Journal of Computer Science 3 (6): 410-418, 2007
ISSN 1549-3636
© 2007 Science Publications

Pre-editing and Recursive-Phrase Composites for a Better English-to-Arabic Machine Translation

Mansoor Al-A'ali
Department of Computer Science, College of Infromation Technology, University of Bahrain
P.O. Box 32038, Kingdom of Bahrain

Abstract: This research presents an approach for an English-to-Arabic Machine Translation System based on Building correct grammar and phrase structures first and then automatically deriving Translation Rules for phrase translation. For every English phrase, the grammar is first analysed and then a corresponding Arabic translation is given which would be used by the machine learning system to produce a translation rule with the help of a dictionary and the user. These same derived rules can partially be used for other phrase sequences especially in the case of a phrase consisting of a number of smaller phrases and thus implemeting the idea of recusive phrase strucutres. The approach was implemented and tested on simple cases and the results are given which indicate that this approach is successful for small to medium phrases. Our approach is an enhancement on existing phrase translation techniques because it analyses the source language grammar first, then builds a syntactic structure before proceeding with the machine learning process of learning the translation rules. Our approach is enhancement on existing phrase based translations in two directions: the grammar editing before the translation rules and the derived translation rules can be complete for complete phrases or are rules for translating smaller phrases which are subsets of larger phrases. The approach has improved the speed and correctness of phrase translations.

Key words: Translation, english-arabic translation, translain rules

INTRODUCTION

Machine Translation techniques have recently achieved some success but in spite of this success, MT still has many hurdles to overcome. Recent research contributed to the basic principles of machine translation systems by using phrase translation[1-3]. Most phrase-based machine translation systems rarely use linguistic knowledge of the structure of the languages involved. At the same time, the cost for using these techniques is high relative to the small improvements gained in performance. Some researchers attempted to employ parsers for tree-to-tree and tree-to-string alignment, respectively. Various experiments have been attempted on the effect of varying amounts of morphosyntactic information, including the techniques used. Linguistic knowledge from NLP tools can be used effectively with relatively small training data and in a limited domain. The problem is how reliable and accurate the knowledge is, and how to employ it.

A major challenge in machine translation is how to build phrase translation which takes into consideration all aspects of the translation process such as semantics and context. There are normally three different methods to build phrase translation probabilities: learning phrase alignments from a word-aligned corpus, learning syntactically motivated phrase pairs from a word aligned and parse tree annotated corpus, and learning a joint phrase model. Phrase pairs consistent with a word alignment has so far yielded the highest performance[1,4]. Syntactic phrase pairs can be restricted by both the word-alignments and the source and target parse trees.

Various experiments exploiting syntactic features in Chinese-to-English translation for example, were proposed in[2,5]. They examined the effect of using tags and syntactic chunks, and treebank-based syntactic parses of source and target sentences within an n-best re-scoring framework based on a log-linear model. Chinese sequences were projected onto the English words using the word alignment. Relative positions were indicated for each Chinese tag. Then a trigram language model was built on these projected Chinese tags and positions. The projected information was one of the most beneficial syntactic features used. Among the syntactic features evaluated, the simple Markov model achieved the most significant performance improvement. They take the approach of building a language model with the morphosyntactically enriched word sequences and interpolate the language model

Corresponding Author: Mansoor Al-A'ali, Department of Computer Science, College of Infromation Technology, University of Bahrain, P.O. Box 32038, Kingdom of Bahrain

with a class-based language model to overcome the problem of data sparseness. The classes are learned from the enriched word corpus by clustering with respect to the context. They have shown that using a conventional dictionary is useful for improving the alignment performance.

A method for improving the alignment performance by adding common word pairs extracted from the asymmetrically learned tree-to-string translation models was presented in[6]. They address the issue of enlarging the corpus using only words, but not chunks. They reported that adding the common word pairs extracted from the translation models contributed to the improvement of the alignment performance. However, since they simply add all of the extracted common word pairs to the training corpus without any validation, they could introduce many erroneous translation pairs that will degrade the translation probability distribution.

The practicality of NLP techniques, and use a base phrase chunker to perform chunk alignment based on the word alignments was considered in[3]. They define a phrase as a word sequence that is covered by a base phrase sequence, not by a single sub-tree in a syntactic parse tree. They make chunk alignment without using base phrase label and extract all phrase translation pairs consistent with the word alignment. Since full parsers often have phrase attachment errors, using the chunker may be more robust than using the decomposed sub-tree pairs and enable to reduce the loss of valuable phrase translation pairs by relaxing the strong syntactic constraint for the phrase alignment. For the purpose of reducing the translation ambiguities and generating grammatically correct and fluent translation output, they address the use of shallow linguistic knowledge, that is: (1) enriching a word with its morphosyntactic features, (2) obtaining shallow linguistically-motivated phrase pairs, (3) iteratively refining word alignment using filtered phrase pairs, and (4) building a language model from morphosyntactically enriched words. Previous studies reported that the introduction of syntactic features into MT models resulted in only a slight improvement in performance in spite of the heavy computational expense; however, this study demonstrates the effectiveness of morphosyntactic features, when reliable, discriminative features are used. Their experimental results show that word representations that incorporate morphosyntactic features significantly improve the performance of the translation model and language model. Moreover, they show that refining the word alignment using fine-grained phrase pairs is effective in improving system performance.

A phrasal lexicon to supplement a small training corpus was presented in[7]. The phrases in the lexicon were added to the training corpus as well as used during phrase translation. The phrasal lexicon consisted of a list of English phrases and their translations into Spanish/Catalan. They utilized morphosyntactic information for semi-automatically constructing and extending the phrasal lexicon, especially for verbal phrase expressions. Since they were extracted partly from various corpora/web-sites and partly created manually, it was expensive to build the lexicon. Therefore, in this study, they tried to automatically extract the chunk translations from the word-alignment by utilizing shallow syntactic information. Since the shallow parsing and the word alignment are not 100% accurate, they filter out the unreliable chunk translations by means of a statistical test at a certain confidence level. They then enlarge the training corpus by adding the most reliable chunk pairs. Then they iteratively train the word alignment on the enlarged training corpus. They expect that the added reliable chunk correspondences will play a role of boosting the accuracy of the unsupervised word alignment process.

Finally, the concept of a decoding algorithm was implemented whereby the decoder's job is to find the translation that is most likely according to a set of previously learned parameters (and a formula for combining them)[6]. Since the space of possible translations is extremely large, typical decoding algorithms are only able to examine a portion of it, thus risking to miss good solutions. Unfortunately, examining more of the space leads to unacceptably slow decodings. Our approach is based on recursive phrase based translations which takes into consideration a number of factors:

- The grammar of the source (English) and target (Arabic) languages.
- Grammar checking is achieved by previously stored structures and by the user intervention for new or partly new structures.
- Editing of the source language grammar if needed before the translation process is activated.
- The use of a dictionary for direct translation is only used during the learning process which is used to derive the translation rules.
- Phrase translation is a combination of direct translation and a user entered semantically correct translation.
- Translation rules are generated for smaller phrases first and then for larger phrases consisting of one or more of the phrases and thus implementing the concept of recursive phrases.
- The number of possibilities for translations is reduced by first checking the existing rules database for matching structures and phrases for composite phrases first and if that does not exist then for their smaller sub-phrases.

- The translation is based on the grammar structure, the semantics, the phrase translation and the dictionary.
- Extracting phrase translation pairs using recursive phrase based information.

English-To-Arabic Machine Translation: The concept of English-to-Arabic machine translation in particular received limited attention in recent years[8,12]. There are some English-to-Arabic translation systems such as Alwafi and Almutarjim, but they are at the beginners' stage when compared with the available translation systems for translating between European languages[13]

One aspect of translation is linguistic difficulties of automatic translation[14]. Some of the other difficulties of Arabic-English automatic translation were reported in[11]. Amongst the general difficulties of translation is morphological analysis, which includes: Homonyms or Homographs, Polysemy, Homophones, Idioms, Prefixes and suffixes. Another problem is the semantic structure, which requires the understanding of the general meaning of the context. Other difficulties include the meaning of phrases, the indication of an omitted part and the effect of pronouns in the sentence and between sentences. An example of difficulties in English to Arabic automatic translation is pronouns references in the sentence and between various sentences. For example, the sentence 'Ahmed has two books’ can be translated into Arabic as follows: ‘ ‘أحمد يملك كتابينor ‘‘أحمد لديه كتابين. If the word 'two' is replaced by 'three' then the Arabic pronoun, which refers to ‘two’ would be changed to refer to the plural, and the Arabic translation would be as follows: ‘ ‘أحمد يملك ثلاثة كتبor ‘‘كتب أحمد لديه ثلاثة. The change is affecting the word 'book' and not the number of books.

Neglecting accentuation may cause ambiguity in understanding the meaning of the word. For example, the Arabic word 'كتب' can have the following forms: ' كتب' means wrote, ' كتب' means written, or ' كتب' means books. The way we would read this word depends on its position in the sentence or on the semantics of the sentence or based on some special symbols placed on top or below each letter, which indicates its grammar status, which in turn is based on semantics. These symbols are not normally written but are derived during reading, yet they influence the semantics of the sentence. The word ' كتب' written on its own could mean: 'he wrote', 'was written' or 'books' or 'asked for it to be written'.

The use of machine translation as a tool for professional and skilled translation still remains for the most part limited and requires a great deal of post editing. The achievements so far do not match the efforts exerted in developing professional MT systems. Examples of recent good attempts to produce such products are: ArabTrans, Arab Translator and Al Wafi. In translation, the system must distinguish several kinds of linguistic knowledge, the Phonological knowledge, syntactic knowledge, Pragmatic knowledge, and Discourse knowledge.

Our approach as implemented in TARJEM can understand the meaning of the word by putting it in a grammar sequence and by learning more examples from the user. For example the article always comes at the begnning of the sentence and before the subject or the object but never comes before a verb. For example, The sentence: 'The boy ate an apple' has the grammer sequence 'Article, Masculine Single Subject Noun, Irregular Past Verb, Article, Masculine Single Object Noun'. By learning from more examples, Tarjem can understand that the subject noun can never be an object, and the past verb can not happen in the present time.

The most difficult issue in translation is the understanding of semantics. Tarjem gives each English word a Grammar Type, to make the computer understand it at least grammatically and can classify it. For example, the word 'boy' represents 'Masculine Sngle Subject Noun'. In this case, 'The boy ate an apple' is correct, but 'the apple ate a boy' would be semantically incorrect. However, TARJIM derives the semantics translation rules from the example based concept of previous instances of sentences and this puts the source and target language sentences in grammar context, semantic context and thus in syntactic context.

Pragmatic means the understanding of words in situations and context. At this level, TARJEM has the ability to learn and then understand the meaning of the word in certain situations and context, because when the word is repeated frequently with certain Grammar Types of words and in a certain position of the grammar sequence, the system builds its rules with the help of the user on those concepts.

Grammar Checking Before English-To-Arabic Translation: In TARJEM, the user is led by the system step-by-step “in the Grammar stage" to build the entered sentence grammatically before shifting to the translation stage. The Grammar stage checking is important and helpful to avoid any grammar errors leading to bad translation. Checking the grammar of the sentence before translation prevents the translation system from building wrong translation rules that may cause translation problems in the future. The advantage of allowing the user to share the translation system building the grammar and translation rules is to make the translation system learn from the expert user and as a result, the translation system avoids building wrong translation or grammar rules.

The rule extrating technique based on the example based technique used in TARJEM will parse a sentence, usually creating an intermediary, symbolic representation (Type-Word), from which it then generates a sentence in the target language. This approach requires extensive lexicons with morphologic, syntactic, and semantic information (Dictionary table and Grammar Rules Table), and large sets of rules (Translation Rules Table).

Profound knowledge of the grammatical rules that govern the source and target languages is essential. These are the analytical tools that we need to correctly disassemble a text in one language and reassemble it from scratch in another language. In fact, English and Arabic are not so 'relatively' similar to each other and this would force us not to skip the complete process of disassembly and reassembly. Lexical knowledge of the source and target languages and the complicated relationships between the two lexicons is another pillar that we can't do without. There is no one-to-one relationship between source and target term and/or phrase. An example 'run', has different meanings in arabic, if it is a noun (ركض, مرة, اندفاع, اجل) and if it is a verb (ركض, أدار, جرى). Understanding language structure is very important in translation by a machine, because a computer needs Grammar Rules And Translation Rules in order to work properly in the translation processes.

The concept of post editing (PE) is used to edit, modify and/or correct pre-translated text that has been processed by a machine translation system from a source language into a target language. The notation of post editing system has been adequately defined as the term used for the correction of machine translation output by human linguists is that it is done after the translation and there is no pre-editing in the system. Thus post editing becomes difficult without 'preparing' the entered sentence before it is entered to the system. In TARJEM, the pre editing and post editing are done during the translation processes and that the context, syntax and semantics are learnt from the user.

Translation rules can automatically be derived from example translations from English to Arabic through the Artificial Intelligence technique of machine learning. These derived rules are later used for automatic translation. Table 1, illustrates an example of the pre-editing concept.

Translation rules can automatically be derived from example translations from English to Arabic through the[10] Artificial Intelligence technique of machine learning . These derived rules are later used for automatic translation. In TARJIM, the user is led by the system step-by-step “in the Grammar stage" to build the entered sentence grammatically before shifting to the translation stage. The Grammar stage checking is important and helpful to avoid any grammar errors leading to bad translation. Checking the grammar of the sentence before translation prevents the translation system from building wrong translation rules that may cause translation problems in the future. The advantage of allowing the user to share the Translation System building the Grammar and Translation rules is to make the Translation System learn from the expert user and as a result, the Translation System avoids building wrong Translation or Grammar Rules.

The rule-based technique, used in TARJIM will parse a sentence, usually creating an intermediary, symbolic representation (Type-Word), from which it then generates a sentence in the target language. This approach requires extensive lexicons with morphologic, syntactic, and semantic information (Dictionary table and Grammar Rules Table), and large sets of rules (Translation Rules Table). Profound knowledge of the grammatical rules that govern the source and target languages is essential. These are the analytical tools that we need to correctly disassemble a text in one language and reassemble it from scratch in another language. In fact, English and Arabic are not so 'relatively' similar to each other and this would force us not to skip the complete process of disassembly and reassembly.

Lexical knowledge of the source and target languages and the complicated relationships between the two lexicons is another pillar that we can't do without. Even less than in the grammatical/syntactical world, we know that there is no one-to-one relationship between source and target term and/or phrase. An example run, it has different meanings in arabic, if it is a noun (ركض, مرة, اندفاع , اجل) and if it is a verb (ركض, أدار جرى). Understanding language structure is very important in translation by a machine, because a computer needs Grammar Rules And Translation Rules in order to work properly in the translation processes.

The Tarjim Machine Translation Approach: The overall TARJIM system is ullustrated in Fig. 1, which shows that TARJIM consists of two main processes; these are the Grammar Analysis process

Table 1: TARJIM

The entered sentences	Pre editing (Grammar checking)	Translation	Remarks
The boy is eaten my apple	The boy has eaten my apple or The boy is eating my apple	أكلَ الولدُ تفاحُتي أو الولدَ يَأكلُ تفاحَتي	The grammar of the entered sentences is not correct, so the translated sentence is meaningless and too difficult to edit. Jeff Allen approach will stop here and translation will be meaningless.

and the translation Analysis Process. These two processes are further elaborated in Fig. 2 and Fig. 3.

Figure 2 describes the Grammar Analysis process which consists of four sub-processes. When an English sentence is typed and entered to TARJIM by the user, it will be split into its words, then put into an array. The English words that are received from process (1.1) are sent by the system to the next process (1.2) to generate an English word Query. At this process, the system creates a query for each English word and sends all the queries to the next process (1.3). The aim of the query is to check each word of the English sentence if it is new or exists in the database (English Type Dictionary). At process (1.3), the system checks with the database (English Type Dictionary) to retrieve the English Grammar Type of the first English word, then the second, and so on to the end of the English sentence. If the word is not found in the database (English Type Dictionary), the system writes the word into the database, at the same time asks the user to enter it's Grammar Type and Arabic translation in a loop until it reaches the last word of the English sentence. After that the sequence of Grammar Types are sent to process (1.4) to check or generate a grammar rule. At process (1.4) the system checks the recieving grammar sequence from process (1.3) with another database (Grammar Rule Table). If the grammar sequence exists, then it is given a new grammar rule name by the system. But if the sequence of the Grammar Type does not exist in the database (Grammar Rule Table), the user is asked by the system to add a new grammar name or select from a list the grammar names. The grammar rule name and its grammar sequence is saved in the database (Grammar Rule Table). Next time if the user enters an English sentence which has the same sequence of Grammar Types, he will not be asked by the system to name it, because the system will recognise it.

Figure 3 describes the Translation Analysis process which consists of two sub-processes. Process (2.1) recieves the English sentence that hes been checked Grammaticaly in the previous process (1.4). At the current process, the system retrieves from the database (English Word-Type) the Grammar Type of the current English word (1) and the next English word (2). The system then sends the Grammar Type of the word (1) and word (2) to process (2.2). At process (2.2) the system retrieves the Arabic translation word of English word (1) and English word (2), then generates a relationship between the Arabic two words in an indirect way. This realshionship is created by using the English Grammar Types that are equivalent to the two Arabic translation words. The translation rules are imported from the database (Translation Rules). The loop is contionus between the current process and the previus process to genrate a relationship between the Grammar Type of word (2) and word (3) till the end of the sentence. After compliting the sentence, it is displayed as an Arabic sentence to the user.

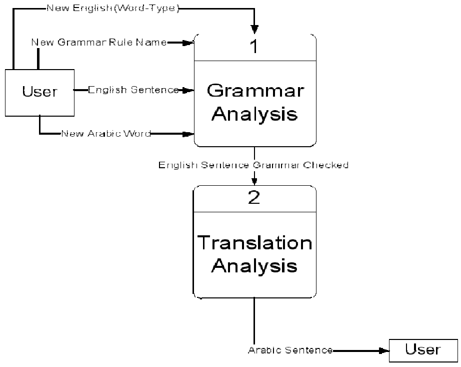

Fig. 1: The main TARJIM processes

TARJIM method for learning translation rules: The system will ask the user to enter a new sentence.

The system will take each word in a sentence (the boy went to school) and search for it in the main dictionary to find it. Therefore, 'The boy went to shool' consists of (Article, Human Single, Past Verb, Proposition, Thing Single).

Then the system will open the word's record to take it's type, meaning and other information.

The system will display the information of the sentence on the screen.

The system will ask the user to enter the correct human Arabic translation sentence.

Enter the Arabic translation: (*ذهب الولد إلى المدرسة*).

The system will break the translation sentence into it's words.

The system will break the sentence as follows: ذهب (*ال ولد إلى ال مدرسة*)).

The system will identify each sub-sentence or phrase and check if it's rule is stored in the Rules file, if it is found, then the system will take it's index number, otherwise it will generate the rule by matching the translation sub-sentence with the Arabic meaning that were taken from dictionary for this sub sentence and the translation given by the user.Finally the rules will be stored in the rule's file.

Example 1: The phrase: The boy
The final translation : الولد

Word	the	boy
Type	Article	human single
Translation	ال	الولد
Sequence	1	2

The Rule: Article noun-human-single J 1 2 .
This means, the translation of this phrase to Arabic is:
Start with the Arabic Article
Then insert the Arabic noun-human-single
Then join the Article with the human single noun in the sequence: Article then the noun.

Example 2: The phrase : His school
The final translation : مدرسته

Word	his	car
Type	Pronoun	Thing-single
Translation	ه	مدرسته
Sequence	1	2

The Rule : Prounoun thing_single J 2 1 .
This means, the translation of this phrase to Arabic is:
Start with the Arabic pronoun
Then insert the Arabic thing-single
Then join the Arabic pronoun with the thing single in the sequence thing-single then the Arabic pronoun.

Example 3: The phrase: The boy went to his school
The direct Arabic translation : إلـى هـ مدرسـة ال ولـد ذهـب
The final translation : ذهـب الولـد إلـى مدرسـته
Partial phrase combinations: (The boy الولـد) (went ذهـب) (to إلـى) (his school مدرسـته)
The system will split the longer phrase to sub-phrases, each one has it's own rule, and passes each one to the translation function to translate it to its corresponding rules. The rules for the different sub-phrases or phrases used for translating longer phrases consisting of smaller phrases for which translation rules already exist, Table 2. The same example phrase can be learnt and stored in a different set of sub-phrases as follows:
(The boy went) (to his school) or (The boy) (went to) (his school) or (The boy went to) (his school) or
(The boy) (went to his school)

Table 2: Rules for phrases making up a longer sentence – a set of sub-phrases

English sub sentences	Sequence No.1	Sequence No.2	Rule Index	The Rules	Arabic Translation
He went	He	went	1	prn Past_verb S 1 2.	هو ذهب
his school	his	school	2	prn thing_s J 2 1.	مدرسته
to his school	to	his school	3	prop r2 S 1 2 .	إلى مدرسته
He went to his school	He went	To his school	4	r1 r3 S 1 2 .	هو ذهب إلى مدرسته

TARJIM translation method

- The system will ask the user to enter an English sentence.
- The system will take each word and search for it in the main dictionary to find it's record.
- Then the system will open the word's records to take it's type, meaning and other information and store them.
- The system will display the information of the sentence on the screen.

The English sentence : The boy reads his book
ال ولـد يقرأ هـ كتاب
The system will break the sentence into sub-sentences which make up the whole sentence, eg, 'the boy', 'his book' which would already have existing rules from a previous traninig exercise.
The system will take each sub sentence and search for it's rule, then translate it according to it's rule and the stored rule's index.

Example 1: The phrase: 'The boy'
Rules File

Rule 1) Article hum_single J 1 2 .

Rule 2) 1 Pronoun S 1 2 .

Rule 3) Pronoun thing_s J 2 1 .

Rule 4) r2 r3 S 1 2 .

The rule will translate it as : (The boy) (Article human_single)
ال ولـد
Then the next part of the rule: (J 1 2) which means join the two words in reverse order which produces 'The boy' translated to (الـولـد).

Example 2: The phrase: '((The boy) (reads))'.
Rules File

1) ART human_single J 1 2 .

2) 1 Pronoun S 1 2 .

3) Pronoun thing_single J 2 1 .

4) r2 r3 S 1 2 .

A rule already exists for translating the sub-phrase 'The boy' and hence will be translated using Rule 1 (r1) to (الـولـد). 'The boy reads' consists of two sections, 'The boy' and 'reads'. A rule already exists for 'The boy'. Another rule exists for 'The boy' and 'reads' as a rule by itself, ie, one rule consisting of another rule and a word. This means that we have a recursive phrase consisting of two ssub-phrases. 'reads' translates into ' يقرأ'. Rule 2 (r2) had been extracted for such a phrase. A new rule is now generated for this sentnece.

Example 3: The phrase: 'his book'

Rule's File

1) Article hum_single J 1 2 .

2) 1 Pronoun S 1 2 .

3) Pronoun thing_single J 2 1 .

4) r2 r3 S 1 2 .

The rule will translate it as : (his book) (Pronoun thing_s)

(كتاب هـ)

Then the next part of the rule is (J 2 1) where 2 1 means join the two words in reverse sequence which produces the translation (كـتـابـه) for the phrase 'his book'

Example 4: The sentence 'The boy reads his book'

Rule's File

1) Article hum_single J 1 2 .

2) 1 Pronoun S 1 2 .

3) Pronoun thing_s J 2 1 .

4) r2 r3 S 1 2 .

(The boy reads) (his book) Rule 4 (r2 r3 S 1 2)

(كتابـه) (يقرأ الولد)

The translation of the example phrase and the use of rules is determined by the set of existing rules for the corresponding sub-phrases. TARJIM reads left to right and will identify the rule for each phrase as it is composed, but if it reads on and identifies a longer phrase containing the previous sub-phrase then the rule for the longer phrase will be used. The possibilities are:

(The boy reads) (his book)

(The boy) (reads his book)

(The boy) (reads) (his book)

At the end of the translation process, the new example itself will be stored as a new structure and a new rule and thus when it is met in the future it will be considered and not its sub-phrases. This approach guaranted that sub-phrases do not give wrong semantics as they could be out of context. This is a major contribution of the TARJIM approach.

TESTING AND RESULTS

The rules shown in Table 2 are a sample of the rules, which have been derived by the machine learning system. This machine learning system works by being fed with English sentences, their corresponding Arabic translations and a dictionary. Therefore, there is no limit as to how many rules that can be produced by the system. Once a rule is produced for any type of sentence, then this rule can be used for translating matching structures. Rules are generated for simple phrases, eg, 'his apple', he went', 'she ran' as well as for longer sentences, eg, 'she ate her apple yesterday'. The rules for longer sentences can be made up of a combinatin of smaller rules for smaller phrases.

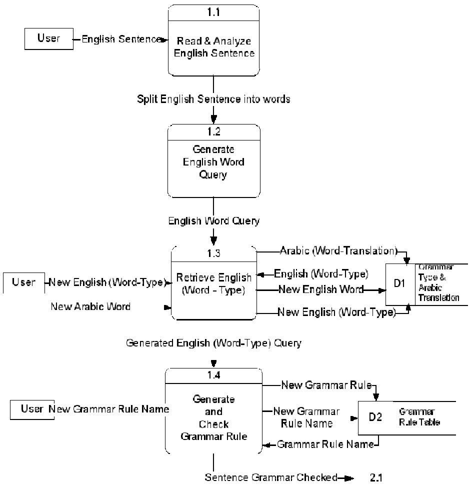

Fig. 2: Grammar checking phase

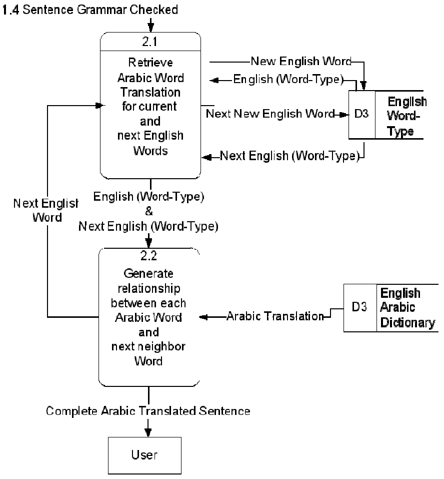
1.4 Sentence Grammar Checked
2.1
Retrieve Arabic Word Translation for current and next English Words
New English Word
English (Word-Type)
Next New English Word
D3
English Word-Type
Next English (Word-Type)
Next English Word
English (Word-Type) & Next English (Word-Type)
2.2
Generate relationship between each Arabic Word and next neighbor Word
Arabic Translation
D3
English Arabic Dictionary
Complete Arabic Translated Sentence
User

Fig. 3: Translation phase

By direct translation we mean the translated words as given by the user in conjunction with the translation of the words in the dictionary. Of course the word translation from the dictionary will have many alternatives and without tying these different translations with the user's given translations, we would not have a proper trnsalation since we would not know which is the appropriate translation we would choose.

Finally, we can show the results of translating part of the beauty and the beast story as produced by TARJIM. The same sentences were fed to Al-Wafi, the most popular English-to-Arabic translation system available today. We can now selectively compare the two translations in order to demonstrate that the TARJIM approach and the method of deriving rules can prove to be more effective in producing a more realistic and near human translation in most cases.

Consider the stroy of the beauty and the beast. Consider the phrase 'Gaston lost his balance' which has the structure (Animate verb pronoun noun). The direct translation of this sentence is (جاستون فقد ه توازن) .

We add (ه) to the noun (توازن) by using (Rule 7 - R7), to be (جاستون فقد توازنه). Now put the verb (فقد) first in the sentence by using (R2) to be (فقد جاستون توازنه). Al-Wafi translation of 'his balance' to be (ميزانه) which is incorrect because this is the translation of 'his scales' as in weighing scales. TARJIM translation is more accurate (توازنه)

Consider the example 'But the beast didn't answer' which has the structure (Conjunction article noun verb). The direct translation of this sentence is (لكن ال وحش لم يجيب). Now add (ال) to the noun (وحش) by using (Rl) to be (لكن الوحش لم يجيب).

Consider the example 'She was gathering her strength ' which has the structure (Pronoun verb verb pronoun noun). The direct translation of this sentence is (هي كان تستجمع ها قوه). Now delete (هي) from the sentence by using (R9) to be (كان تستجمع ها قوة). Add (ت) to the verb (كان) by using (R25) to be (كانت تستجمع ها قوة). Add (ها) to the noun (قوه) by using (R17) to be (كانت تستجمع قوتها). Al-Wafi translation of 'She was gathering her strength' was (هي كَانتْ تَستجمعُ قوها) which is incorrect because the word (قوها) here is semantically meaningless. TARJIM gave (قوتها) meaning her strength, which is semantically correct.

Consider the example 'They both remained silent ' which has the structure (Pronoun noun verb adjective). The direct translation of this sentence is (هم اثنان بقي صامت). Now delete (هم) from the sentence by using (R9) to be (اثنان بقي صامت). Add (ال) to the noun (اثنان) by using (Rl) to be (الاثنان بقي صامت). Put the verb (بقي) first in the sentence by using (R2) to be (بقي الاثنان صامت). Add (ان) to the adjective after (both) by using (R24) to be (بقي الاثنان صامتان). Al-Wafi gave the translation of the sentence ' They both remained silent' as (كلاهما بَقيا صامتَ) which is weak and gramatically incorrect translation compared to TARJIM which was (بقي الاثنان صامتان).

The sentence 'The map was upside-down' which has the Structure (Article noun verb adjective). The direct translation is (ال خارطه كان مقلوبه رأسا علي عقب). Add (ال) to the noun (خارطه) by using (R1) to be (الخارطه كان مقلوبه رأسا علي عقب). Put the verb first (كانت) using (R2) to be (كانت الخارطه مقلوبه رأسا علي عقب).

The sentence 'Good luck' has the structure (Adjective noun). The Direct translation (جيد حظ). Put the noun first (حظ) followed by the adjective (جيد) to be (حظ جيد). Al-Wafi translation of ' Good luck' was (الحظّ السعيد) which is really the translation of 'The good luck' which is a different meaning from the original sentence.

The sentence 'He is not for me' has the structure (Pronoun verb negation preposition pronoun). The direct translation is (هو يكون ليس لأجل ي). Delete (يكون) from the sentence by using (R9) to be (هو ليس لأجل ي). Add (ي) to the preposition (لأجل) by using (Rl6) to be (هو ليس لأجلي).

The example 'He saw the yellow eyes' has the structure (Pronoun verb article adjective noun). The direct translation (هو رأي ال صفراء عيون). Delete (هو) fron the sentence using (R9) to be (رأي ال صفراء عيون). Add (ال) to the adjective (صفراء) and to the modified noun (عيون) using (R5) to be (رأي الصفراء العيون). Put the modified noun (العيون) first followed by the adjective (الصفراء) using (R2) to be (رأي العيون الصفراء).

The example 'At first no one recognized him' has the structure (Preposition noun negation noun verb pronoun). The direct translation is (في بداية لا أحد لاحظ ه). Add (ه) to the verb (لاحظ) by using (R8) to be (في البداية لا أحد لاحظه).

The example 'Please help me' has the structure (Verb verb pronoun). The direct translation is (أرجوك ساعد ني). Add (ني) to the verb (ساعد) by using (R18) to be (جوك ساعدني).

The example 'Her answer was no' has the structure (Pronoun noun verb negation). The direct translation is (ها جواب كان لا). Add (ها) to the noun (جواب) by using (R19) to be (جوابها كان لا). Put the verb (كان) first in the sentence by using (R2) to be (كان جوابها لا).

Consider the example 'They watched the beast'. This has the structure (Pronoun verb article noun). The direct translation is هم شاهد ال وحش)). Delete (هم) from the sentence by using (R9) to be (شاهد ال وحش). Add (وا) to the verb (شاهد) by using (R26) to be (شاهدوا ال وحش). Add (ال) to the noun (وحش) by using (R1) to be (شاهدوا الوحش).

The sentence 'The beast replied sadly' has the structure (Article noun verb adjective). The direct translation is (ال وحش رد حزن). Add (ال) to the noun (الوحش) by using (R1) to be (الوحش رد حزن). Add (ب) to the adjective (حزن) by using (R20) to be الوحش رد بحزن))). Put the verb (رد) first in the sentence by using (R2) to be (رد الوحش بحزن). Al-Wafi translation of 'The beast replied sadly' was (أجابَ الوحشُ من المحزن) which actually means 'The beast answered from sadness' whilst TARJIM gave the correct translation رد الوحش بحزن)).

We have demonstrated through the TARJIM approach that a number of enhancements on current phrase based machine translations are possible.

The inclusion of the analysis of the source language grammar in the learning phase and in the translation phase is a way to guarantee that the derived rules are based on phrases which are syntactically and grammatically correct. These correct syntactic and grammatical structures are stored and thus learnt such that translation of future text is not just based on previously stored phrase sequences which can and may exist but can be out of context. During the translation process editing of the source language grammar is needed before the translation process is activated. If the system hits a new phrase or a new partial phrase then the user is given the chance to train the system by entering the correct grammar and translation.

The use of a dictionary for direct translation is only used during the learning process which is used to derive the translation rules. This saves time during the translation process. This is possible because the phrase sequences and the phrase structures are stored and are enough for the translation process.

Phrase translation is a combination of direct translation and a user entered semantically correct translation. Translation rules are generated for smaller phrases first and then for larger phrases consisting of one or more of the phrases. This approach is superior in that newer longer phrases are not treated as an alien new phrase but as a phrase which has translation for a number of its sub-phrases and only the translation and grammar of the extra new sub-phrase is needed. Further, our approach has the ability to generate the translation in parallel by identifying the sub-phrases and getting their translations. If a long phrase consists of a number of sub-phrases and recursively one of its sub-phrases is itself consistent of two or more sub-phrases, then if the translation and structure of the longer phrase already exists, the system directly extracts the longer phrase translation. This approach obviously contributes to the speeding up of the translation process. Thus, the number of possibilities for translations is reduced by first checking the existing rules database for matching structures and phrases based on the optimal minimum number of sub-phrases.

The translation is based on the grammar structure, the semantics, the phrase translation and the dictionary. Therefore, a word meaning depends on its position in the phrase and structure and may have different meaning in different positions and phrases.

CONCLUSION

This study presented an approach for English-to-Arabic machine translation. The approach was based on training the TARJIM system to accept English phrases, their structures, their Arabic translation and automatically derive the translation rules. The approach was implemented in a system called TARJIM and experimental results were produced. A set of phrases were then fed to TARJIM and compared to the currently leading English-to-Arabic translation software called Al-Wafi and the results presented clearly demonstrate that TARJIM produced better translations for most of the phrases. We demonstrated that the TARJIM approach is clearly an enhancement on existing approaches in that it allows pre-editing for the source language grammar before deriving the translation rules. Another enhancement is achieved by allowing larger phrases trnslations to consist of smaller phrase translations and thus improving the speed of translation. The use of recursive phrase structures where phrases can consist recursively of sub-phrases proved to a more effecient way of translation especially in terms of speed. If a sub-phrase already exists as part of a larger phrase then the larger phrase structure and translation is used.

REFERENCES

1. Koehn, P., F.J. Och and D. Marcu, 2003. Statistical phrase-based translation, Proc. Human Language Tech. Conf. (HLT/NAACL), Linguistics, 30: 181-204.
2. Och, F.J. and H. Ney, 2004. The alignment template approach to statistical machine translation, Computational Linguistics, 30: 417-449.
3. Hwang Young-Sook, Finch Andrew and Sasaki Yutaka, 2007. Improving statistical machine translation using shallow linguistic knowledge, Computer Speech and Language, 21: 350-372.
4. Marcu, D. and W. Wong, 2002. A phrase-based, joint probability model for statistical machine translation, Proceedings of EMNLP.
5. Och, F.J. and H. Ney, 2000. Improved statistical alignment models, Proc. 38th Annual Meeting of the Assoc. Computational Linguistics, 440-447.
6. Germann Ulrich, Jahr Michael, Knight Kevin, Marcu Daniel and Yamada Kenji, 2004. Fast and optimal decoding for machine translation, Artificial Intelligence 154: 127-143.
7. Popovic, M. and H. Ney, 2005. Exploiting phrasal lexica and additional morpho-syntactic language resources for statistical machine translation with scarce training data, Proc. 10th Annual Conf. European Assoc. Machine Translation, pp: 212-218.
8. Ahmad, T., Al-Taani and M.H. Eyad, 2005. A direct english-arabic machine translation system, Information Tech. J., 4: 256-261.
9. Papineni, K., S. Roukos, T. Ward, J. Henderson and F. Reeder, 2002. Corpus-based comprehensive and diagnostic MT evaluation: Initial Arabic, Chinese, French, and Spanish results, Proc. Human Language Conf.
10. Al-A'ali, M., 1998. Machine learning for machine translation: English to Arabic, Int. Conf. Multi-Lingual Computing ICEMCO, University of Cambridge, UK.
11. Ebrahim, M. A. and J.D. Clarke, 1990. Arabic-english machine translation of figurative expressions, Proc. Second Cambridge Conf. Bilingual Computing in Arabic and English, Cambridge, U.K.
12. Ebrahim, M. A., J.D. Clarke and A.A. Fahmv, 1989. Arabic machine translation proceedings of the seminar on bilingual computing in arabic and english, Univ. Cambridge, Cambridge, U.K.
13. María Calzada Pérez, 2005. Applying translation theory in teaching, new voices in translation studies, 1: 1-11.
14. Mohamed Abdel Fattah, Fuji Ren, Shingo Kuroiwa. 2006. Stemming to improve translation lexicon creation orm bitexts, Information, Proc.and Manag., 42: 1003 - 1016.