ANN and Rule Based Method for English to Arabic Machine Translation

Marwan Akeel and Ravi Mishra
Department of Computer Engineering, Indian Institute of Technology (BHU), India

Abstract: *Machine translation is the process by which computer software is used to translate a text from one natural language into another language with or without minimal human intervention. This definition involves accounting for the grammatical structure of each language and using rules and grammars to transfer the grammatical structure of the source language into the target language. This paper presents an English into Arabic Machine Translation (MT) system for translating simple well-structured English sentences into well-structured Arabic sentences using a rule-based approach and feed-forward back-propagation Artificial Neural Network (ANN). Our system is able to translate sentences having gerunds, infinitives, prepositions and prepositional objects, direct objects, indirect objects, etc. Neural network works as bilingual dictionary, which does not only store the meaning of English word in Arabic but it also stores linguistic features attached to the word. The performance is evaluated by different MT evaluation methods. The n-gram blue score achieved by the system is 0.6029, METEOR score achieved is 0.8221 and 0.8386 on F-measure.*

Keywords: *MT, neural network, back-propagation, rule based translation, English-Arabic machine translation system.*

Received April 21, 2012; accepted March 20, 2013; published online April 4, 2013

1. Introduction

World wide web, globalization and international business bring the world together. In order to have smooth communication, the need of breaking the language barriers appears. Here comes the study of machine translation systems, which is the study of designing systems that translate text from one human language into another language with or without minimal human intervention. Machine translation is an automated process, in which translation job is done by the computer software. Machine translation is an application of computer linguistics. Computer linguistics is an interdisciplinary field of computer science and requires language and computer experts. Translation as an art of rendering a work of one language into another is as old as written literature [7]. Computer technology has been applied to technical translation to improve one or both of the following factors [25]:

- *Speed:* Translation by or with the aid of machines can be faster than manual translation.
- *Cost:* Computer aids to translation can reduce the cost per word of a translation.

If MT researchers are able to develop a perfect multilingual machine translation system, people with different languages can share ideas and information worldwide on every topic as business, economic, educational, political, socio-cultural, etc. The purpose of a translation process (whether machine translation or human translation) is that meaning of the text being translated should not change. There are many different machine translation systems available online or desktop systems. One of the machine translation studies is concerned with translating from English into Arabic. Arabic language is important because of the current economic and political issues and also the big number of countries that speak it. During the last decade, many researches have been conducted in machine translation, focusing on Arabic language.

There are some English-Arabic translation systems based mainly on the transfer model such as Alwafi and Al-Mutarjim, but they are at the beginners' stage when compared with the available translation [3]. Al-Mutarjim is the commercial software that is available from ATA software which translates English text into Arabic. Some studies discussed the problem of the English-Arabic translation of the embedded proverb expressions and idioms in the English sentences [21-9]. Rafea *et al.* [19] has developed an English-Arabic MT system which translates a sentence from the domain of the political news of the Middle East. A machine translation was developed by Pease and Boushaba to translate medical texts from English into Arabic [18]. Mokhtar [16] has developed an automated English-Arabic MT system of scientific text, which was applied to a set of abstracts from the field of artificial intelligence. There are systems that are used to translate Web pages from English into Arabic such as the system that uses commercial machine translation system to translate the textual part of a Web page from English into Arabic automatically [27]. Then, it displays a web page containing the

Arabic translation with all tags inserted in the right places and thus the layout and content of original English. Most recent studies focus on word order and agreement problem to enhance the translation quality of English into Arabic [1, 2, 3].

2. Arabic and English Language Features

2.1 Arabic Salient Features

Arabic belongs to the Semitic language family. The members of this family have a recorded history traced back thousands of years and one of the most extensive continuous archives of documents belonging to any human language group [5]. It is spoken by more than 356 million people as a native language, in an area extending from the Arabian Gulf in the East to the Atlantic Ocean in the West [12]. There are different varieties of this language depending on the regions. But contemporary Arabic or the standard Arabic (an offshoot of the classical Arabic language) is the language that is taught in the schools and universities. Even the Arabic media use the standard Arabic as the medium.

Arabic language has always been considered due to its morphological, syntactic, phonetic and phonologic properties which is one of the most difficult languages for written and spoken processing [1, 4, 28]. It has two main types of sentences: Verbal sentences and equational or copula sentences. It is a relatively free word order language, structured under the combinations of SVO, VSO, VOS and OVS for the elements of subject (S), verb (V) and object (O). The most common synchronize structures are SVO and VSO [6]. Its alphabet consists of 28 characters, where the shape of each character depends on its position within a word:

(SVO) Sami takes the book - سامي أخذ الكتاب
(VSO) Takes Sami the book - أخذ سامي الكتاب

The modern Arabic dialects are well-known as having agreement asymmetries that are sensitive to word order effects. Word agreement and ordering plays an important part in constructing Arabic sentence. Subject and verb must agree in number, gender, and person, taking into consideration the features of the subject, which are important factors in the derivation of the verb as well as the features of the verb itself. Other agreements are required between the adjectives and the nouns where Arabic adjectives depend on the number, gender and person as well as the definiteness and indefiniteness of the nouns. Some other agreements also exist between the numbers and the countable nouns [1]. Arabic is not alone in showing word-order asymmetries for agreement. Similar asymmetries have been documented in Russian, Hindi, Slovene, French and Italian [8].

Here, are some examples of Arabic sentences having adjectives and nouns:

1) beautiful car - سيارة جميلة .
   car (tp-sing-fem) beautiful (tp-sing-fem).
2) the beautiful car - السيارة الجميلة.
   the car (tp-sing-fem) the beautiful (tp-sing-fem).
3) the beautiful cars - السيارات الجميله.
   the cars (tp-plur-fem) the beautiful (tp-sing-fem).
4) beautiful kids - أطفال جميلون.
   kids (tp-plur-masc) beautiful (tp-plur-masc).

In the above examples, the adjective agrees with the noun or the head word in number, gender, person and definiteness. We use the abbreviation 'sing' and 'plur' to represent the singularity and plural features respectively, the gender features denoted by 'masc' for masculine and 'fem' for feminine, and 'tp' is the abbreviation denoting the third person. One of the exceptions in Arabic words agreement is when the plural noun or head word does not have the human feature then the adjective should be in singular feminine form. In the third example, the cars are plural and not human; so, the adjective beautiful comes in singular feminine form where in the forth example the kids is plural and human; so, it follows the rule and the adjective comes also in plural form and both agree on the gender feature.

The nouns in Arabic are divided into feminine and muscular. The feminine nouns are used while referring to the female and masculine ones are used while referring to the male. In some cases, the feminine noun is formed by adding a character "ta marbuta - ة" at the masculine noun's end:

doctor (fem) - طبيبة ; doctor (masc) - طبيب

The above example shows how the use of the 'ta marbuta' changes a noun from a masculine to a feminine. However, there are also some pairs that use two totally different nouns for referring to the masculine or feminine. For instance, Arabic nouns used to refer to the word "man" and "woman" is completely different as given below:

woman - امرأه ; man - رجل

The genders are not only used to refer to people but they are also used for other objects that are considered to be feminine or muscular. "ta marbuta" is again the character that is used to distinguish between a masculine and feminine in object like table, room or paper. Yet, there are nouns that do not use the "ta marbuta" to change a muscular noun to feminine. In addition to the masculine/ feminine distinction, Arabic has singular, dual and plural forms of nouns, pronouns, verbs, adjectives, etc.

Arabic has three cases, then: The nominative, the accusative and the genitive. The use of cases in Arabic is complicated by the fact that the Arabic script only allows the writer to show the consonants and long vowels of a word. The short vowels can be indicated by a system of straight and curved lines placed above and below the letters, but these are time consuming to write and are normally included only in texts where it is important to indicate correct pronunciation.

The verbs of Arabic differ from those of English as well, particularly in how their tenses (whether they

refer to past, present or future actions) are perceived. In Arabic, the basic distinction of verb tense is between "completed" and "not completed" actions [5].

2.2. Comparative Linguistic Features of English and Arabic

English language is a member of the Indo-European family of languages. It becomes extremely important in the world as a universal communication language. It is used in business, politics, literature, science and academic studies. Arabic language is the official language of most of the Middle East countries and widely used throughout the Muslim world. Both languages are among the United Nations official languages. Table 1 shows some of their basic features.

Table 1. English and Arabic basic features.

Feature	English	Arabic
Letters	Contains 24 letters; 5 vowels. G, P and V do not exist in Arabic language.	Contains 28 letters 3 vowels (Alif, Wāw, Ya') in addition to aspiration marks Fatha, Damma, Kasra, Sukūn, Shadda (or tashdīd).
Word Type	Noun, verb, adjective, adverb, pronoun, conjunction, interjection.	It is either verb, noun or article.
Read and Write Direction	Left to write.	Right to left
Letter form and Punctuation	Comes in capital or small cases.	Changes when occurring alone, in the beginning, middle, or end of a word. Rules for punctuation are much looser than in English.
Gender	The noun is either (male, female or neuter); The verb form does not differentiate between male and female.	(Male, female) It differentiates between male and female in verbs and nouns.
Number	(Singular, plural).	(Singular, dual, plural)
Numbering System	Arabic numbers.	Uses 2 numbering system of its own, the first set of numbers is originally borrowed from India and the second is its original one.
Base form of a Verb	The present-tense form for all persons.	Past form of the singular third person masculine.
Verb and Sentences	The sentence must have a verb.	The sentence either starts with a verb (verbal) or it starts with a noun (nominal).
Case System	(Nominative, accusative, genitive, dative)	(nominative, accusative, genitive)
Word Order	The general pattern is: Subject (S), verb (V), object (O)	Free word order allows the combinations of SVO, VSO, VOS and OVS. The most common synchronize structures are SVO and VSO [6].
Definite and Indefinite Article	It has both definite and indefinite article.	There is a definite article but The indefinite article does not exist.
Tense	(Past, present, future); It focuses on the action.	(past, present); If the action is completed or not completed.
Spelling	Hearing what you think is spelled.	Write what you hear, there are no silent letters except in few rare cases.

3. Problem Description

Our method used in building the English to Arabic machine translation is rule based and neural network approach. Rule based approach is the classical approach of machine translation. It consists of collection of rules called grammar rules, lexicon and software programs to process the rules. It is extensible and maintainable. Rule based approach is the first strategy ever developed in the field of machine translation. Rules are written with linguistic knowledge gathered from linguists. Rules play major role in various stages of translation: Syntactic processing, semantic interpretation, and contextual processing of a language. Systems that use rule-based transformations are based on a core of solid linguistic knowledge. The linguistic knowledge acquired for one natural language processing system may be reused to build knowledge required for a similar task in another system. Shaalan mentioned tow advantages of the rule-based approach over the corpus-based approaches that are [22]:

1. Less-resourced languages, for which large corpora, possibly parallel or bilingual, with representative structures and entities are neither available nor easily affordable.
2. For morphologically rich languages, which even (with the availability of corpora) suffer from data sparseness.

In rule based machine translation approach, the system is fed with linguistic rules and bilingual dictionaries. System parses and analyzes the grammatical structure of the source language text, in which the structure is transformed to the target language structure with the help of the linguistic rules. When the structure is transformed, target language text is generated by the use of bilingual dictionaries and linguistic rules. Many systems have been developed using rule based machine translation, such as Systran, Eurotra and Japanese MT System. Neural networks are a possible solution to the machine translation problem. Neural networks have the ability of learning through examples. Neural network has been proven to be very useful in various natural language processing tasks [10]. PARSEC [11], JANUS [26], English-Sanskrit MT system [15] and English-Urdu MT [13] use neural network approach for natural language processing task and automated machine translation. Our English-Arabic machine translation system uses Feed-Forward Back-Propagation Neural Network with rule based machine translation approach. Neural networks are very efficient in pattern matching. Machine translation using rule based approach is consistent with predictable quality.

Our system uses neural network as knowledge base for bilingual dictionary. Neural network maps Arabic words/ tokens (such as verb, noun/ pronoun etc.) equivalent to English words/ tokens. These words/ tokens are then processed using the rules in rule base side of the machine translation to construct its right form.

4. System Architecture and Description

The block diagram of our English to Arabic machine translation system is shown in Figure 1. There are two models, the ANN model and the rule based model. The ANN model works as a bilingual dictionary consist of a train object. In rule based model, there are seven main modules: Contractions removal, parser and tagger, knowledge extraction, grammar and sentence structure recognition, ANN and rule based word mapping, Words selection and syntax addition, and Arabic sentence generation. The function of each module is explained below:

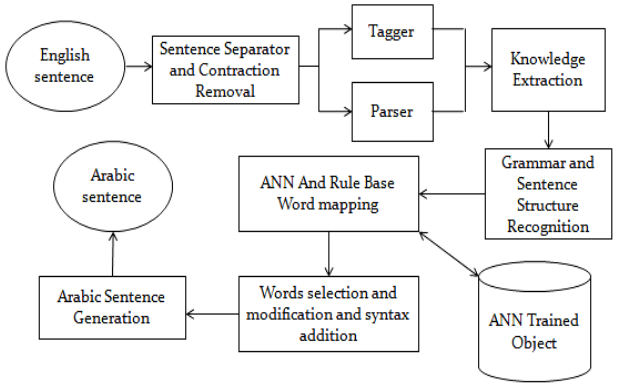

Figure 1. System architecture.

- *Sentence Separator and Contractions Removal:* The translation process starts with English text input to this module. This module first separates the paragraph into sentences. Then each sentence is processed. If any contraction is present in the sentence, it is removed. Contractions are common in spoken English and now becoming informal in written English too. In this step, we replace contractions with their respective full form.
- *Parser and Tagger:* The output text from the sentence separator and contraction removal is given as input to the Parser and Tagger module. Stanford typed dependency parser is used for parsing the English Text. Stanford parser is the implementation of probabilistic natural language parsers, both highly optimized PCFG and lexicalized dependency parsers, and a lexicalized PCFG parser. Probabilistic parsers use knowledge of language gained from hand-parsed sentences to try to produce the most likely analysis of new sentences. These statistical parsers still make some mistakes, but commonly work rather well [23]. The parser provides Stanford typed dependencies as output. The dependency is written as abbreviated_relation_name (governor, dependent). We are using Stanford POS tagger for tagging the English text. A Part-Of-Speech Tagger (POS Tagger) assigns parts of speech to each word (and other token), such as noun, verb, adjective, etc. The Stanford POS tagger uses the Penn Treebank tag set and is implemented using maximum entropy tagging algorithm [24]. POS tagger adds part of speech information to each word (and other tokens) in the text.
- *Knowledge Extraction:* The function of this module is to process the typed dependency obtained from parser and to process tagged text from the tagger. Each word of the sentence and all the information related to it are collected and encapsulated in a knowledgeable object so the sentence is represented as a collection of knowledgeable objects.
- *Grammar Analysis and Sentence Structure Recognition:* This module processes the collection of knowledgeable objects and recognizes parts of the sentence, such as subject, main verb, auxiliary verb, object, indirect object, etc. Each recognized part of the sentence contains one or more word called chunk. So, each sentence is divided into chunks. The tense of the sentence is recognized with the help of main verb and auxiliary verb. Sentence voice, whether it is passive or active, is also recognized in this phase. Sentence type is detected from the knowledge present in the collection of knowledgeable objects. On the basis of knowledge obtained, sentence parts and attributes (tense, voice, type etc.) are analyzed and sentence grammatical structure is generated with the help of rule base. The corresponding Arabic sentence structure is obtained using the sentence tense and the English sentence structure and type.
- *ANN and Rule Based word Mapping:* Sentence chunks (or parts) are translated according to the grammar structure obtained from last module. Each sentence chunk has to be taken and translated separately. Word mapping module encodes each word in the chunk into numeric form and send it to the ANN trained object at the ANN model, which is trained for word mapping, and gets the corresponding Arabic meaning and associated information in numeric form. The numeric returned value is decoded to textual form and passed to the word selection and modification module. The work is done when all the sentence chunks are processed.
- *Words Selection and Modification and Syntax Addition:* In this module, we select the correct form out of the received Arabic translation of each chunk word. The selection is done depending on the sentence tense or the features of the chunk main word or the subject main word. Then the necessary modification is done by adding affixes to the selected form or word.
- *Arabic Sentence Generation:* In this module, the chunks translation are arranged in accordance with the Arabic sentence structure obtained previously and a necessary leading word is added when needed such as كان, ان. Finally, the Arabic sentence is generated.

5. Encoder-Decoder

To build the ANN trained object, We created a data set of input-output pairs of English-Arabic words with associated knowledge. Encoder-Decoder converts this

training data into numeric coded form, which is suitable to be used as input for the ANN trainer. Each English and Arabic alphabet and special character is represented as a unique integer number between 1 and 72 (a = 1, b = 2, c = 3, …,ا = 27, ب = 28 , ت = 29, … and so on) as shown in Table 2. 72 is the total number of English and Arabic alphabets, numbers from 1 to 9 and some special characters. Converting the characters into integer numbers is to train the neural network. The output of the network training is the ANN trained object.

Table 2. Characters encoding.

Character	The Numeric Integer Number	Character	The Numeric Integer Number
a	1	ز	37
b	2	س ش	38
c	3	ص ض	39
d	4	ط ظ	40
e	5	ع غ	41
f	6	ف ق	42
g	7	ك	43
h	8	ل م	44
i	9	ن ه	45
j	10	و ي	46
k	11	آ ى	47
l	12	ؤ ء	48
m	13	ة	49
n	14	Space	50
o	15	|	51
p	16	*	52
q	17	#	53
r	18	1	54
s	19	2	55
t	20	3	56
u	21	4	57
v	22	5	58
w	23	6	59
x	24	7	60
y	25	8	61
z	26	9	62
ا	27	0	63
ب	28	!	64
ت	29	?	65
ث	30	;	66
ج	31	:	67
ح	32	'	68
خ	33	"	69
د	34	(	70
ذ	35	)	71
ر	36	,	72

The text word into numeric and decode back the numeric into text as shown in Figure 2.

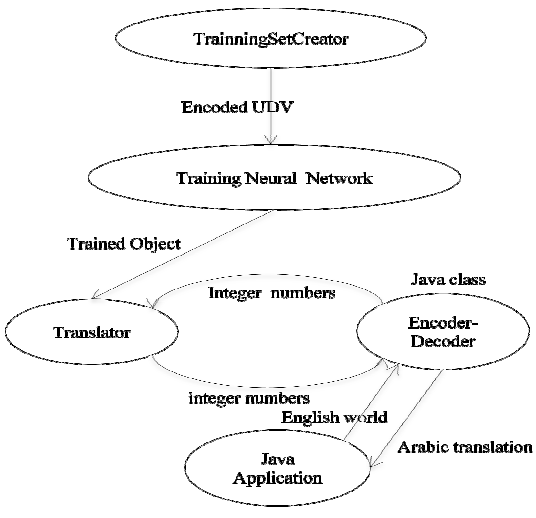

Figure 2. Mapping process.

6. ANN Based Mapping Process

We used feed-forward back-propagation artificial neural network for the selection of Arabic words/ tokens (such as verb, noun/ pronoun etc.) equivalent to English words/ tokens. Each English word is matched to only a single meaning in Arabic.

For a word, which is noun or adjective, we may have up to six forms of the Arabic word returned from the ANN along with its coupled information. The six forms are two for the singular masculine and feminine, two for the dual masculine and feminine and remaining two forms are for the plural masculine and feminine of the Arabic noun or adjective. In some cases, a flag value attached to digit 8 is also returned with the Arabic word forms which may be needed later for agreement purpose. The coupled information which may come with the noun or the adjective is the number, person and gender. For example, "blue" is mapped to "أزرق1 زرقاء2 زرقان3 زرقاوتان4 زرق5 زرق6" and "pen" is mapped to "قلم |tp #sing *masc" where tp is third person, sing is singular and masc is denoting the masculine. For an English verb, the ANN returns seven forms of the corresponding Arabic translation and a letter variable. These Arabic verb forms are one for the past tense, one for the present tense, one for imperative, three passive forms (past, present, adjective), and last is the noun form of the verb. The letter variable is to indicate whether the verb belongs to a special criterion or not. These verb forms are mainly of the singular third person masculine. For example, the verb "brought" is mapped to "أحضر1 يحضر2 أحضر3 أحضر4 يحضر5 محضر6 احضار7 ن8". Verbs in neural network model are trained in their entire possible occurrence in English (bring, brings, brought, bringing). The words which may come as noun or verb are mapped into two corresponding meanings, one for each. For example, the word "watch" comes as a verb or noun so it is mapped to two different meanings ("watch" → "شاهد1 يشاهد2 شاهد3 يشاهد4 شوهد5 مشاهدة6 مشاهدة7 ن8 : ساعه |tp #sing *fem"), one when it comes as a verb and the other when it comes as a noun. The both meanings are separated by a colon mark. The following are three main steps in mapping a word as shown in Figure 2:

1. Encoding of English words/ tokens or User Data Vector (UDV).
2. Mapping of English numeric code: Data sets are fed to neural network from which ANN selects the Arabic equivalent of the English words/ tokens provided for translation.
3. Decoding the code of the obtained Arabic words/ tokens or decoding of UDV. Once we get the equivalent words/ tokens, Arabic meaning and information are extracted and processed.

7. Translation Rules

Translation rules have been created for various classes of the sentences. From the input of the English

sentence to the output of the Arabic translation, the sentence and its parts are processed through a huge number of rules. Our system is able to handle simple English sentences (assertive, interrogative, imperative) with the forms of (affirmative, negative) for both voices (active, passive). Sentence structure, tense, type, form and voice are found using the part of speech of the sentence and its type dependency. Selection, syntax and affixes addition to the verb, and affixes addition to the subject and object will be done on the basis of information of tense, subject and object gender, number and person. We are going to demonstrate some of the rules, which are applied by the MT by taking the following sentence as an example: She will not send my old car to the mechanic. First we obtain the sentence chunks/ sentence parts using the Parser and Tagger:

***S**ubject = "She".*
***A**uxiliary = "Will not".*
***V**erb = "Send".*
***O**bject = "My old car".*
***Prepos**ition = "To the mechanic".*

Some of the rules applied on the sentence at the MT are: At the Grammar Analysis and Sentence Structure Recognition module, to get the sentence tense and type, and the Arabic sentence structure, the following rules are applied: If (auxiliary equals to will), then (the tense is Future-Indefinite), if (the sentence auxiliary contains not), then (the sentence is negative), if (sentence tense is Future-Indefinite and sentence is in negative active voice and English sentence structure is "SVOPPOS" starting with SV), then (the Arabic sentence structure is starting with VS replacing SV → "VSOPPOS"). The auxiliary chunk is not considered on the sentence structure and it is only used to determine the tense of the sentence and wither the sentence is affirmative or negative.

At the word mapping and selection module, the sentence is translated according to the Arabic sentence structure generated on the previous step. Starting with the first chunk which is "send", and after getting the translation from the ANN module we apply the following rules: If (sentence tense is Future-Indefinite and voice is active), then (select the present form or the second form of the Arabic verb "أرسل1 أرسل2 يرسل3 أرسل4 يرسل5 مرسل6 ارسال7 ن8" → "يرسل"), if (subject is pronoun and equals to she), then (remove the leading "ي" from the verb "يرسل" and add prefix "ت" to it → "ترسل"). Second chunk, if (the subject is pronoun), then (do nothing). If the subject is pronoun we do not individually translate it but we represent it in the verb form by adding affixes. Processing the object chunk, "my old car", starting from the most right word, which is the chunk main word "car" and after getting the Arabic corresponding match, "سيارة |tp #sing *fem", we apply the following rule: If (the chunk contains "my" as possessive pronoun and the word "سيارة" gender is fem ending with "ة"), then (remove "ة" and suffix "تي" → "سيارتي"). After getting the meaning of the word old from the ANN module, which contains six forms, "قديم1 قديمه2 قديمان3 قديمتان4 قدماء5 قديمات6", apply the following rules: If (the chunk main word is singular and feminine), then (select the singular feminine form or the second form → "قديمه"), if (the chunk contains possessive pronoun and the word is not the chunk main word), then (add prefix "ال" to the word "قديمه" → "القديمه"). The Preposition chunk, "to the mechanic", is also processed from the right most word, mechanic, and the return value from the ANN is " ميكانيكي1 |tp #sing *masc, ميكانيكيه2|tp #sing *fem". The following rules are then applied: If (masculine and feminine form of the noun mechanic is available), then (select the masculine as default → "ميكانيكي"), if ("the" is existing in the chunk), then (prefix "ال" to the word "ميكانيكي" → "الميكانيكي"), if (the chunk contains "to" and the chunk main word is noun), then (prefix the chunk translation "الميكانيكي" by "الى" → "الى الميكانيكي"). The, a, an and possessive pronouns are not translated one to one but they change the form of the related nouns and adjectives by adding some features to them. The final output will be: If (sentence tense is Future-Indefinite and negative), then (add "لن" at the beginning of the sentence → " لن ترسل سيارتي القديمه الى الميكانيكي").

8. Implementation

We have implemented our English-Arabic machine translation system on java platform. We used java jdk1.5 version for its compatibility with Matlab 7.1. System is implemented in java except the neural network module. Neural network model is trained, tested and successfully implemented, using Matlab 7.1 neural network library. Neural network works as the knowledge base for bilingual dictionary. Bilingual dictionary does not only store the meaning of English word in Arabic but also store linguistic knowledge (e.g. number, person and gender, etc.) attached to the Arabic word. We trained the two-layer feed-forward neural network with Levenberg-Marquardt back-propagation algorithm. The neural network for knowledgeable bilingual dictionary has been trained with a data set of around 3000 input-output pair of English-Arabic words with its associated knowledge. A java class is created to help building and training data in numeric form. The java class encodes data present in human readable form in a text file to data in numeric form to be used in training the neural network. The created java class also does the encoding and decoding of the tokens before sending it and after receiving it to/ from the neural network. Neural network gets numeric input, mapping its value and then returning the result, which is in numeric form back to the java class that converts it to readable form. This knowledge is further processed and Arabic meaning and attached information are extracted. Figure 2 shows the steps.

9. Performance

Our MT system is first separates the paragraph into sentences. The sentences are processed one by one. Each sentence is gone through number of steps: The MT system checks if contraction is present in the English input sentence. It removes any contraction and rewrites it again in full form, for example, we're, they've, she'll and I'd, etc., will be replaced by we are, they have, she will and we had/ would etc., respectively. Similarly, negative contractions aren't, needn't, won't etc., will be replaced by are not, need not, will not etc., respectively.

The output of contraction removal module is passed to the parser and tagger module which uses Stanford parser and tagger for parsing and tagging the input English sentence. Example of the parser's output for an English sentence is shown below:

I brought my blue pen to write the notes
[nsubj (brought - 2, I - 1), poss (pen - 5, my - 3), amod (pen - 5, blue - 4), dobj (brought - 2, pen - 5), aux (write - 7, to - 6), xcomp (brought - 2, write - 7), det (notes - 9, the - 8), dobj (write - 7, notes - 9)]

The output of the tagger for the same English sentence is as follows: I/ PRP brought/ VBD my/ PRP$ blue/ JJ pen/ NN to/ TO write/ VB the/ DT notes/ NNS. Knowledge extraction module processes the result obtained from Parser and Tagger module and converts each word of the sentence to a knowledgeable object containing the word and all its associated information. Sentence is now represented as a collection of knowledgeable objects which are then given as input to the Grammar Analysis and Sentence Structure Recognition module. This module analyzes these objects and identifies the attributes of the grammar for the English sentence as tense, voice, sentence type, subject, main verb, auxiliary verb, object, indirect object, etc. Some of the extracted sentence parts and attributes value, of the above example sentence are:

S = "I".
V = "Brought".
O = "My blue pen".
Infinitive = "To write".
O = "The notes".
Tense = Past indefinite.
Voice = Active.
Type = Assertive – affirmative.
English Structure = SVOInfO.
Arabic Structure = VSOInfO.

The ANN and rule based word mapping will deal with the sentence as individual chunks and starts translating the content of the chunks word by word, for the above example, the chunks obtained are: I, brought, my blue pen, to write, and the notes. Each word/ token in a chunk is mapped from the neural network as explained earlier in the implementation section. Verbs, nouns, adjectives and other parts meaning are stored in ANN base form so each Arabic word will be sent to the words selection and modification and syntax addition module to add the necessary affixes according to the sentence tense, gender and number of the subject, object and the chunk main word. Then the Arabic word mapping will process the next word or chunk and the circle will continue until no more words or chunks remains. For example, out of the Arabic forms received for the verb brought the first form احضر is selected because the sentence is active and the tense is past Indefinite. A necessary modification should be done to the verb to match the subject "I" and the verb form becomes احضرت.

All the parts of the sentence are then arranged at the Arabic sentence generation module according to the Arabic sentence structure obtained previously and a necessary leading word will be added and the output is presented in Arabic script form. The translation of the example sentence is:

I brought my blue pen to write the notes.
احضرت قلمي الأزرق لأكتب الملاحظات.

The input of English paragraph: *I have seen my brother in the morning. He was eating his breakfast. He wanted to go to the market to buy clothes. He did not have money. Leena gave him his brown bag. He went to study math. The output of Arabic translation:*

قد رأيت أخي في الصباح، كان يأكل أفطاره، أراد أن يذهب الى السوق ليشتري ملابس، ما ملك نقود، منحته لينا حقيبته البنيه، ذهب ليدرس حساب.

10. Results and Evaluation

One of the most difficult things in machine translation is the evaluation of a proposed system. We have a problem that there is not only one good translation. On the other hand, there may be many perfect translations of a given source sentence. These translations may vary in word, choice or in word order even when they use the same words. And yet humans can clearly distinguish a good translation from a bad one [17].

Our system is capable of dealing with different kind of simple sentence structures of assertive, imperative and interrogative type. We have selected a test set of thirty sentences covering most of the structures and types that the system can translate or under its scope. The translation was correct or acceptable to most of the test set sentences shown in Table 3.

The errors, observed on the system output translations, generated from different sources. Some errors found on the meaning given to some words, which are not the right one and that because every English word is given only one corresponding Arabic meaning. The words, which are not present in the bilingual dictionary are printed as it is in the translation in capitals. The error made by parser or tagger for any sentence at earlier stages will be propagated throughout the translation process and will result in wrong translation. Syntactical features, which have not been yet covered, are also causing errors.

Table 4 shows sentences from the test set which generate wrong translation outputs demonstrating some of the error cases. The candidate sentence is the machine output and the reference sentence is the human translation. The differences between the candidate and the reference translation sentences are underlined.

Table 3. Correct sentences translation.

Sentence Type	English	Generated Arabic Translation
Assertive Active Affirmative	The tree leaves fall in Autumn.	تسقط اوراق الشجره في الخريف.
	They are doctors.	انهم أطباء.
	I need a single room.	أحتاج غرفه فرديه.
Assertive Active Negative	The students did not want to come to my school to play cricket.	ما أراد الطلاب ان يحضروا الى مدرستي ليلعبوا كريكت.
Assertive Passive Affirmative	Coffee is grown in Brazil.	تزرع القهوة في البرازيل.
Assertive Passive Negative	The message has not been deleted.	لم تمسح الرساله.
Interrogative Active Affirmative	Where do you keep your money?	اين تحفظ نقودك؟
	Were the players ready to enter the field?	هل كان اللاعبون مستعدين ليدخلوا الميدان؟
	Had they taken the umbrellas with them?	هل كانوا قد أخذوا المظلات معهم؟
Interrogative Active Negative	Why will not Sara participate in the conference?	لماذا لن تشارك ساره في المؤتمر؟
	Are not they twins?	أليسا توئمان؟
Interrogative Passive Affirmative	When were the glasses broken?	متى كسرت النظارة؟
	Has my room been cleaned?	هل قد نظفت غرفتي؟
Interrogative Passive Negative	Why was not the food been eaten?	لماذا لم يؤكل الطعام؟
	Will not the printer be repaired?	ألن تصلح الطابعه؟
Imperative Active Affirmative	Translate the following sentences.	ترجم الجمل التاليه.
Imperative Active Negative	Do not write on the wall.	لا تكتب على الجدار.

It has been seen from the results that system performs efficiently on those classes of sentences that grammatically and syntactically covered and it can be also improved. Precision and recall are widely used to evaluate NLP systems including machine translations. It compares a set of candidate items *Y* to a set of reference items *X* independent of word order:

$$Precision\ (Y|\ X) = |\ X \cap Y\ |\ /\ |X|$$
$$Recall\ \ (Y|\ X) = |X \cap Y|\ /\ |Y|$$

BLEU [17] is an IBM-developed metric. It uses a modified form of precision (modified n-gram precision) to compare the candidate translation against reference translation. It takes the geometric mean of modified precision scores of the test corpus and then multiplies the result by exponential brevity penalty factor to give the BLUE score. Papineni *et al.* [17] also indicated that BLEU correlates very highly with human judgments. It is the most commonly used MT evaluation method [16].

METEOR [20] is a machine translation evaluation metric developed at Carnegie Mellon University. The Meteor metric is based on the weighted harmonic mean of unigram precision (P = m / wt) and unigram recall (P = m/ wr). Where m is number of unigrams, wt is the number of unigrams in candidate translation and wr is the reference translation.

F-Measure [14] in is a metric developed on the New York University. The F-measure is defined as the harmonic mean of precision and the recall as: F-measure = (2 * Precision * Recall) / (Precision + Recall).

Table 4. Translations containing some errors.

English Sentence	Candidate Translation	Reference Translation
He is on the way to here.	انه على الطريق الى هنا.	انه في الطريق الى هنا.
Sami has lost his wallet.	قد خسر سامي محفظته.	قد اضاع سامي محفظته.
Jogging is good for health.	ان هروله جيده للصحه.	ان الهروله جيده للصحه.
The mountains are beingcoveredwith snow.	ان الجبال تغطى مع الثلج.	ان الجبال تغطى بالثلج.
The policeman was coming to catch the thief.	كان الشرطي قادم ليمسك السارق.	كان الشرطي قادما ليمسك السارق.
The professor will leave to Amsterdam.	سيغادر البروفيسور الى Amesterdam.	سيغادر البروفيسور الى امستردام.
Is he studying with you?	هل انه دراسة معك.	هل يدرس معك.
University students have written their exams.	قد كتبت طلاب جامعه أمتحاناتهم.	قد كتب طلاب الجامعه أمتحاناتهم.
When the sun rises in the morning?	متى ترتفع الشمس في الصباح.	متى تشرق الشمس في الصباح.
They had washed their clothes with water.	كانوا قد غسلوا ملابسهم مع ماء.	كانوا قد غسلوا ملابسهم بالماء.
That boy does not want to run.	ذلك لا يريد ولد ان يركض.	لا يريد ذلك الولد ان يركض.
Has he decided to visit the museum?	قرر زر المتحف.	هل قد قرر يزور المتحف.
My mother will give the old man a coat	ستمنح أمي الرجل القديم معطف.	ستمنح أمي الرجل العجوز معطفا.
Will the new bridge be inaugurated by the king?	هل Inaugurated الجسر الجديد بواسطة الملك؟	هل سيفتتح الجسر الجديد بواسطة الملك؟

Our system scored an average 0.7143 in BLEU method, 0.8734 in METEOR and 0.8928 in F-measure.

11. Conclusions

The developed translation system is capable of translating English sentence to Arabic language using rule based and ANN method. The system is able to handle simple English sentences (assertive, interrogative, imperative) with the forms of (affirmative, negative) for both voices (active, passive). The system is able to translate sentences having gerunds, infinitives, prepositions and prepositional objects, direct objects, indirect objects, etc. The output is the corresponding Arabic translation of the English sentence.

The working and architecture of our English to Arabic Machine translation system is discussed in this paper. All the modules have been implemented successfully. This paper describes the use of neural network with rule based machine translation approach. Our system uses neural network for dictionary lookups. Every English input to the neural network has one output containing its meaning in Arabic, in one or more forms, in addition to some information attached with the words which makes it efficient and fast. Our system works efficiently on the sentences under study and words which are available in the neural network. If the word is not present in the dictionary, English word is printed as it is in the translation in capitals. The translation results obtained from the system evaluated using machine evaluation methods show that the system works. The n-gram BLEU score obtained for the system is 0.6029; METEOR score achieved is 0.8221 and F-score of 0.8386.

References

[1] Abu-Shquier M. and Sembok T., "Word Agreement and Ordering in English-Arabic Machine Translation," *in Proceeding of the International Symposium on Information Technology*, USA, pp. 1-10, 2008.

[2] Abu-Shquier M., Al-Nabhan M., and Sembok T., "Adopting New Rules in Rule-Based Machine Translation," *in Proceedings of the 12th IEEE Computer Society International Conference on Computer Modelling and Simulation*, Washington, USA, pp. 62-67, 2010.

[3] Abdulraheem E. and Ab-Aziz M., "English to Arabic Machine Translation Based on Reordring Algorithm," *Journal of Computer Science*, vol. 7, no. 1, pp. 120-128, 2011.

[4] Ben-Othmane Z., Aroua T., and Ben-Ahmed M., "A Multi-Agent System for POS-Tagging Vocalized Arabic Texts," *the International Arab Journal of Information Technology*, vol. 4, no. 4, pp. 322-329, 2007.

[5] DeYoung T., "Arabic Language & Middle East/ North African Cultural Studies," available at: http://www.indiana.edu/ ~arabic/ arabic_history. htm, last visited 2011.

[6] Elming J., "Syntactic Reordering in Statistical Machine Translation," *PhD Thesis*, Copenhagen Business School, Denmark, available at: http://openarchive.cbs.dk/bitstream/handle/10398/ 7922/jakob_elming.pdf?sequence=1, last visited 2008.

[7] Homiedan A., "Machine Translation," *Journal of King Saud University (Languages and Translation)*, Saudi Arabia, vol. 10, pp. 1-21, 1998.

[8] Hutchins W. and Somers L., *An Introduction to Machine Translation*, Academic Press, London, 1992.

[9] Ibrahim M., "A Fast and Expert Machine Translation System Involving Arabic Language," *PhD Thesis*, Cranfield Institute of Technology, UK, 1991.

[10] Imperial N., Koncar N., and Guthrie D., "A Natural Language Translation Neural Network," *in Proceedings of the International Conference on New Methods in Language Processing*, pp. 71-77, 1994.

[11] Jain A., "Parsing Complex Sentences with Structured Connectionist Networks," *Neural Computation*, vol. 3, no. 1, pp. 110-120, 1991.

[12] League of Arab States, "Arab Countries, Figures and Indicator," Third Addition, 2011.

[13] Khan S. and Mishra R., "Translation Rules and ANN Based Model for English to Urdu Machine Translation," *INFOCOMP Journal of Computer Science*, vol. 10, no. 3, pp. 36-47, 2011.

[14] Manning C. and Hinrich S., *Foundations of Statistical Natural Language Processing*, MIT Press, Cambridge, MA, 1999.

[15] Mishra V. and Mishra R., "Ann and Rule Based Model for English to Sanskrit Machine Translation," *INFOCOMP Journal of Computer Science*, vol. 9, no. 1, pp. 80-89, 2010.

[16] Mokhtar H., "An Automatic System for English-Arabic Translation of Scientific Text (SEATS)," *Master Thesis*, Cairo University, Egypt, 2000.

[17] Papineni K., Roukos S., Ward T., and Jing W., "BLEU: A Method for Automatic Evaluation of Machine Translation," *in Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Stroudsburg, USA pp. 311-318, 2002.

[18] Pease C. and Boushaba A., "Towards an Automatic Translation of Medical Terminology and Texts into Arabic," *Technical Document*, Translation in the Arab World, King Fahd Advanced School of Translation, Morocco, 1996.

[19] Rafea A., Sabry M., El-Ansary R., and Samir S., "Al-Mutarge: A Machine Translator for Middle East News," *in Proceedings of the 3rd International Conference and Exhibition on Multi-lingual Computing*, 1992.

[20] Satanjeev B. and Alon L., "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments," *in Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/ or Summarization*, Ann-Arbor, Michigan, USA, pp. 65-72, 2005.

[21] Shaalan K., Rafea A., Moneim A., and Baraka H., "Machine Translation of English Noun Phrases into Arabic," *the International Journal*

of Computer Processing of Oriental Languages, vol. 17, no. 2, pp. 121-134, 2004.
[22] Shaalan K., "Rule-Based Approach in Arabic Natural Language Processing," *the International Journal on Information and Communication Technologies*, vol. 3, no. 3, pp. 11-19, 2010.
[23] Stanford_University_Project, available at: http://nlp.stanford.edu/ software/ lex-parser.shtml, last visited 2011.
[24] Stanford_University_Project, available at: http://nlp.stanford.edu/software/tagger.shtml, last visited 2011.
[25] Trujillo A., *Translation Engines: Techniques for Machine Translation*, Springer Verlag, Berlin, 1999.
[26] Waibel A., Jain A., McNair A., Saito H., Hauptmann A., and Tebelskis J., "A Speech to-Speech Translation System using Connectionist and Symbolic Processing Strategies," *in Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*, USA, vol. 2, pp. 793-796, 1991.
[27] Zantout R. and Guessoum A., "An Automatic English-Arabic HTML Page Translation System," *Journal of Network and Computer Applications*, vol. 24, no. 4, pp. 333-357, 2001.
[28] Zughoul M. and Abu-Alshaar A., "English/ Arabic/ English Machine Translation: A Historical Perspective," *Translators' Journal*, vol. 50, no. 3, pp. 1022-1041, 2005.

Marwan Akeel received his BSc in software engineering from Budapest University of Technology and Economics, Hungary, in 2003, and his Master of Computer Applications from Bangalore University, India in 2008. Since July 2010, he has been pursuing his PhD at the Department of Computer Engineering, IIT-Banaras Hindu University. His research interest include natural language processing and machine translation.

Ravi Mishra received his BSc in electrical engineering from BIT Sindri and M.Tech. in control system from IT-BHU and his PhD in AI in medicine from IT-BHU. Dr. Mishra has 33 years teaching experience and presently serving as a professor and head in the Department of Computer Engineering, IIT-BHU. His areas of research interest include artificial intelligence a multiagent systems and their applications in medical computing, machine translation, e-commerce, robotics and intelligent tutoring systems. He has authored around 150 papers in journals and conferences.