bahasa lain. Proses ini mengambilkira struktur bahasa menggunakan peraturan, contoh dan tatabahasa ke atas bahasa sumber kepada bahasa sasaran. Terdapat sedikit sistem terjemahan mesin yang memenuhi kehendak terjemahan pengguna dan ia terhad kepada domain tertentu. Prestasi terjemahan Bahasa Inggeris ke Bahasa Arab yang lemah telah memotivasikan kajian ini untuk membina satu sistem terjemahan mesin yang memfokus kepada pengendalian persetujuan dan susunan perkataan. Kesilapan terjemahan berpunca dari keputusan analisis bahasa sumber atau masalah penjanaan bahasa sasaran. Pengendalian persetujuan perkataan perlu dilihat antara subjek dan kata kerja seperti bilangan, jantina, orang dan fitur lain subjek dalam terbitan serta fitur kata kerja. Pengendalian persetujuan juga diperlukan antara adjektif dan kata nama, di mana adjektif dalam Bahasa Arab bergantung kepada bilangan, jantina dan orang bagi kata nama. Persetujuan lain juga mengambilkira antara nombor dan kata nama yang boleh dibilang. Kajian ini bertujuan menjana terjemahan mesin dari perkataan Bahasa Inggeris ke Bahasa Arab yang berstruktur baik menggunakan teknik berasaskan tatabahasa dan contoh terjemahan dalam mengendalikan masalah persetujuan dan susunan perkataan. Teknik ini menggabungkan terjemahan mesin berasaskan peraturan *(RBMT)* dan terjemahan mesin berasaskan contoh *(EBMT)* yang dipanggil terjemahan mesin berasaskan teknik hibrid *(*HERBMT*)*. Kaedah kajian yang dicadangkan adalah fleksibel dan berskala. Kelebihannya ialah HERBMT menggabungkan kelebihan teknik *RBMT* dan *EBMT*, dan ia boleh diaplikasikan pada bahasa lain. *EBMT* mengekstrak contoh dari bahasa sasaran yang hampir dengan perkataan dar*i* bahasa sumber yang diinput. Analisis ke atas penguraian bahasa sumber ke fragmen yang sesuai dilakukan dahulu sebelum pengekstrakan dibuat. RBMT pula digunakan apabila contoh bahasa sumber yang perlu diterjemah ke bahasa sasaran tidak dijumpai dalam pangkalan data. Teknik RBMT ini juga boleh diguna pada bahasa lain. Pangkalan data direka dengan fleksibel di mana peraturan didefinisikan dalam jadual untuk menjana kod. Penghurai OAK diguna untuk menganalisis input teks berbahasa Inggeris ke *part of speech* (POS) dalam proses pra-terjemahan dan pengesahan peraturan diaplikasikan dalam pangkalan data serta kod atur cara. Ini bertujuan agar data dapat diintegrasikan. Matlamat reka bentuk sistem ialah sistem dapat digunakan secara bersendirian malah dapat diintegrasikan dengan sistem terjemahan mesin umum untuk bahasa Inggeris. Penilaian dibuat ke atas 250 data ujian bebas dan analisis mendapati prestasi teknik HERBMT secara purata mencapai 97.2% kejituan.

CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

MT	Machine Translation
POS	Part Of Speech
OAKP	OAK system is a total English analyzer Parser
SL	Source Language
TL	Target Language
SVO	Subject Verb Object
VSO	Verb Subject Object
DB	Data Base
FAHQMT	Fully Automated High Quality Machine Translation
ALPAC	Automatic Language Processing Advisory Committee
NLP	Natural Language Processing
CEC	Commission of the European Communities
UNL	Universal Network Language
GAT	General Analysis Technique
EBMT	Example-Based Machine Translation
REMT	Rule- Based Machine Translation
SMT	Statistical Machine Translation
KBMT	Knowledge - Based Machine Translation
HREBMT	Hybrid (Example &Rule-Based) Machine Translation
RTL	Right to Left
LTR	Left to Right
GAT	General Analysis Technique
F	Feminine
M	Masculine
S	Singular
D	Doula
P	Plural
ASCII	American Standard Code for Information Interchange
OEP	Overall Errors Percentages

CHAPTER I

INTRODUCTION

1.1 INTRODUCTION

The Machine Translation (henceforth referred to as MT), which is also known as the Automatic Translation (AT) employs the application of computers to perform translation of texts from one natural language into another language. MT utilizes computer software to translate a text from one natural language into another language (Systran 2007).

There are two approaches in the translation process using MT; one of the approaches is Rule-based, the other one is Example-based approach. Rule-based approach encompasses several methods; direct transfer method, transfer-based method, and Interlingua method. This approach focuses on the structural aspects of the language. The emphasis is on producing correct grammatical translations from the source language to the target language. The main setback of this approach it requires much time in the development phase and fine tuning the rules for the translation process. Example-based MT is data-driven technique. This approach is based on non-structural translation whereby grammar rules are not explicit. The translation process is based on a limited bilingual corpus. Compared to the rule-based approach, this method saves time as it relies primarily on word-for-word translation.

Hence, the aim of this thesis is to strike a balance between both approaches in the use of MT for the translation of texts. The writer is of the opinion that it can be achieved by making inferences of transfer rules automatically from bilingual texts, targeting texts with little or limited bilingual data. The rules are obtained from various sources, such as parses that are available for one of the languages, or morphological

processes that are available for the source language (SL) and the target language (TL). The term, Machine Translation has been used by the European Association, for it is (one of the very earliest pursuits in computer science). MT has proven to be an obscure target, but currently a number of systems are available and are capable of producing output, even with some imperfections, as well as possessing adequate quality to be useful in a number of specific sectors. According to Arnold et al. (1994) in performing the whole or part of the translating process from one human language to another, the flow of the process can be in several directions. Unidirectional translation occurs in one direction as in the case of English into Arabic, bi-directional translation occurs in both directions as from English into Arabic and from Arabic into English or even multi-directional translation that takes place in both directions, involving more than two languages or language pairs.

The multi-lingual challenge is one of the reasons that many recent approaches to MT have opted to learn translation information automatically from bilingual corpora (i.e. text that is given in parallel in two languages), most commonly using one approach such as; example-based, rule-based or statistical methods that derive models of translation. Rule-based and example-based approaches focus on analyzing the sequence of words and their translation. Moreover, they have the advantage of being derivable from bilingual text, thus overcoming the development bottleneck. However more recently, the community has noted once again the potential beneficial impact of structural and feature information on translation quality.

Many recent approaches in MT are in favors of translation of texts from bilingual corpora due to multi-lingual challenges. In bilingual corpora, the texts are provided in two languages; hence the preferred translation approaches are rule-based and example-based. Both these approaches focus on analyzing the sequence of words and their translation with emphasis on the application of rules and word-for-word translation which can speed up the translation process of bilingual texts. It is also worth noting that the focus on structural and word features can improve the quality of translation work.

The MT process consists of several steps. Before the translation process takes place, the words in SL are translated into TL and are stored as examples. Also the syntactic structures of both SL and TL are stored as input. The other step is the matching of the structural systems which requires the analysis of the input source-language sentence to identify its syntactic structure. It is followed by the search of the translation-examples with the equivalent on the syntactic level. In non-structural systems, the translation-examples are stored as pairs of strings, with additional information at word level. Usually in the morphological analysis, the words in the input text are identified initially according to the word class or word form. For instance, the word "*indices"* is identified as the plural form of *index* and the word "*ate"* is identified as the past tense form of the verb "*eat"*. After the initial stage, based on the morphological stored information, parsing of the lexical structures follows with the identification of Noun Phrase, Verb Phrase, and other phrases of a sentence and the understanding to clearly define where one phrase ends and the other begins. Besides identifying and parsing, the translation process must also consider analyzing the sentences during the composition and order organization of the sentence, and rules are given for how transfers into another language are. "Knowledgeable" enough to determine the grammatical functions of, "*who is doing what to whom at where and when*". This is necessary as it has to select suitable and appropriate words from a huge number of choices to convey the meaning in the input text.

The process of generation of the target language will only start when the source language is completely analyzed, matched and understood. This is followed by the representation process at the lexical, syntactic, and semantic levels which are eventually transferred to the target language. Upon transference of meaning, rigorous rules in the target language are applied; in which words are to be situated in the correct sequence and to be grammatically appropriate in the target language. Agreement on the entire structural string and important features of the words plays a very significant role in the construction of the target sentence or phrase that is coherent and meaningful.

1.2 PROBLEM STATEMENT

Generation process of sentences in the target language must conform to the specific requirements of the target language's grammatical rules, ensuring sense and meaning in the output dictates fulfillment of ordering features Attia (2002).To enhance performance at the output in MT, strict adherence is inevitable to avoid mistakes in the translation process, both in analytical phase at the source language and/or generation phase at the target language.

The crucial criterion of word ordering rules for the target language complement its sole importance for reordering of sentences constituents. The advantage of the rule-based approach is clear for Azza (2000) to construct accurate sentences for the target language a grammatical knowledge is of vital importance, such syntactic knowledge are contained in these rules. Machine translation between languages with rich morphological variations and large differences results in more complicated agreement rules and agreement requirements. For example, when translating from English to Arabic, which is a language that is dominantly sensitive to agreement features, especially in achieving grammatical accuracy, MT system faces a dilemma when dealing with the information about gender in the present tense English verb because it does not indicate whether the subject is singular or plural whereas in other tenses, the verb is completely neutral English verbs and determiners are also, to a great extent, neutral to number and/or gender. Moreover, Arabic language's adjectives, verbs, and most determiners are highly reactive to the nouns they modify, whether singular, dual, plural, feminine, masculine, human or non-human, alive or non-alive as well.

(Yumout & Karim 2006) and Rayner (2000) developed translation program using one method only which is Example-Based as a 'mono' approach like (TARJIM) MT system. (Shaalan et al., 2010) and (Shquier & Sembok 2008) and Attia (2002) also Azza (2000) proposed another 'mono' approach using Rule–Based method instead, as in (ALKAFI) MT system.

One of the attractive features in MT system is the ability to extract useful information in certain cases from the translation source language to positively influence the accuracy in the agreement for the target language as an output. To enhance performance of the MT system when handling lexical items, special attention must be drawn to the specific determination of the proper nouns and titles in the source language, namely, gender and number. For example, *Halim* حليم in the source language is masculine, while *Halima* حليمةis feminine. Some cases, indicates that the information of the lexical items required for the agreement purposes can be provided by the target language, especially in the Arabic language, where the gender of common nouns is clearly defined; masculine or feminine, singular, dual or plural. For example, the target language imparts that باب [*bab*], which is the equivalent for *door*, is masculine while عتبه [*ataba*], the equivalent for *doorstep*, is feminine. When the information needed for agreement cannot be extracted neither from the source language nor from the target language, unfortunate problem arises, where; the deciding factor in such case becomes the context. Fortunately, the MT system inherently comprises reverse-tracking in the process (going forth and backward) looking for the information it needs. For example, in the sentence '*The teacher fetches his colleague to the school'*, here it is not clear whether the teacher is masculine or feminine until the system moves forward and identifies the pronoun *his* to establish the gender of the teacher is masculine. Nevertheless, the gender of the friend still cannot be determined from the sentence. Then the system has to look for clues in the previous or following sentences, if any; otherwise it will take the default meaning of masculine.

1.3 OBJECTIVES OF RESEARCH

The MT process involves several complications when handling agreement and reordering features, however when looking into solutions to overpower the status problem arising as a consequence represent one of the primary objective of this research study. The objectives of this research are:

- To form a method that empower pragmatic means and methods through employing grammatical and lexical transformation process that has to

conclusively set the rules of the agreement and reordering process in influencing the construction of consistent Arabic structures in the MT output.

- To design a hybrid platform that accommodates the rule architecture and translation-examples to handle the problem of word agreement and ordering in the translation of sentences from English to Arabic.

- To adopt synergism in implementing hybrid-based Machine Translation system prototype that handles the word agreement and ordering problems.
- To investigate means where the algorithm can automatically learn transfer rules for Machine Translation from bilingual text. Via targeting difficult scenario where rules are learned from; 1) a very small corpus and 2) for disproportionate language pair where a parser is available for only one of the information necessary for agreement and reordering in the course of translating from a language with little morphological variations, such as English language into a morphologically rich language, such as Arabic language.

- Mobilization of the computer knowledge acquisition feature and enhancing Machine Translation system to acquire information from the input texts.

Finally this research aims to utilize this system solely as a stand-alone tool that can be very well integrated with a general Machine Translation system for the translation of English language into Arabic language, and vice versa.

1.4 THE SCOPE OF WORKS

This research study is based on two languages; namely English language as the source language and Arabic language as the target language. The texts in English language are analyzed and translated into Arabic language. In the process of translation, all problems related to the transference of English language into Arabic language, including the construction of Arabic structures will be fully discussed. However,

aspects that are related to the analysis of Arabic language as a source language is out of the scope of this research study.

There are three basic operational strategies in MT research, namely; 1) direct, 2) transfer and 3) interlingua. For the purpose of this study, the main focus is on the analysis of MT based on transfer strategy with example-based strategy as the other choice. The reason for this choice over the other two strategies is its wide usage. The transfer strategy or approach has made its presence to be both theoretically powerful and practically available for MT application where most MT systems that are currently available in the market have been designed according to the transfer strategy system (Trujillo 1999).

As for the other two strategies or approaches, their uses are rather confining. The direct strategy is both theoretically and practically insufficient to meet the needs of MT development, as it has no conceivable theoretical background. Meanwhile, the interlingua approach is highly and academically hypothetical, so its use is only confined to a few systems that are not available for widespread commercial application as yet.

The basic idea for adopting hybrid techniques in machine translation is to emulate human translation in some cases, Nagao (1984), such course becomes fairly common techniques, where transfer-based combined with the example based to form a paradigm for natural language processing (NLP) applications. A major difference between our work in this thesis and the approaches described is the ability to learn transfer rules from extremely small corpora. Statistical machine translation SMT is inherently large bilingual corpora dependent. We show that carefully designed composition of much smaller corpora can be utilized so the transfer rules can be learned from. Combined strategy where transfer and example based techniques is adopted to form this paradigm gain popularity for natural language processing (NLP) applications. Thus, initiates a shift towards quasi human like thinking strategy during the translation process.

Our intention in this thesis is to show that the transfer rules are obtained from very small bilingual corpora and can be learned from corpora that are detailed and carefully designed in composition despite its small size.

Finally the choice of texts for translation purposes is given consideration for the purpose of the study. An only electronic text which is written in machine-readable format is included, while voice or paper document inputs to a computer will not be discussed.

1.5 RESEARCH METHODOLOGY

In order to achieve the goal of the research in a practical essence but pragmatically the MT system of transfer from English into Arabic language will be exclusively discussed, references to the English and to the Arabic MT systems are represented by SYSTRAN, GOOGLE, and TARJIM, while other references MT systems are represented by ATA Software AL-KAFI.

The MT system is used only as the basis of testing ground for the salient points introduced in this thesis, therefore, the research does not involve in the evaluation of any MT system.

Language features advantages in example –based strategy have been indicated. The flexibility of language features such as; agreement and word ordering with ability to effectively contend with the MT system requirements. To support these features several examples from the MT system were presented here to fortify the issues discussed. It is also to be noticed that the capability of MT systems in generating logically consistent and correct agreement and reordering of words; nevertheless, in some cases, this task will cease to perform in successful manner. Therefore, we will show both cases in the methodology, namely, successful performance of agreement and reordering features, failure to perform these features by the system due to some inabilities will also be illustrated.

The process of MT system starts with initial identification phase where the system perform matching analysis of the source text, the output of this step will results in a phrase structure of the sentence and the part of speech POS for each word in the sentence. Then this output is passed through the lexicon and rules which are employed to make the production of translated output possible. OAK Parser software is used to run the whole process. Adopting the rule-based and example-based strategies as a hybrid approach in the MT system gives the advantage of flexibility of the design that can be expanded to cover all words/phrases and possible patterns of sentences as well as the addition of new words to the lexicon and the addition of new patterns to be included in the file of grammars.

Otherwise, the rules which are necessary to handle the derivations of the suitably correct meanings for verbs and adjectives that depends on gender, number and person of the subject and object, will be to a feasible extent contained in these tables. As a matter of fact, these tables in the database will have the flexibility to accommodate any additional new rule(s).

1.6 RESEARCH STRUCTURE

- **Step I -Theoretical Study**

This stage Figure 1.1 involves the current state of art research filed in the machine translation research and the applications that can assist in this field such as English to Arabic machine translation with GOOGLE (Google on-line translation page http://translate.google.com.my/?hl=en&tab=wT, 2010-2012). ALKAFI (Alkafi on-line translation website http://www.filecrop.com/al-kafi-translator.html, 2010-2012) SYSTRAN (Systran on-line translation website http://www.systranet.com/translate/2010-2012) and TARJIM (Tarjim online translation website http://translate.sakhr.com/sakhr/MainView.aspx?lang=1, 2010-2012) these systems have been studied and analysis for the features of agreement and ordering. The process of analysis is based on the translation of the target language without investigating their technique used or how do they work.

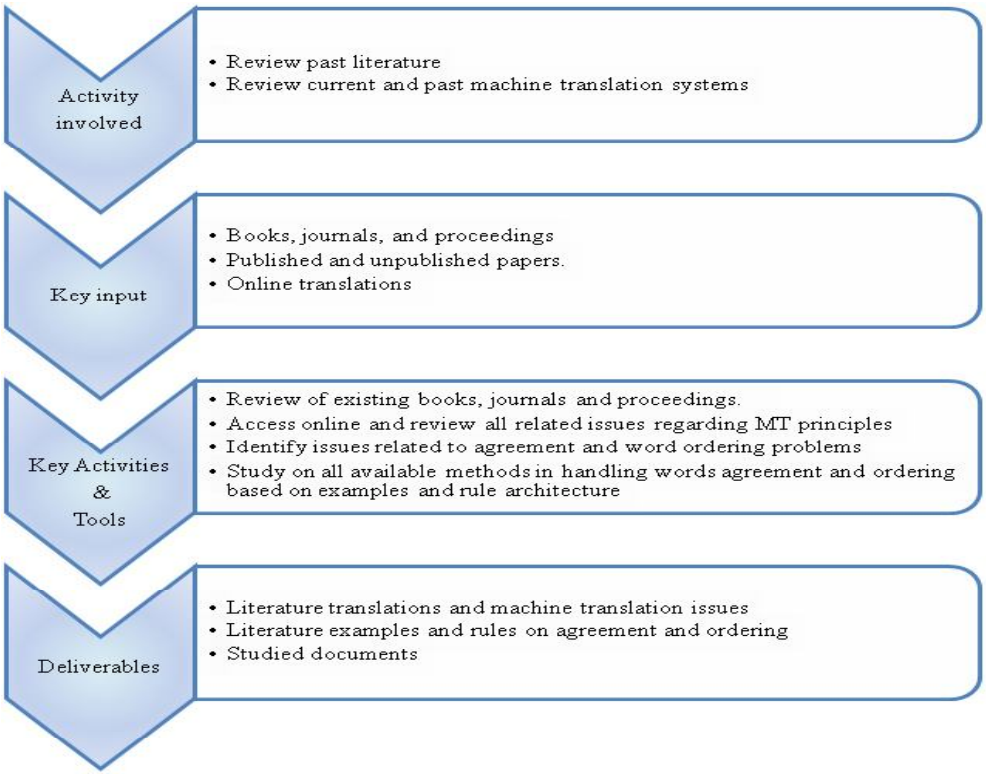

Figure 1.1 Theoretical Study

- **Step II -Framework Development**

In step II Figure 1.2. Theoretical study of the English to Arabic machine translation research area and components involved in MT system is used as the basis for the development of our system " EA-HREBMT" analysis transfer and generation processes with the example and rule-based architecture are studied and developed then depicted in the conceptual framework of " EA-HREBMT" .

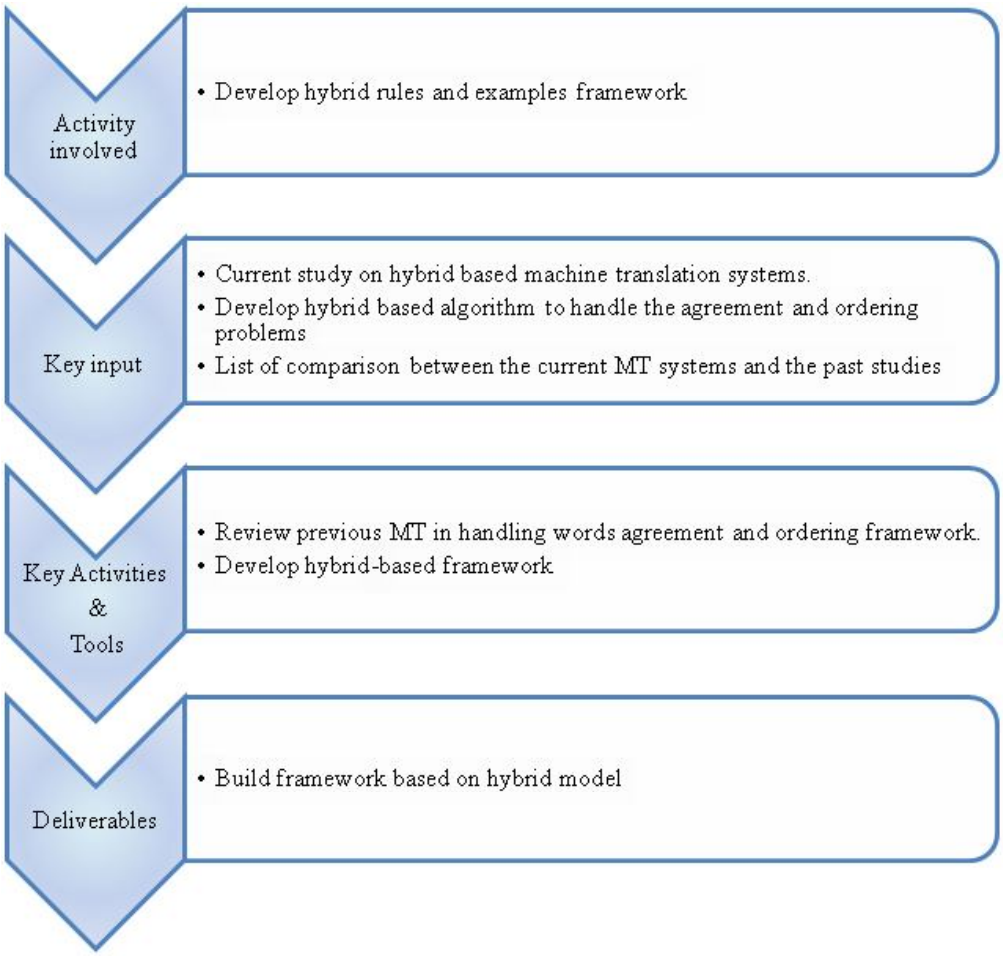

Figure 1.2 Framework Development

- **Step III -Prototype Implementation**

In step III Figure 1.3 a prototype is developed based on the framework for the hybrid based Architecture. The prototype has been designed to handle the word agreement and ordering with machine translation from English to Arabic.

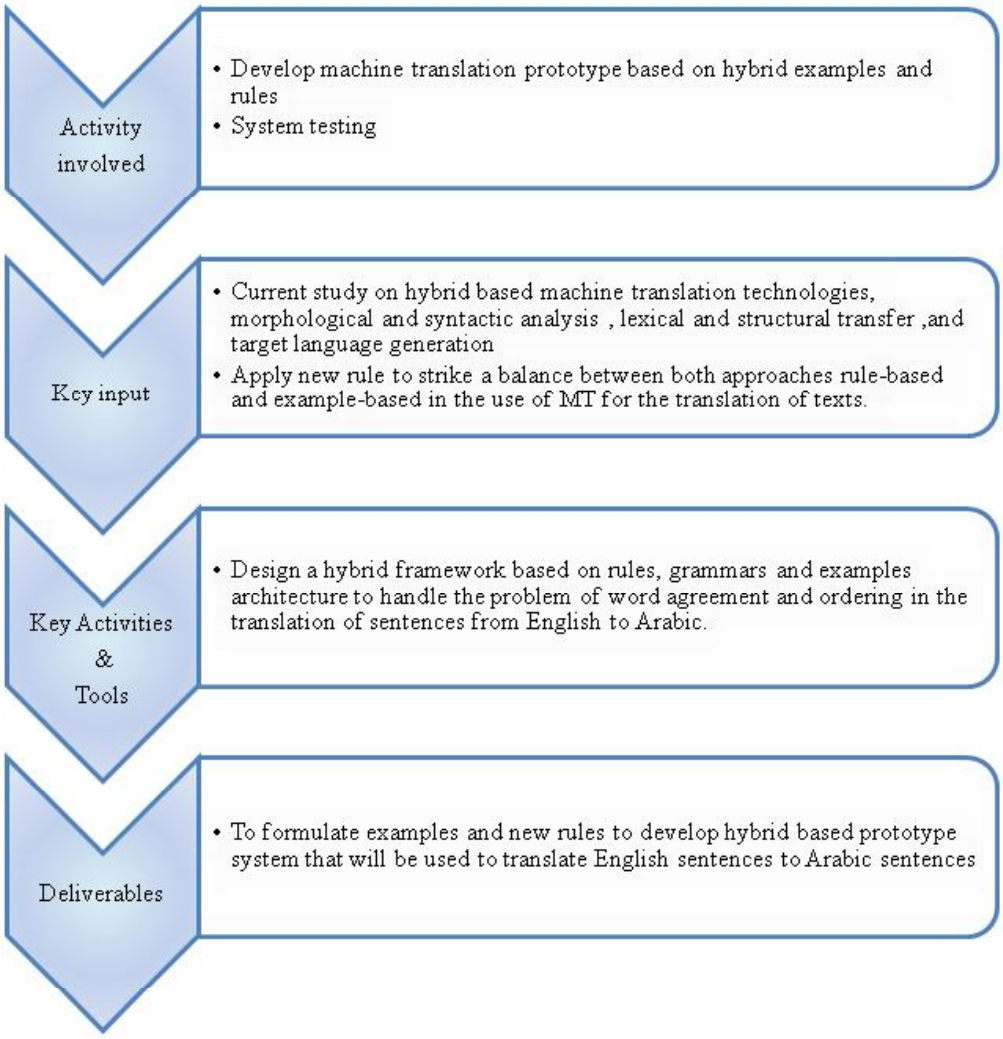

Figure 1.3 Prototype Implementation

- **Step IV –Experiment**

In step IV Figure 1.4 Two experiments have been conducted will be discussed in chapter VII . the experiments is conducted using a test suite consisting of 250 test examples to prove that the selected technique and the system developed are suitable in handling the word agreement and ordering in this steps we have classified the type of error in MT system .

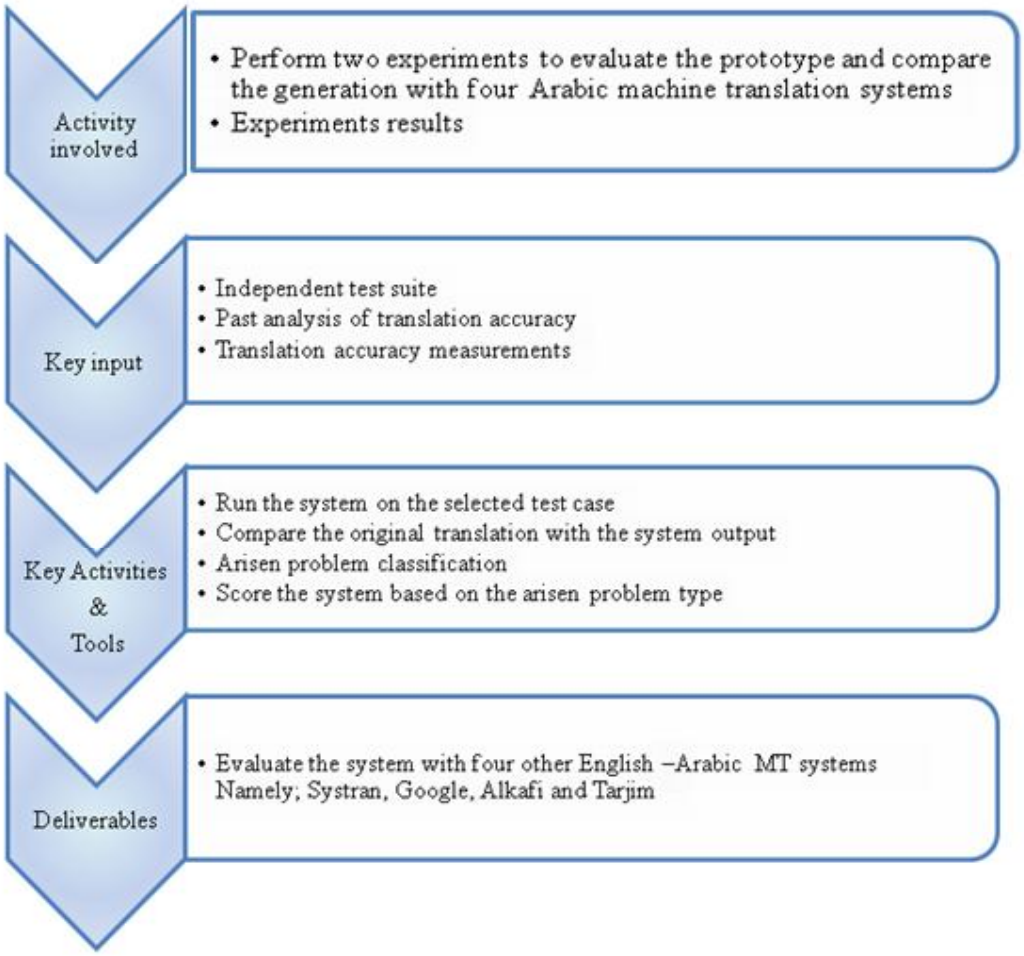

Figure 1.4 Experiment

1.7 ORGANIZATION OF THE THESIS

The outlines of the remaining chapters of the thesis is organized into seven chapters where the literature review marks the starting while conclusions, work contributions and suggested future works designate the ending of this thesis. The other chapters covers languages features and elements pertinent to MT, their problems and handling techniques, then the proposed hybrid approach with the design, methodology, and evaluation . The chapters are presented as follows:

Chapter II is a concise collection of several research areas of machine translation (MT), spanning various eras from first attempts and the following

development activities and achievements into the present stage, shading some light to the ongoing activities and the current primary focus area. Brief historical background of machine translation MT, relevant to the scope of our research in this thesis, emphases on Arabic MT history and English to Arabic translation MT system was presented. A close examination of the hybridization approach in machine translation MT with clear explanation to the choice of rule based strategy to combine with example-based on MT. Dwelling and explaining hybridization necessitates the introduction of various approaches and strategies in machine translation systems.

Chapter III, the agreement and word ordering problems were addressed in both English language and Arabic language. Agreement is a basic property of language. Whereas languages are varied in the agreement requirements, some of them like Arabic require number, gender, person, and case agreements while others need some of these agreements. These two features greatly affect the output of MT, due to the fact that text orientation and sentences compositions depend on language. Various areas in what is referred to as grammatical phenomenon which are anticipated to cause problems during translation were explored.

Chapter IV, the problem of agreement and word reordering were presented together with proposed solution to handle each problem, and for this purpose, a collection of specially constructed test sentences designated as 'test suit' was presented to explore this problems in the translation process from English into Arabic. To ensure achieving enhanced results specific modifiable rules were suggested to evaluate performance of the hybrid MT system.

Chapter V is to elucidate the logical phases of hybrid machine translation process. With a special focus on the requirements needed and procedures involved in the process. It will explain how the source language text is analyzed as an input by detecting constituent structures and resolving lexical and syntactic ambiguities and processed until at the end rendering the target language in a grammatically acceptable form generated as an output. Analysis occurs at different hierarchical levels. To link both steps, transfer stage starts with the output of the analysis phase by applying bilingual rules to the representations which result from analysis phase and ends where

the phase of generation starts. Generation comprises of two components; syntactic generation and morphological generation. It involves only target language information and operates independently of the source text.

Chapter VI presents the framework of the hybrid-based English to Arabic MT system based on the example-based and rule**-**based. Three main components were employed to combine example-based with rule-based methods, namely; analyzer, transfer, and generation components. To further indicates the strength of the developed hybrid method in this thesis, an example was selected with three cases were part of the sentence cannot be handled by a mono approach like the example-based adopted by well-known existed machine translation system.

Chapter VII describes the evaluation methodology for assessing machine translation system. The discipline applied to three experiments with the analyses of the results. The basic evaluation concept simply assesses the system output in comparison with already existed reference translation or original translation. A proper categorization to the arisen mismatch problems was defined, and then a suitable scoring will be assigned by human for each problem.

Finally, in Chapter VIII, we present the conclusions of our work, the major contributions achieved in this research, and suggestions of number of avenues for possible future research.

CHAPTER II

LITERATURE REVIEW

2.1 INTRODUCTION

This chapter provides the tools to grasp the required knowledge in various areas of this research. A thorough characterization of machine translation (MT) functions, a concise history of MT and historical background of Arabic MT as a field of study, sweeping its early stage through various phases of development into the present stage with a rigorously attentive examination of the conventional approaches of MT systems with special attention to clarify the choice of transfer based, also on why combining with examples based on MT. Due to the importance of these research areas, different related fields have been studied for clearer understanding and better insight to the problem under research.

2.2 THE ARDUOUS TASK OF MACHINE TRANSLATION

Machine translation is an active field of research with many competing paradigms to manage core translation problems. In recent years, an important topic for research has been investigating how hybrid machine translation engines as well as integration systems including some translation engines can be designed and implemented so that the resulting translations provide an improvement over the component parts (Melerob et al. 2012). Machine translation (MT) is not only an computer application to translate one natural language to another natural language automatically and without human involvement except for the part of preparing the electronic format of the source language as an input to the system Therefore, but also it is an automation process, in the process the computer is programmed 'taught' to acquire the required knowledge that human translators require to execute their work such as, the most suitable procedures and routines to successfully accomplish the translation process. In order to understand the complications and requirements, we must excavate the roots of the

problem; therefore, translation assignment can be efficiently accomplished if the human translators have possession of the consummate skill of the following distinctive knowledge according to Eynde (1993):

a. Cognition to comprehend the meaning of the source language text through Linguistics terms and components (lexicon, morphology, syntax, and semantics).
b. Cognition to produce intelligible, adequate translation to satisfy the required standard, and constructed in accordance with the syntactic rules of a particular system 'grammatically correct text' as per target language grammatical usage (lexicon, morphology, syntax, and semantics).
c. Knowledge of establishing the relationship between source and target language to efficiently perform the transference step of lexical items and syntactic structures of the source language to the closest counterpart in the target language
d. The knowledge to understand the specific and contextual usage of the subject matter terminology in the translation process.

2.3 HISTORY OF MACHINE TRANSLATION

Machine Translation history can be traced back to the 17th century (Hutchins 1995). The first suggestion is to use mechanical dictionaries to solve the language barriers and this represented the early machine translation system, yet a concrete proposal for machine translation was not introduced until a patent issued. In the second quarter of the last century, two precursors are identified through their works pertaining to our subject MT. Two patents were issued in (1933), the first one dated to (22nd July 1933 to Georges Artsrouni) in Paris, while the other dated to (5th September 1933 to Petr Petrovič Trojanskij)in Moscow. They have essentially made a reference in their two separately issued patents to construct mechanical multilingual dictionaries. Artsrouni invented a general purpose mechanical storage device called a 'mechanical brain', the storage device contains facilities for retrieving and printing stored information with several potential applications such as automatic production of paper documents which are used in many industrial and commercial sectors like ; railway timetables, telephone directories, commercial telegraph codes, banking statements, and even of

anthropometric records, automatic production of railway timetables, of telephone directories, of commercial telegraph codes, of banking statements, anthropometric records. It was claimed to be particularly suitable for cryptography, for deciphering and encrypting messages, and finally it was claimed to be a device for translating languages (Hutchins 2000).

It is worth mentioning that Trojanskij's concept of automated translation process (mechanized at his time) coincides to a great extent with the current philosophy of MT, the three stages he described (analysis, transfer, and synthesis), with recognition of some basic problems of translation before many years of the MT pioneers in the (1950's. In addition to that, Trojanskij) stated clearly many major advantages of mechanization (in particular with multilingual output).

2.3.1 Inceptive Idea

In order to understand this research study in more clear way, it is useful to consider some background information on relevant materials to MT research portraying the development course from beginning to the current time. World War II witnessed the emergence of the first developed ideas of automated translation. Deciphering coded messages in the language of the counterparts 'enemies' requires human decoders with laborious efforts and a very slow output in their premonition of the encrypted calamitous results.

Taking into account the imperativeness of situations and conditions, a decoding machine was invented. This created the motive to further develop the idea of the language translation automation using a machine with appropriate and scientific research work. This idea was suggested for research through a correspondence written by Warren Weaver in 1949, he argued that human languages can be translated automatically using a computer without reliance on human translators. This brought about the wave of interests and the subject matters based on his theories began to elucidate.

Many critics in the linguistic society have opposed the argument presented by Weaver's analogy of translation and cryptography. It is clearly true that both disciplines requires subtleness and perceptiveness, however, and due to the fact that coding is a 'one-for-one permutation and uses different symbols for the words of the same language, whereas translation is a far more complex discipline. Despite these facts, several research projects and programs have been initiated in the United States, Russia, and other countries to pursue the research idea of translation automation.

2.3.2 Sanguine and Auspices

Chronologically, optimism and high expectations era have emerged with the sudden and strong rush of interest in the MT. However, such lively eagerness within scientific researchers of MT started to confront some basic problems in MT. Machine translation works requires more developed approaches and strategies involving grammatical rules employment to analyze source texts and this represent one of the preeminent dismaying problems which face researchers were only available a restricted and simple dictionary-based systems that could not support translation works. The problem is not only isolated as the lack of technical facilities or hardware, but also is identified as of a linguistic nature. These problems slacken investigative and fact-finding research work which resulted in the shortfall to implement an operational system within a realistic time frame.

This dilemma begun in the early 1960's to severely mount in intensity that caused some scientists and funding agencies to believe that the realizable practical MT system is not feasible. One of the participating scientists in MT research and development since its early beginnings, (Bar-Hillel, published a report in 1959) entitled 'Report on the state of machine translation in the United States and Great Britain', where he criticized the methodology and goals pursued by MT research groups at that time and consequently strongly argued against the development of the MT describing the problem of developing fully automatic high quality MT (FAHQMT) was an unattainable goal in principle, not just at the present moment (Arnold et al. 1994). He also suggested that MT should only focus on moderate translation that involved human interaction.

After the publication of Bar-Hillel's report a lot of uncertainness, precariousness, and ambiguity began to surface. The report seriously impaired and weaken both MT researchers and the public's perception of MT. The impact of this vehement view that emerged to describe that MT was unachievable, impossible and unattainable task began to rise and most scientists and people considered this report as proof of impossibility of MT. This gloomy state of affairs prevailing MT during this period resulted in that the book publishers argued that any effort to mechanize human thinking processes, for instance translation, is destined to inevitable failure (Hutchins 1986), and (Mortimer Taube 1961). In an unprecedented step to rejuvenate the hopes of this research discipline the National Academy of Sciences formed the Automatic Language Processing Advisory Committee (ALPAC) in (1964) to report on the current status of MT research and to advice on its feasibility.

This action, in addition to the request from the funding agencies to continue the exploratory research activities in MT has revived from its hibernation this discipline but just for a short period of time, and after two years the committee have investigated several issues; the existing demand, supply, costs of translation, the availability of translators, and the cost and output quality of MT, the committee issued a report, known as the (ALPAC) report in (1966) with specific negative conclusion on the MT development, the report described MT process was slow, less accurate and twice as expensive as human translation, (Hutchins 1992). They also announced that there was no immediate possibility of future success for MT and as such, there was no necessity for further research and investment in this field. It recommended that researchers were to alternatively focus on the development of machine aids for translators.

The MT supporters both individuals and scholars have animadvert stated different perspective where they considered that (APLAC) statement as unfairly biased and argued that improvements to MT systems were achievable. They stressed that the conclusion in (ALPAC) report about the failure of MT, at the most, was hastily made and require the essence of a sound scientific fundamental principle.

2.3.3 Invigoration

Several factors have contributed to the revival, invigoration and progress of Machine Translation which helped increase the pace of development; the researchers' perseverance and contention, the worldwide new technological development which started in the 1980s, information technology rapid advancement, a swift fall in the cost of computing power, cross interaction effects of globalized market and increasing demands from multinational companies and governments to meet the rising needs of translation, and in addition to funding the MT dilemma was rightfully brought strongly to the forefront.

Since the 1980s commercial markets demands and its increasing expansion has prompted several new operational MT systems to appear, however, their need for further development and improvement is inevitable. Some of these systems are; The 1960s developed Georgetown system, the French multilingual system; (TITUS) system, the Chinese English (CULT) system; the Spanish-English system; (SPANAM), that was initiated by the Pan American Health Organization and the tailor made systems developed by the New York based Smart Corporation. Also, the US Air Force and the European Community adopted the use of the Systran Russian-English system and the System of Logos Corporation respectively. The Commission of the European Communities (CEC) in Europe supported the English-French version of the Systran. The French German System (ASCOF) and (SEMSYN) for the translation of Japanese scientific articles into German were developed in Germany known as (SUSY) (Saarbrucker Ubersetzungssystem). The (EUROTRA) project is considered an ambitious and reputable system, which has been successfully developed by the European Communities in this era. This project has become the basis of the development of a multilingual transfer system for translating between all the Community languages. In the 1980s, Japan dominated the commercial activity in operational MT systems for most computer companies developed software's for computer aided translation that was mainly targeted for the Japanese-English-Japanese markets, Hutchins (1995).

The swift reduction in the cost of computing power, increasing demands from governments and multinational corporations and the increasing stretch of globalization have created new factors in the 1980s which ultimately resulted in the competitive resurgence and revival of research and interest in MT. The Interlingua systems have shifted away from the syntax of different languages, the major contribution to the advancement of this approach is greatly due to the introduction of translation system where the source text is translated into an intermediate language or symbolic representation from which it could be translated into any other languages. Hybrid architectures intend to combine the advantages of the individual paradigms to achieve a better overall translation. (Hunsicker et al. 2011) have shown that using a substitution based approach can improve the translation quality of a baseline RBMT system.

2.3.4 Impressive Enhancements and Inspirations

At the turn of the new millennium an unforeseen development and significant improvements to the MT systems, where personal computers and Internet started to enforce its presence in daily applications, compelling speed and performance enhancements to the MT systems. However, and after more than two decades the desired vision for a fully automatic high quality machine translation (FAHQMT) is still a target to be realized. Nonetheless, useful and interesting outputs continues to provide the users of MT systems a great chance to capture the essence of the translated materials to understand or to prepare a draft for post-editing to get more meaningfully accurate translation. Hybrid MT is a recent trend (e.g. Federmann et al., 2009; Chen et al., 2009) for leveraging the quality of MT. Based on the observation that different MT systems often have complementary strengths and weaknesses, different methods for hybridization are investigated that aim to "fuse" an improved translation out of the good parts of several translation candidates.

(Huyssteen et al. 2009) presented a bidirectional rule-based machine translation between Dutch and Afrikaans, two closely-related Germanic languages. The system gives promising results, and offers an improvement in translation quality in the Dutch to Afrikaans direction over another publicly available system, but does not offer any

improvement in translation quality in the Afrikaans to Dutch direction. In the institute of Advanced Studies of the Tokyo-based United Nations University, extensive exploratory investigations have been initiated to design a Universal Networking Language (UNL). This (UNL) is of great interest because it resembles the transfer language which requires 'editors' to convert text from one language into UNL and then to another language, Guessoum and Zantout (2001).

The developers of this new language had high hopes that the Internet users will learn this translation system which would convert users' language text to (UNL), then display the converted version to the users who would then reconvert the (UNL) version to best describe the intentions of the users. After that, The (UNL) version would then be displayed on the Internet and subjected to conversion to any other language by needs of who browsed the (UNL) version. One company (SAKHAR صخر) is involved in the Arabic language part in a big and complex language translation project, the company intends to incorporate deeper and more abstract levels of representations, including the discourse of structure and interpersonal pragmatic activities that are included in the transfer of related structures.(Harshawardhan, et.al 2011). Presented a novel framework for English to Tamil translation system. This is a phrase based translation using translation memory and concept labeling. The given input text are labeled and converted into phrases. These phrases are searched in the parallel corpus and the related phrases are extracted from it. Among the related phrases the best one is chosen as the target output sentence.

2.4 HISTORY OF ARABIC MACHINE TRANSLATION

In recent years, machine translation (MT) systems have achieved increasingly better translation quality. Through Hybrid architectures intend to combine the advantages of the individual paradigms to achieve an overall better translation (Sabine et al. 2012). Arabic language attracts a lot of attention from the major powers in the world due to several reasons, In this study, we are not going to dwell on these reasons; instead we shall highlight the important features of Arabic languages which makes the principle players in the world arena to focus on studying this rich and complicated pleasant

language when exploring MT system development. To better understand the actual reason, it is worth to mention that MT was primarily used for code breaking. Russians and American (USA), shows a great interest in Arabic language as one of the earliest major languages which had undergone extensive experiments in MT. Codes breaking activities related to military, technical, and the scientific literature, in order to monitor the various opponent's fronts. In spite of the limited availability of the published literature on the history of MT, there is clear evidence that Arabic language scored remarkably high on the list of languages where MT tools were being researched and developed in the US.

When three-step scheme parsing algorithms was attempted in the late 1950s in the Operational Grammar Encoding Project ('COMIT'), Arabic was one of the languages besides English, German and French were the subjects of research, Yngve (2004). Similarly, in the development of Georgetown Automatic Translation project 'GAT' which was implemented in the late 1950s, Arabic language had been designated as a priority language by the US government Vasconcellos (2000), the programming of GAT on Arabic was taken over by Nancy Kennedy, a graduate student at the Institute, Vasconcellos (2000).

In the central laboratory for agricultural expert systems, this tool is found to be essential in developing bidirectional (English - Arabic) expert systems because both English and Arabic versions are needed for development, deployment, and usage purpose. The tool follows the rule based transfer MT approach. A major design goal of this tool is that it can be used as a stand-alone tool and can be very well integrated with a general (English-Arabic) MT system for Arabic scientific text Shaalan (2010).

Arabic language considered as one of the most difficult languages written and spoken Nemah (1973). In spite of this fact, Arabic language perpetually received praiseworthy consideration when translated into other languages principally due to its strong and challenging features (morphological, syntactic, phonetic and phonologic) Jihad Abdullah (1996). Written Arabic language processing was initiated in the 1970s. Several aspects of Arabic language had been researched in the early days of machine translation. Before the problems of Arabic text editing were completely solved,

Boualem (2003) had proceeded to call for working papers for a major conference on the Arabic language processing research which reflects the grand interest in Arabic language as far as MT is concerned.

Initial studies predominantly focused on lexicons and morphology. Since the internationalization and prevalent of WWW for more than a decade or so and the multiple productions of communication tools in Arabic language, the demand for a large number of Arabic NLP applications dramatically increased, consequently resulting in considerable research activities. In pursuit for this task, the researchers were dedicated to include diversified sectors comprehensively of Arabic language processing, including but not limited to; information retrieval, machine translation, document indexing, and syntactic analysis.

SAKHAR (2004) which is most famous Arabic speaking group working consistently on Arabic language translation, principally stated that, the staggering language differences of Arabic language from other languages in terms of its characters, morphology and shades of different connotative meanings, and to claim contrarily would be a mistake. Moreover, it is hardly realistic to import solutions from other languages without jeopardizing the uniqueness of the Arabic language features.

Guidere (2002) distinctively underlined two approaches for the study of machine processing of Arabic at large; "particularistic", and "universalistic". The first approach, depict the peculiarity of Arabic. Furthermore, it provides better working platform to the local processing approach that specifies the internal linguistic system of the Arabic language. In other words, it is interestingly involved with the semantic and the morphological aspects of Arabic language, especially the trilateral root system. Another well known company which is Systran has involved for a very long time in the development of software applications for different languages in the world including Arabic language, This company provided a list denoted as facts that help in translating Arabic (TranslationSoftware4u.com 2004), this list is characterized as very concise and with precise brevity. When engaged in translating the language, these facts list comprise crucial information about Arabic language that have to be handled with special care and attention. The facts are as follows:

a. Arabic writing sits on the line and direction from right to left and in a horizontal form. additionally have no capital letters, the Similarity in Punctuation between Arabic and English except for commas, commas in Arabic sit on the line, while in English under the line. All known nouns are assigned a gender in Arabic, with no neutral ones. There is spacing between words in a sentence.
b. Some letters change shape depending on their location in the word; namely, whether they are at the start, in the middle or at the end of the word.
c. There are 29 letters in Arabic, with 3 letter sounds which are non-existent in the English language.
d. Arabic does not distinguish between vowels and consonants
e. The use of a small sign on the top or under the letter indicates the pronunciation.

The second approach is the "universalist". This approach involves in adopting already available and tried methods on other languages like English, German or French and probing the potentiality for applications into the subject language with or without recasting the methods. Because the research works in both approaches "particularist" and "universalist" is directed at the syntactic characteristics of the linguistic system in general, they viewed as complementary pairs.

With minor contemplation to the approaches being used, (Guidere 2002) expressed a view stating that in regards to the available machine translation systems to and from Arabic are concerned mainly with the English-Arabic pair. As a matter of fact, this particular strategy has advanced the progress and development to develop electronic dictionaries with improved performance. Arabic linguistic phenomena had confined ambit in the other available applications by well known companies as such, he further alleged. For this limitation reason, such applications were predominantly based on specialized dictionaries. Therefore, their application was directed towards technical translation aid applications and not considered as machine translation software packages.

One can notice that most of the companies developing and then producing Arabic software applications have adopted the "universalist" approach, while, SAKHR proposed works using the "particularist" approach which clearly indicates the polarization status on both ends of the continuum. With the recent technological advances in MT, Arabic has received attention in order to automate Arabic translations (Farghaly et al., 2009)

2.5 RULE-BASED MACHINE TRANSLATION

Basic machine translation strategies described by the classic pyramid architectures was presented by Vauquios (1968). The architectures can be classified or categorized by three main disciplines as;

a. Direct or transformer architecture
b. Transfer based architecture
c. Interlingual architecture

Figure 2.1, principally describes the traditional approaches of machine translation systems, Vauquios (1968)

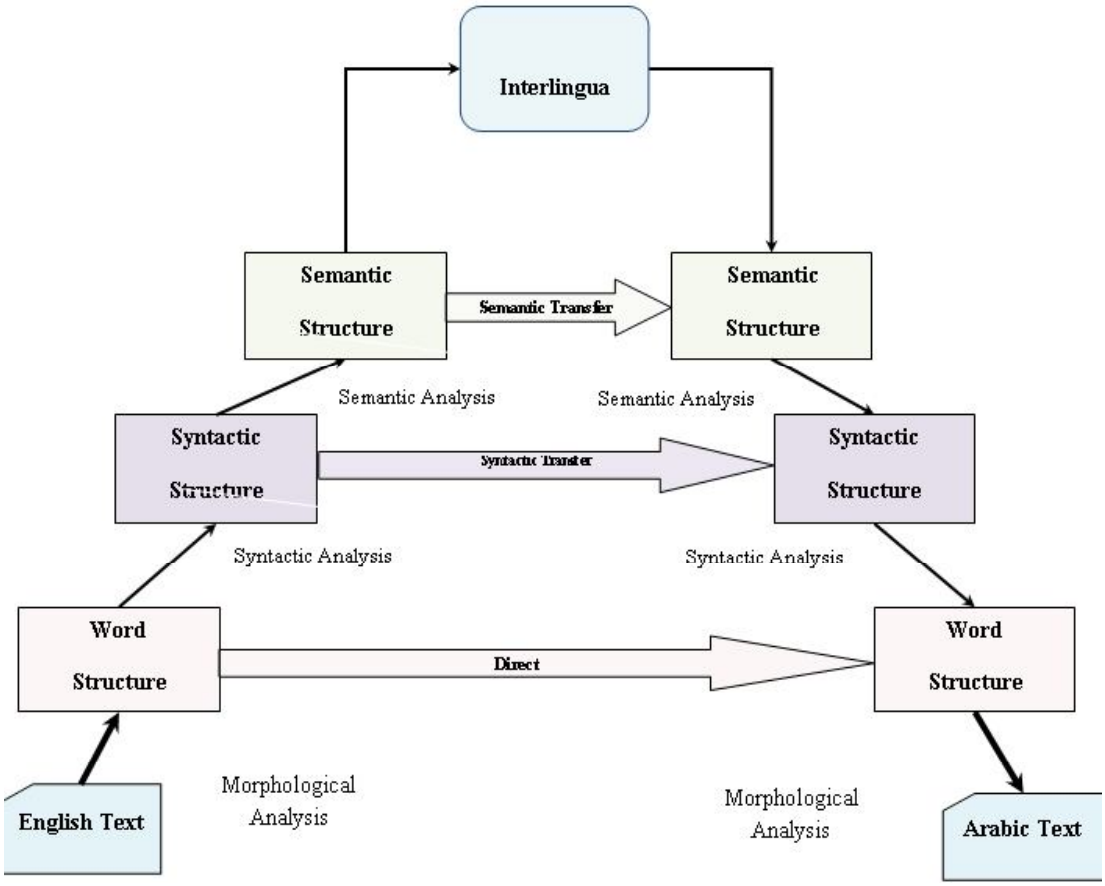

Figure 2.1 Machine Translation Systems Classification Vauquois (1968)

From Figure 2.1 we can observe that the bottom of the pyramid machine translation systems involves some sort of word-level analysis in the translation process which is designated as "direct machine-translation". Due to its bounded features, those systems are suitable for processing single target language cases. For The concept of direct machine translation MT system tailored for a pair of language; source and target. In this system, there is no need to analyze the structure of the sentence or to determine the structure of word formation, "morphological" analysis. Instead, word for word translation performed using large dictionaries. Post treatment for the produced translation to adjust the output to comply with the grammatical rules in word ordering and morphologically. Systran is a direct MT system.

The next level of this pyramid is the "indirect machine-translation" as transfer systems which involves three steps; parser and the source-language text is used to

conduct the first 'analysis' step on syntactic or even semantic level in order to create corresponding structures. Upon capturing the structure of the source language text whether syntactic or semantic, transfer-based method is used as the second step in the transfer strategy which implicitly administers changing the underlying representation of the source-language structure to the synonymous target-language structure.

Generation of translated text considered as the final step, where generator and target rule is used to produce the translated text from the transferred target-language structure. At the top of the hierarchy of translation methods in the classification diagram, we can notice the Interlingua systems which lie on top of this taxonomy; Also classified as "indirect machine-translation" this strategy involve using some kind of "universal" language as an intermediate step (either a common-logic representation or even a slightly modified natural language, along the lines of Esperanto), and then generate the target-language text from the universal representation.

Although Bar Hillel (1960) openly declares his commentary on the total concept of automatic high quality translation system and considered it as inappropriate, he was the first to comprehend with systems that were based on this approach. As mentioned earlier direct transfer is good for single language-pair translation system, while transfer and Interlingua systems are designed to deal with more than one pair of languages, Transfer-Based systems are smartly described through several examples by many researchers of MT Hutchins and Somers (1992);

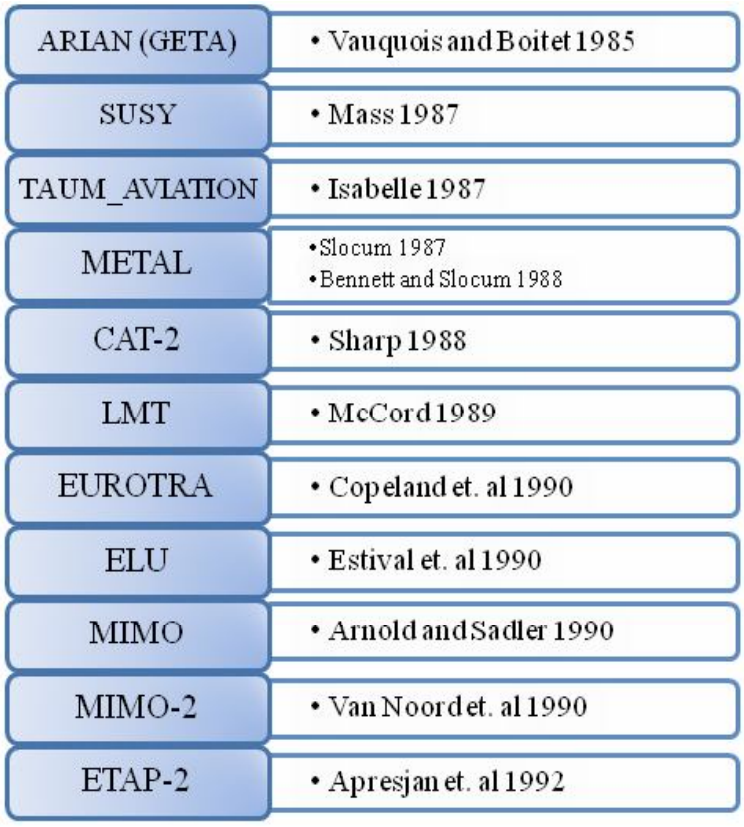

Figure 2.2 Transfer-based system examples

Classification is a problem of *enormous* dimensions, therefore, research paradigm considered as distinctly different dimension in machine-translation system classification. Two significant paradigms of current machine-translation presented by Dorr, Jordan and Benoit (1999);

a. Linguistic-based
b. Corpus-based

Predominant use of predefined set of rules, with or without some kinds of linguistic knowledge-bases to produce the translation is known as 'Linguistic-based' paradigms. In this paradigm, comprehensive knowledge of linguistic theory for the source language in addition to the target language is a perquisite to work with such discipline. While, all known three strategies; direct, transfer and Interlingua use 'Corpus-based' paradigms. Corpus-based paradigm is referred to as Example-Based

Machine Translation. As shown in Figure 2.1, a whole process can be envisioned here, were a small parallel bilingual corpus can be utilized for offline statistical learning, assisting in translation prediction of a newly introduced input, or for producing a recombined translation via searching readily available 'translated' examples.

The key concept of the work in this thesis is the utilization of both Example-Based with Rule-Based approach which enable us to design and implement a Hybrid machine translation MT system in consonance with other research works.

2.5.1 Direct or Transformer Architecture

When processing source language input text leads 'directly' to the required target language output text without any intermediate analysis, we refer to that as the direct strategy approach (W. Hutchins & H. Somers, 1992). A distinguishing attributes to this strategy is the resemblance with the machine translation systems that were developed in the 1950s and 1960s denoted as the 'first generation' MT systems. From its commencement, the core principle in the design of direct translation system is to deal with specific source and target language pair not necessarily accompanied by general linguistic theory or parsing principles, Tucher (1987).

The constituents of direct strategy processing in the MT systems have three stages;

i. First Stage: performing morphological analysis of the input text of the source language to determine word ending to minimize the inflected forms to their uninflected base forms.
ii. Second Stage: exploitation of the bilingual dictionary which is an 'enormous and abundant information source' to choose the right words and then replacing them with their corresponding match in the target language from the words in the source language.

iii. Third stage: The system processes the output text as an adjustment measures and to locally reorder the replaced equivalent words 'from stage ii above' and put them in their right order using the application of rules.

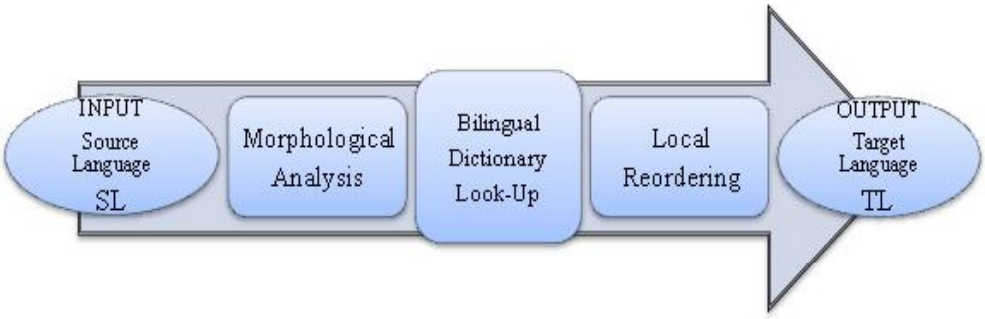

Figure 2.3 Direct MT Strategy, Hutchins & Somers (1992)

This process is ingeminated in Figure 2.2**.** We can observe in this process, the missing analysis utilization of syntactic structures or semantic relationships.

2.5.2 Transfer Based Architecture.

The transfer-based method is obviously viable if the system conduct the translation operation of the source language to target language is due to the execution of the following steps;

a. An input text analysis uses the parser and the source grammar in the analytical phase.

b. Switching the underlying representation of the sentence in the source language to the target language during the transference phase.

c. Target language grammar and generator are used to switch the underlying representation of the sentence from the source language to the target language during the synthesis phase.

In the middle position of the taxonomy, the **Transfer** method lies as an intermediate strategy between the **Interlingua** system and the **Direct** translation system of the MT. Figure 2.3, conceive the difference between those strategies Trujillo, (1999).

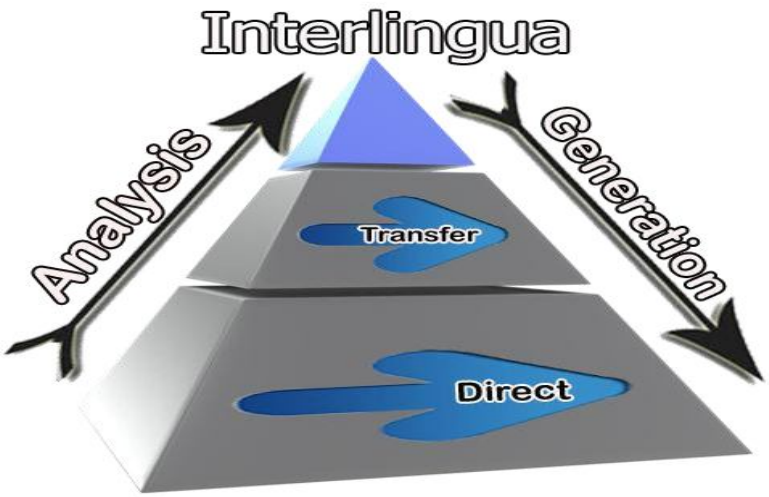

Figure 2.4 Direct, Transfer, and Interlingua MT Methods correlation, Trujillo (1999)

From the Figure 2.4, we can clearly deduct that direct method utilizes a set of rules for direct translation and has no modules for the analysis of the source language or the generation of the target language. However, in the Interlingua method, full analysis of the source language rendering language-independent representation.

Direct systems are characterized by ease of implementation, while the Interlingua systems contain enormous resources. Functional compromise can be achieved using the Transfer method as a strategy to efficiently utilize the favorable features of both methods, namely; resources of the Interlingua systems and the ease of direct systems. Accurate results can be obtained using transfer-based MT system, but high quality translation can be achieved when combined with example based, where agreement can be handled to solve translation problem, Somers (1999). In this combined method, features of the source language rendered by analyzing and extracting language-dependent representation, then, this representation transformed into the target output as a second representation by employing a set of transfer rules.

This new representation held target language features which are the product at the end of the generation module, Trujillo (1999).

The outstanding preponderance of research and development systems are transfer-based leaving out a small number of commercially available systems which remain essentially `Direct', and almost an equally small number of Interlingua systems that are particularly knowledge – based MT, Hutchins and Somers (1992) cited examples of the Transfer systems that have been explained in detail.

The transfer-based system has facilitated and empowered the design and implementation of a MT system that can deal with agreement and word ordering problems in sentence construction when confronted throughout the translation process of the English language to the Arabic language, however, engagement of human translators or unite with other strategy such as example-based to overcome certain complexities is still needed. Due to the fact that the transfer-based system alone cannot produce sufficiently successful translation output, therefore, we combined the two strategies and suggested new hybrid approach to overcome the encountered difficulties of both approaches when used alone and enhance each other features interactively via exploiting their favorite features and minimizing their drawbacks and shortcomings.

Scholars and researchers favorably choose the transfer method for its definite advantages:

a. Ease of implementation. Compared to the Interlingua system the transfer MT system faster to develop with much less effort. Thus, and due to this fact the market witness several operational transfer type machine translation systems.
b. Applicability. The level of analysis in transfer models is achievable. While, the abstruseness and complexities of the Interlingua systems render it hard to comprehend and attain.

No system is free from drawbacks and setbacks, despite the distinctive advantages of the transfer method, the approach still suffer from many shortcomings

like; it incurred high cost when translation process involves many languages due to the need for a substantial bilingual component that is tailored for a specific source language and target language pair (A.Tucher, 1987). Every time new language added to the system, component installation requires substantial time and effort. For example, if a transfer system is developed and designed to translate four languages, 20 modules are required, because the system needs 12 transfer modules in addition to four analysis modules and four generation modules, Hutchins & Somers (1992). Mathematically, for *n* languages;

Number of Language = n

The required number of transfer modules = N

The total number of transfer modules 'N',

$$N = n \times (n - 1) + n \text{ 'for analysis modules'} + n \text{ 'for generation modules'}$$

$$N = n \times (n - 1) + (2 \times n) \quad (2.1)$$

However, optimum design can alleviate the wearisome and arduous works in the transfer modules and the creations of new modules. The transfer method can be greatly improved and the cost-effectiveness in a multilingual environment can be immensely enhanced by employing a number of techniques, such as:

i. Emphasis on in-depth analysis: in-depth analysis of the source language greatly reduced the requirement of work in the transfer component.
ii. Reversibility of transfer rules: This can reduce workload if used in a transfer module. Required work can be reduced by half if the translation from English language into Arabic language can be reversed. Unfortunately, not all transfer rules in a component are reversible.
iii. Maximal Allowed Shareability and Reusability: closely related languages can share and reuse some transfer rules by different transfer components, Trujillo (1999).