Handling Agreement and Words Reordering In Machine Translation from English to Arabic Using Hybrid-Based Systems

Mouiad Alawneh and Tengku Mohd,

National University of Malaysia, Faculty of Information Science and Technology, Bangi, Malaysia

Summary
Machine Translation has been defined as the process that utilizes computer software to translate text from one natural language to another. This definition involves accounting for the grammatical structure of each language and using rules, examples and grammars to transfer the grammatical structure of the source language (SL) into the target language (TL). This paper presents English to Arabic approach for translating well-structured English sentences into well-structured Arabic sentences, using a Grammar-based and example-translation techniques to handle the problems of ordering and agreement. The proposed methodology is flexible and scalable, the main advantages are: first, a hybrid-based approach combined advantages of rule-based (RBMT) with advantages example-based (EBMT), and second, it can be applied on some other languages with minor modifications. The OAK Parser is used to analyze the input English text to get the part of speech (POS) for each word in the text as a pre-translation process using the C# language, validation rules have been applied in both the database design and the programming code in order to ensure the integrity of data. A major design goal of this system is that it will be used as a stand-alone tool, and can be very well integrated with a general machine translation system for English sentences

Key words: *MT, Agreement, Word reorder, Hybrid-based, POS.*

1. Introduction

The current Machine Translation system facilitates the end user to understand the English textual sentences clearly by generating the precise corresponding Arabic language. Agreement is a basic property of language. In the most basic sense, agreement occurs when two elements in the appropriate configuration exhibit morphology consistent with their co-occurrence. Perhaps the most transparent case of this linguistic mechanism is number agreement between a subject and a verb: A singular noun in the subject position regularly co-occurs with a singular verb (e.g., "the dog runs"), and a plural subject noun regularly co-occurs with a plural verb (e.g., "the dogs run"). If the language has number marking on other elements, such as determiners or adjectives, these should also exhibit morphology that is consistent with their relationship to the subject head noun, and this co-occurrence relationship holds for gender and person agreement as well.

Manuscript received June 5, 2011
Manuscript revised June 20, 2011

The modern Arabic dialects are well-known as having agreement asymmetries that are sensitive to word order effects. These asymmetries have been attributed to a variety of causes, first, by the analysis problems at the source language, second, the generation problems at the target languages. However, Arabic is not alone in showing word-order asymmetries for agreement, Similar asymmetries have been documented in Russian, Hindi, Slovene, French and Italian (Hutchins and Somers 1992). Languages are varied in the agreement requirements. Some of them like Arabic require number, gender, person, and case agreements while others need some of these agreements. Machine translation system develops by using four approaches depending on their difficulty and complexity.

These approaches are: rule based, knowledge-based, corpus-based and hybrid MT. Rule-based machine translation approaches can be classified into the following categories: direct machine translation , interlingua machine translation and transfer based machine translation (Abu Shquier and Sembok, 2007). Our purpose of this paper is to design a hybrid-based (rule-based and example-based) framework based hence, to strike a balance between both approaches in the use of MT for the translation of texts and to handle the problem of word agreement and ordering in the translation of sentences from English to Arabic.

2. Grammars

In the grammars contains the English patterns and their equivalent Arabic patterns as well as some other information that will be used to apply the agreement rules and the reorder of the words in the produced Arabic translation.
Following the example explained the contents of this table and their purposes.

Example: That mad woman loved that crazy driver
Translation: احبت تلك المرأة الحمقاء ذلك السائق المجنون
[*ahbat telka almr`ah alhmka dalek alsae`g almjnoon*]

English Pattern	DT/1;JJ/2;NNX/3;VBX/4 ;DT/5;JJ/6;NNX/7	In this pattern the subject is preceded by a determinant and adjective, the verb is followed by an object which is preceded by a determinant and adjective
Subject	3	This means that the subject is the 3rd word
Main verb	4	This means that the main verb is the 4th word
Object	7	This means that the object is the 7th word
Verb agr.	¾	This means that we should handle the agreement between the 3rd word (NNX: subject) and the 4th word (VBX: main verb)
Adj. Agr.	2/3;6/7	This means that we should handle the agreement between the 2nd word (JJ: adjective) and the 3rd word (NNX: subject) and the agreement between the 6th word (JJ: adjective) and the 7th word (NNX: object)
Arabic Pattern	VBX/4;DT/1;XAL/1;NN X/3;XAL/1;JJ/2;DT/5;X AL/5;NNX/7;XAL/5;JJ/6	This represent the pattern in Arabic language, observe that the order is not just a reverse order of the English pattern, I used the symbol XAL to represent the addition of "AL alta'reef ال" before the nouns and adjectives depends on the category of the related determinant. The integers given here are not serial but they are the same integers given to each POS in the English pattern

3. Hybrid-Based MT Systems

The expansion of methodologies in the past decade and the introduction of new applications for automated translation processes have highlighted the limitations of adopting one single approach to the problems of translation. In the past, many MT projects were begun by researchers who saw MT as a test bed for a particular theory or particular method, with results that were either inconclusive or of limited application. It is now widely recognised that there can be no single method for achieving good-quality automatic translation, and that future models will be 'hybrids', combining the best of rule-based, statistics based or example-based methods.

Figer1 shows that Our approaches In RBMT, the process of conversion is based on the use of bilingual dictionaries and rules for converting SL (source language) structures into TL (target language) structures or using the dictionaries and rules for deriving intermediary representations from which output can be generated. Before analysis the input SL strings are interpreted into appropriate translation units and relations. After synthesis TL texts are derived from the TL structures or representations produced by the conversion process stage. The operation of the EBMT translation design system is founded on finding or extracting examples of target language sentences that are analogous to input source language sentences. The extraction of appropriate translated sentences is preceded by an analysis stage for the decomposition of input sentences into appropriate fragments. The processes of analysis (decomposition) and synthesis (recombination) are designed, respectively, to prepare input text for matching against database and to produce output text. In HYBRID MT, rules could be used, such as in RBMT (Rule-Based Machine Translation) when an example of the source language to be translated into the target language happens not to be found in the machine database. According to this way of functioning, HYBRID MT has been seen as a better way of translation EBMT combine with RBMT. The matching examples approach for translation can work more successfully on more different types of languages.

A hybrid-based can be applied on some other languages with minor modifications. The database has been designed to be flexible where most of the rules are defined in tables, basically, the lexicon, grammar, morphology, derivation, irregular and translation-example tables, while the Parser will be used as a part of the morphological analysis phase to give the part of speech for the source language words.(V.Ambati and Rohini 2007) proposed a system for English-Indian machine translation system with high quality, to handle the Indian language problems .

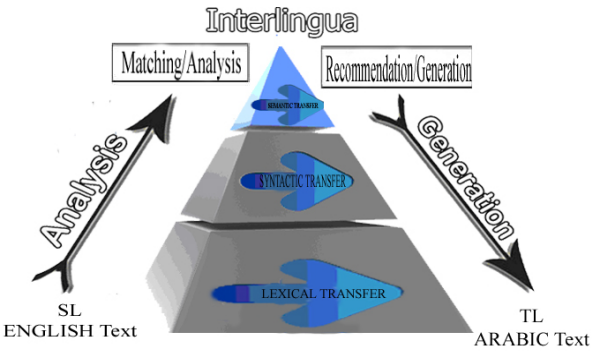

Fig 1: Hybrid based MT

4. Rules

The Rules are included within "English Intelligent-Rules" & "Arabic Intelligent -
Rules" Steps and they are divided as follows:

Grammar Rules
English-Arabic Rules
Linguistic Rules
Translation Rules

Rules are what add to automated translation the overall meaning of the text and thus our automated translator wouldn't perform inaccurate word by word translation. We believe that knowing, processing and adding thousands of rules specific to the English and Arabic Languages gives the automated translator a higher probability of providing an accurate translation. In what
follows are some of the rules that we started generating through analysis of many texts.
In many cases we prefer to put general rules that we implement by coding, which would increase speed by reducing search time, decrease database words and provide same translation results.

1)

a) Example: He is kind.
(GOOGLE) انه لطيف [*enaho latif*]
(SYSTRAN) هو لطيف [*hoa latif*]
The correct translation: هو لطيف [*hoa latif*]

b) Example: He speaks French.
(GOOGLE) انه يتحدث الفرنسية [*enaho etahadath alfaransiah*]
(SYSTRAN) هو يتحدث اللغة الفرنسية [*hoa etahadath allogah alfaransiah*]
The correct translation: يتحدث الفرنسية [*etahadath alfaransiah*]

Rule example one:

a) Translate "he" into Arabic when it is followed by an auxiliary

b) Omit the translation of "he" when it is not followed by an auxiliary.

2)

a) Example: The book is the best friend.
(GOOGLE) هذا الكتاب هو افضل صديق
[*hatha alketab hoa afdel sadeq*]
(SYSTRAN) ال كتاب ال صديق جيد
[*al – ketab al- sadeq jaeead*]
The correct translation: ان الكتاب افضل صديق
[*ena alketab afdal sadeq*]
Book كتاب [*ketab*]
Friend صديق [*sadeq*]
Best الافضل ا و افضل[*alafdal*] or [*afdal*]

b) Example: This car is the best
(GOOGLE) هذه السياره افضل [*hathe alsearh afdel*]
(SYSTRAN) هذا سياره ال جيد [*hada searah al jead*]
The correct translation: هذه السيارة الافضل
[*hathe alseara alafdel*]

Rule example two:

=> (best is translated as الافضل when it happens to be at the end of a sentence
Like example b.

3)

Example: I heard the two boys speak in a low voice.
(GOOGLE) سمعت
اثنين من الفتيان في الكلام بصوت منخفض[*samato ethnen mn alftian fe alklam bsot mokafed*]
(SYSTRAN) انا سمعت ال اثنان فتى تكلمت في صوه منخفض
[ana *samato al- ethnen fta tklamt fe sot mokafed*]

The correct translation: سمعت الولدين يتحدين بصوت منخفض
او سمعت الولدان يتحدثان بصوت منخفض
[*samato alwldan ethdthan bsot mokafed*]

hear يسمع او يصغي [esma`a]
heard استمعت او سمعت [sama`t]
two اثنان [ethnan]
boy ولد غلام صبي [walad]
boys الاولاد ولدين ولدان [alawlad]
speak يتحدث يتكلم [etklm]
He speaks يتحدث يتكلم [etklam]
She speaks تتحدث تتكلم [ttklam]

Rule example three:

=> (for "she" in the present, we take the same verb conjugated for "he" but instead of ي at the beginning we put ت)

(in the past for "he" we take all the verb as in present except the first letter)

He spoke تكلم

=> (in the past tense for "she" we take the verb in the past for "he" and we add at the end of the verb ت)

She spoke تكلمت

=> (in the past tense, for the pronoun "they" (plural masculine), we take the past of "he" and we add _M at the end of the verb. In the feminine plural case the past tense is again formed by taking the past of "he" but 2 should be added at the end of the verb).

they spoke (masculine) تكلموا
they spoke (feminine) تكلمن

4)

Example: The teacher praised both students who answered correctly.
(GOOGLE) واشاد المعلم كلا من الطلاب الذين اجابوا بشكل صحيح

(SYSTRAN) مدح المعلم كلا طالب الذي اجاب بشكل صحيح

The correct translation: مدح المعلم كلا الطلاب الذين اجابو بشكل صحيح

teacher المعلم المدرس
Praise اطراء او ثناء

Rule example four:
=> (Above, is the translation of "praise" as a noun, and it is as such when the
word happens to come after a preposition for example such as: a praise. When
this word follows a noun then it should be analyzed as a verb and it is translated
in the form shown below:)

praise (to praise) مدح
both كلا كلتا معا

=> ("both" is translated as معا when it happens to refer to the subject.
For Example: They ate both. They went both.
"Both" here refers to they. Usually also when "both" ends the sentence.)

(GOOGLE) اكلت كل منهما
(SYSTRAN) هم اكلوا كلا

student الطالب
who من الذي التي
answer يجيب جواب
correct صحيح
Correctly بشكل صحيح

=> (correctly is an adverb. Adverbs are recognized by "tly" at the end. This is
translated in Arabic by adding بشكلor ب to the word in Arabic.

Example:
Quick سريع
Quickly او بشكل سريع بسرعه

Example: They ate both. They went both.
"Both" here refers to they. Usually also when "both" ends the sentence.)

(GOOGLE) اكلت كل منهما
(SYSTRAN) هم اكلوا كلا

5. Handling the Agreement and Word Reordering Problems in MT

4.1. Agreement and Reordering with Arabic System

In this section we will explore different areas that are expected to cause agreement and reordering problems during translation from English into Arabic. The test example will be put to the Arabic MT system.
Example:
Can I book tickets for two good boys tomorrow?
(GOOGLE) ويمكن حجز التذاكرالاول للصبيان جيدة غدا؟
(SYSTRAN) يستطيع انا كتاب تذكرة ل اثنان فتى جيد غدا ?

4.2 Proposed Solution with Hybrid MT

Let us investigate the translation with Arabic MT system and see how it can handle the agreement and word-ordering, using hybrid – based MT following Methods steps:

STEP 1: Input the source text in English language
Can I book tickets for two good boys tomorrow?"

STEP 2: Pass the source text to the OAK Parser and get the output as (tagged POS)
Can/MD, I/PRP, book/VB, tickets/ NNS, for/IN, two/CD, good/JJ boys/NNS tomorrow/NN

STEP 3: From the output in 2, construct the English pattern in the format of the grammar table. : 1/MD, 2/PRP, 3/VB, 4/NN, 5/NN, 6/CD, 7/JJ, 8/NNS, 9/NN

STEP 4: Check the procedure according to which EBMT is based is the following:
4.1 The alignment of texts. "Can I book tickets for two good boys tomorrow?"
4.2The matching of input sentences against phrase (examples) from stored database. Can I book table, Can I book room
4.3The selection and extraction of equivalent target language or translated phrases. Can I book

4.4The adaptation and combination of translated phrases an acceptable output sentences. "Can I book tickets for two good boys tomorrow?"
4.5 When an example of the source language to be translated into the target language happens not to be found in the machine database go to step5.
"tickets for two good boys tomorrow?"
STEP 5: Retrieve the record of this pattern from the grammar table in order to know the subject, verb, object, agreement requirements, and the equivalent pattern in Arabic language.
VB, VBP, VBZ, VBD, and VBN ➔ VBX

NN, NNS, NNP, NNPS, NNX, and NNXS ➔ NNX

POS for is, are, and am ➔ AUX

POS for was and were ➔ AUXD

STEP 6: From the lexicon get the features and Arabic meaning for all words of the sentence.
STEP 7: Check for irregular word(s)
STEP 8: Apply the agreement rules for verbs and their subjects.
STEP 9: Apply the agreement rules for adjectives and the entities that they describe.
STEP 10: Apply modification rules on the object words.
STEP 11: Construct the Arabic text using the pattern exists in the grammar table.

(1/MD, 2/PRP, 3/VB) ➔ 1
(4/NN) ➔2
(5/NN, 6/CD, 8/NNS) ➔3
(7/JJ) ➔ 4
(9/NN) ➔ 5

STEP 12: Repeat steps 1 to 11on the next sentence.

هل يمكنني حجز تذكرتين لولدين جيدين غدا

6. Conclusion

Many shortcomings in the output of MT have been shown in this paper, due to either faulty analysis of the source language text or faulty generation of the target language text. Enhancement to the output can be done only by formalizing our linguistic knowledge and enriching the computer with adequate rules to deal with the linguistic phenomenon. Fully automated, high quality machine translation (FAHQMT) has not yet been achieved. Yet there is a lot that we can do to improve the quality of MT output and increase its usefulness.

In this paper we have presented the necessity to handle both the agreement and the words reordering in the machine translation from English to Arabic. We proposed a hybrid-based approach to solve those problems; the paper has dealt with two features that greatly affect the output of MT, that are agreement and ordering problem which comes from the fact that different languages have different text orientation where some of them are left-to-right and others are right-to-left. The order of the words in the sentence is also different from one language to another.

References

[1] **Michelle Wendy Tan** ,2008.` acooperating hybrid mt enviroment using RULE-BASED and EXAMPLE-BASED paradigrams, manila ,philipines..

[2] **Attia, M**. 2002. 'Implications of the Agreement Features in Machine Translation'. AL-AZHAR UNIVERSITY

[3] **mohammd, and Sembok**, T. 2007a. 'TOWARD FULLY AUTOMATED ARABIC MACHINE TRANSLATION SYSTEM', IJCSNS International Journal of Computer Science and Network Security, 7 (5): 1-10.

[4] **Franck, J. Lassi, G Frauenfelder, U. & Rizzi, L**. 2006. 'Agreement and movement: A syntactic analysis of attraction'. Cognition, (101): 173-216.

[5] **Hutchins, W. and Somers. L.** 1992. 'An Introduction to Machine Translation'. London: Academic Press. Love P.E.D and Irani Z. 2003. 'A project management quality cost information system for the construction industry'. Information and Management, 40(7): 649-661.

[6] **Mohammad, M.** 1990. 'The problem of subject-verb agreement in Arabic: Towards a solution', Amsterdam, Benjamins, Publishing Company: 95-125.

[7] **mohammd, and Sembok**, T. 2007b. 'HANDLING AGREEMENT IN MACHINE TRANSLATION FROM ENGLISH TO ARABIC'. The 1st International Conference on Digital Communications and Computer Applications (DCCA2007). JUST: 385 – 379.

[8] **Trujillo, A**. 1999. 'Translation Engines Techniques for Machine Translation', Springer – Verlag Berlin Heidelberg, New Work.

[9] **Satoshi, S**. 2008, 'The manual of Apple Pie Parser v7.0' Computer science department, New York university.

[10] **H.Yamout, K.Kanso**. 2006, 'online English-Arabic Translater Electrical & Computer Engineering department, AUB university.