A Rule-based English to Arabic Machine Translation Approach

Ahmad Farhat and Ahmad Al-Taani
Department of Computer Science, Yarmouk University, Irbid, Jordan

Abstract: *In this study, we propose Rule-based English to Arabic Machine translation system for translating simple English declarative sentences into well-structured Arabic sentences. The proposed system translates sentences containing gerunds, infinitives, prepositions, direct and indirect objects. The system is implemented using bilingual dictionary designed in the SQL server. A major goal of this system is to be used as a stand-alone tool and can be integrated with general (English-Arabic) machine translation systems. The proposed system is evaluated using 70 various simple English declarative sentences written by English Language experts. Experimental results showed the effectiveness of the* proposed *MT system in translating English simple declaratives sentences into Arabic. Results are compared with two well-known commercial systems; Google Translate and Systran Systems. The proposed system reached an accuracy of 85.71% while Google got 31.42% and Systran got 20% on the same test sample.*

Keywords: *Machine translation, Rule-based approach, Bilingual dictionary, Natural Language Processing.*

1. Introduction

Since the middle of last century, and particularly in the last ten years, there has been a spurt of research growth in Machine Translation (MT). Various MT systems have been developed in Europe, the USA and in the Far East, but these systems principally involve European languages [5]. Comparatively little work has been done on MT systems involving Arabic as either the source or target language. Also the incorporation of Arabic into MT systems is clearly of importance, not only from economic and trade considerations, but also for social and cultural reasons.

Arabic Natural Language Processing has been the focus of research for a long time in order to obtain an automated understanding of the Arabic language [3]. It is a highly inflectional language with a rich morphology and relatively free word order, and two types of sentences: nominal and verbal [10]. English is a universal language that is widely used in the media, commerce, science, technology, and education. Modern English content (e.g. literature and web content) is larger than the amount of Arabic content available. Consequently, English-to-Arabic MT is particularly important and the systems are mainly based on the transfer classification.

The related work showed that English-Arabic MT approaches were concentrated on Rule-based and Corpus-based approaches. It also showed a small amount of work done on the Arabic language as a target language.

Rule-based MT (RBMT) has several advantages over the corpus-based approaches, which is one of the most widely explored areas in MT [13]. These include:

1. RBMT systems tend to produce better translations from a syntactic point of view [9].
2. RBMT systems deal with long distance dependencies, agreement and constituent reordering in a more principled way, since they perform the analysis, transfer and generation steps based on morphologic knowledge [9].
3. RBMT systems are a less-resourced approach compared with the corpus-based approaches which need very large corpora [11].
4. RBMT systems are extensible and Maintainable [7].

On the other hand, RBMT have problems with lexical selection due to a poor modelling of word level translation preferences [4]. Furthermore, if the input sentence cannot be parsed due to the limitations of the parser or because the sentence is ungrammatical, the translation may fail and produce very low quality results [6]. The literature also showed that the Example-based MT which is one of the corpus-based approaches can cover this loophole.

According to Groves et al. [4] the Example-based machine translation has a main advantage over rule-based approaches; it is usually better at lexical selection and fluency, since it models lexical choice with distributional principles and explicit probabilistic

language models trained on very large corpora. Furthermore, it can be implemented in systems which are not EBMT systems themselves.

Researchers stated that agreement and word ordering are the main problems in MT and play a big role with the quality of translated sentences from English to Arabic. Agreement is main property of language, it occurs when two words in the appropriate pattern exhibit morphology consistent with their co-occurrence. In the English language, the main case of this linguistic mechanism is number agreement between a subject and a verb [2], and there are several agreements we attempt to solve in this study such as: Adjective-noun agreement, Verbs-subject agreement, and Pronouns agreement.

More than two-thirds of the linguistic efforts in analyzing English are spent on the morphology [10]. In most existing systems the incorrect translation occurs because agreement and word reordering problems still exist. In this research, we propose a MT system to deal with agreement and reordering problems. The proposed approach can be extended to include other types of English sentences.

2. Methodology

The proposed methodology is flexible and scalable and the main advantages are: first it depends on the morphological issues which are mainly based on translation rules of English-Arabic languages. Secondly, it can be applied on several different languages.

The proposed machine translation system use the transfer based method. This method attracts me in contrast to other methods, because with the direct method the translation is based on dictionaries and word-by-word translation with the same grammatical adjustment. There is no parsing here, so it is not enough to develop the desired machine translation. Regards to the Interlingua method, it is beyond the need for the desired machine translation because this method has much relevance in multilingual machine translation and this emphasizes a single representation for different languages.

In general, the flow of the transfer-based approach is as follows; it begins with the analyzer which takes the English sentence (source text) that is to be translated and produces a POS tagging for every word in it. Next it transfers this POS tagging to an English sentence pattern to obtain the equivalent Arabic sentence pattern by using reordering rules. Finally, get the meaning of words by using bilingual dictionaries, and from the Arabic sentence pattern that has been generated, and depending on agreement and synthesis rules, it generated the target text. The proposed methodology design is shown in Figure1.

2.1 The Analysis Phase

A sentence is a group of words which starts with capital letter and ends with a dot. A sentence contains or implies a predicate and a subject.
Sentences contain clauses, simple sentences have one clause and sentences can contain subjects and objects. The subject in a sentence is generally the person or thing carrying out an action comes before the verb. The object in a sentence is involved in an action but does not carry out that action. The object comes after the verb. For example: The boy climbed a tree.
If you want to say more about the subject (the boy) or the object (the tree), you can add an adjective, the adjective comes before the noun (whether subject or object). For example: The boy climbed the tall tree.

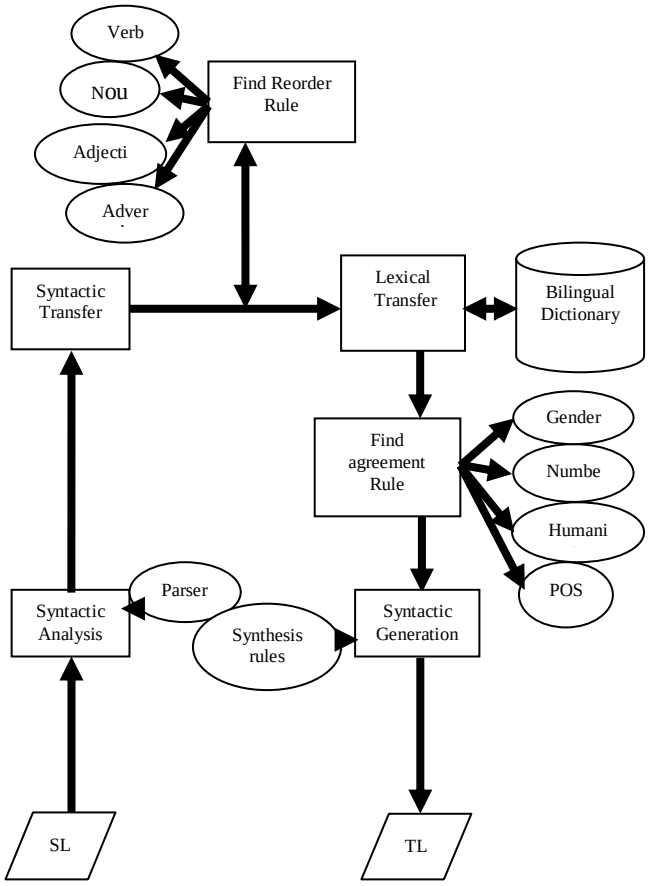

Figure 1: The overall methodology design

In the English language there are many patterns for sentences and according to [8] the simple declarative sentences (SDS) have some pattern as follows:

- Subject - Verb - Object Pattern:
  For example: He likes coffee.

- Subject - Verb – Indirect Object – Direct Object Pattern
  For example: The teacher gave the student a book.
- Subject - Verb - Adverb Patten:
  For example: The boy came quickly
- Adjective - Subject Pattern:
  For example: The small house.
- Subject - Verb - Adjective Patten:
  For example: He is kind.
- Subject - Verb – Adverb - Adjective Patten:
  For example: The girl is very smart

In general, the English language contains eight parts of speech (also called lexical categories). These are the following: (Verbs, Nouns, Pronouns, Adjectives, Adverbs, Preposition, Conjunctions and Interjections) [8]. Computers need many POS to distinguish between words, to deal with the grammatical structure of a given sentence and to help resolve some of the morphological ambiguities of words. There is an urgent need for a tool to handle these POS which is known as a part-of-speech (POS) tagger. Part-of-speech tagging is the process of marking sentence words with their part-of-speech. The tags are taken from a tag set, which is a predefined tag list. Table 1 shows the well-known Penn TreeBank tags [12].

Table 1|: The Penn TreeBank project tag set

Tag	Role	Tag	Role	Tag	Role
CC	Coordinating conjunction	NNS	Noun plural	TO	To
CD	Cardinal number	NNP	Proper Noun singular	UH	Interjection
DT	Determiner	NNPS	Proper Noun plural	VB	Verb, base form
IN	Preposition	PRP	Personal pronoun	VBN	Verb, past participle
JJ	Adjective	PP$	Possessive pronoun	VBP	Verb, non-3s, present
JJR	Comparative adjective	RB	Adverb	VBZ	Verb, 3s, present
JJS	Superlative adjective	RBR	Comparative adverb	WDT	Wh-determiner
MD	Modal	RP	Particle	WPZ	Possessive Wh-pronoun
NN	Noun singular	SYM	Symbol	WRB	Wh-adverb

In the proposed MT system I used the OpenNLP POS tagger which depends on the Penn TreeBank tag set. The OpenNLP (POS) tagger like other natural-language tools was developed based on a rule-based paradigm or a corpus-based one. Rule-based taggers use a set of rules to compute the tags of a new given sentence, while corpus-based taggers learn how to tag new inputs from a large tagged corpus. Hybrid taggers also exist. The OpenNLP (POS) tagger used huge corpus files to distinguish the parts of speech of words. The following are some of them:
(gen.nbin,location.nbin,num.nbin,money.nbin,organization.nbin, person.nbin, time.nbin…etc)
For example, given the sentence, "They are two good boys", the following are the tags of its words, using the OpenNLP (POS) which is based on the Penn TreeBank tag set:
They/PRP are/VBP two/CD good/JJ boys/NNS

2.2 The Transfer Phase

The second phase is the transfer phase, in which a transformation is applied to the English sentence pattern to construct the equivalent in Arabic. Once the POS tagging process is complete, I store the POS for all the words of the given sentence into an array to simplify the handling of each word by its index and through that I can get the English pattern for each sentence depending on the English grammar as mentioned earlier, such as: the subject coming before the verb, the object coming after the verb, the adjective coming before the noun and the adverb coming after the verb [2].

The second step is transferring the English sentence pattern obtained from the first step to its equivalent Arabic sentence pattern depending on the English-Arabic comparison pattern table [2]. This step was done by swapping indexes in the array of POS and the array of words.

Table 2: English-Arabic comparison pattern table

English Sentence Pattern	Arabic Sentence Pattern
S V E.g. The boys ran	**V S** ركض الأولاد
S V O E.g. The child drank the milk	**V S O** شرب الطفل الحليب
S V Oi Od E.g. The teacher gave the student a book	**V S Oi Od** أعطى المعلم الطالب كتاب
S V Cs E.g. Ali is kind E.g. Ali was sick E.g. Ali came quickly E.g. Ali is very smart	**S Cs Or S V Cs *Or* V S Cs** علي لطيف كان علي مريض جاء علي بسرعة علي ذكي جدا

Cs O E.g. The small house	**O Cs** البيت الصغير
S V O Co E.g. They elected him president	**S V O Co** (هم) انتخبوه رئيس
S V Oc E.g. Ali lives a good life	**V S Oc** يعيش علي حياة جيدة

Where S:Subject, V:Verb, O:Object, Od: Direct Object, Oi: Indirect Object, Cs: Subject complement which may be an Adjective, an Adverb or both , Co: Object complement which may be a Noun or an Adjective, Oc: Cognate object which is Adjective followed by noun.
According to this table I wrote the reordering rules to be used in my MT system to got a correct Arabic sentence pattern from the English one. These cases constitute a set of reorder rules as follows:

2.2.1 Rule1

When the English sentence contains a (Noun) as Subject followed by a Verb, in the corresponding Arabic sentence the Verb must precede the Subject [1].
For example: "The boys ran" must translated to Arabic sentence as: "ركض الأولاد"

2.2.2 Rule2

When the English sentence contains a (Pronoun) as Subject followed by a Verb, in the corresponding Arabic sentence the order stay as it is.
For example: "He runs" must translated to Arabic sentence as: " هو يركض "

And there are also 14 other rules.

2.3 The generation phase

The last phase is the generation phase, which is a combination of extracting the Arabic meaning and other features for each word from the English-Arabic bilingual dictionary, then applying syntheses and agreement rules on the sentence to produce the Arabic sentence as a result of the translation process.

In the Arabic language, the verb and adjective invariably change whenever the subject changes in gender and number. The gender in Arabic is basically masculine or feminine, and the number in Arabic is singular, dual, or plural [2]. I added a third feature which is humanity to get more accurate translation; the humanity is true or false. So I designed my own English-Arabic bilingual dictionary which included these fields: English words, Arabic words, POS tags, number, gender and humanity. The English word and the POS tag fields will be filled automatically from the first phase by the POS tagger and the other fields will be filled manually by machine learning form.

There are a lot of cases that arise during the generation process that must be taken into account and fixed before generating the resulting Arabic sentence. These cases constitute a set of grammar rules as follows:

2.3.1 Rule1: Adjective-Noun definiteness Agreement
A sentence containing the article "the" as (DT) followed by a noun (NNX)., in Arabic language there is no separate equivalent word to the article "the", so instead a separate word prefix will be added"ال"to the next (NNX) that follows in Arabic.

For example: The door → البيت

A sentence containing the article "the" as (DT), followed by an adjective (JJ), followed by a noun (NNX). After applying a suitable reorder rule, instead of a separate word being added in Arabic, a prefix will be added "ال"to the next (NNX) then adding "ال" to the next (JJ).
For example: The small house →البيت الصغير

A sentence containing the article "a" or "an" as (DT), followed by an adjective (JJ), followed by noun (NNX), then in Arabic language these articles must not translated.
For example: A small house →بيت صغير
There are also many rules I processed that covering subject - verb Agreement, adjective – noun agreement (for number and gender), cardinal number – noun agreement, cardinal number – Noun and Adjective agreement, cardinal number – pronoun and noun agreement, and personal possessive pronouns - noun agreement.

3. Experiments and Evaluation

I drew a sample consisting of 70 various simple English declarative sentences selected from human experts in the English Language.

3.1 Evaluation Method and Results

In order to evaluate the correctness of the proposed MT system, we developed suitable evaluation methodology. The following steps describe the evaluation methodology:

1. Run the system on the data set.

2. Compare the output translation between the proposed MT system, Google MT and Systran MT by human Expert.
3. Classify the problems that arise from the mismatches between the proposed MT system and other MT systems.
4. Determine the percentage accuracy of the data set for each MT system, by computing the number of correctness test cases over total number of test cases multiplied by 100%.
5. Suggest possible solutions for the identified problems and apply the necessary improvements to the MT system.

3.2 Analysis of Results

The result shows that 60 sentences have been translated correctly using the proposed MT system and 10 have been translated incorrectly, so it needed some improvements, on the other hand 22 sentences have been translated correctly using the Google translator and 14 sentences have been translated correctly using the Systran translator. The proposed MT system has the highest accuracy of 85.71% after that the Google translator with 31.42% accuracy and at lastly the Systran translator with 20% accuracy. Table 3 shows a sample of tested sentences compared with Google translate and Systran system.

4. Conclusions

Enhancement of the outputs of the proposed MT system can be done only by formalizing our linguistic knowledge and enriching the system with adequate rules to deal with the linguistic issues. Fully automated high quality machine translation (FAHQMT) has not been achieved yet. There is a lot of work that we can do to improve the quality of MT outputs and increase its usefulness. In this project I have presented the necessity to handle both the agreement and the word reordering problems in the machine translation from English to Arabic. I proposed a system which uses the advantages of the Rule-based machine translation (RBMT) approach to solve those problems. The project has dealt with the two features that greatly affect the outputs of MTs, which come from the fact that different languages have different text orientations where some of them are left-to-right and others are right-to-left. The orders of the words in the sentence are also different from one language to another.

The proposed MT system is restricted only to simple English declarative sentences and no other sentence type, so extending the current MT system to cover not only simple declarative sentences, but also the compound ones and possibly other types of sentences will be the next step in future works. That depends on more analysis and demands more grammars rules, so the complexity of the translation process will increase.

Finally, Some sub patterns from the main patterns of simple declarative sentence are not yet included in the proposed MT system, not because hard to do it, but since it demands more time to cover, while there is a time constraint to complete the project and get sensible results. For example: the sentence: "The girls will eat the food", it's a sub pattern from (subject-Verb-Object), the verb here is in the simple future tense, the proposed MT system covered the verb in present, past tense and gerund by using equivalent reorder and agreement rules. But in the case of the verb being in the simple future tense, this sub pattern was covered when the subject was singular but not plural, for example: the sentence: "The girl will eat the food".

References

[1] Abdo, Dawod., 'Deep Structure of the Sentence in Arabic: Did Verb Subject Object or Subject Verb Object' . By Dar Al-Karmel. Amman, Jordan, Pages 103-105, 2008.

[2] Alkhuli, Muhammad Ali.,'Comparative Linguistics: English and Arabic'. By the National Library, (ISBN): 9957-401-05-9. Amman, Jordan, 1999.

[3] Al-Sughaiyer, Imad and Al-Kharashi Ibrahim., Arabic Morphological Analysis Techniques: A Comprehensive Survey. Journal Of The American Society For Information Science And Technology, 2004.

[4] Groves, Declan and Way., Hybrid data-driven models of machine translation. Volume 19, Issue 3-4, pp 301-323, 2005.

[5] Hutchins, W.John., Machine Translation: A brief History. Concise history of the language sciences: from the Sumerians to the cognitivists. Edited by E.F.K.Koerner and R.E.Asher. Oxford: Pergamon Press. Pages 431-445, 1995.

[6] Hutchins, W.John., Example-based machine translation: a review and commentary. Published online: © Springer Science and Business Media. Pages 6,7, 2006.

[7] Kaji Hiroyuki., An Efficient Execution Method for Rule-Based Machine Translation. Systems Development Laboratory~ Hitachi Ltdo1099 Ohzenji, Asao, Kawasaki, 215~ Japan, 1988.

[8] Khalil, Aziz M., A Contrastive Grammar of English and Arabic. By Jordan Book Center, (ISBN): 978-9957-604-13-4. Amman, Jordan, 2010.

[9] Labaka, Gorka and Stroppa, Nicolas and Way, Andy and Sarasola, Kepa., Comparing rule-

based and data-driven approaches to Spanish-to-Basque machine translation. Copenhagen, Denmark, 2007.

[10] Ryding Karin., A Reference Grammar of Modern Standard Arabic. Cambridge University Press The Edinburgh Building, Cambridge, CB2 2RU, UK, 2005.

[11] Shaalan Khaled., Rule-based Approach in Arabic Natural Language Processing. International Journal on Information and Communication Technologies, Vol. 3, No. 3, 2010.

[12] The Penn Treebank Project., Computer and Information Science, Penn University. URL http://www.cis.upenn.edu/~treebank/ (viewed on 27/11/06), 2006.

[13] Tripathi Sneha and Sarkhel Krishna., Approaches to machine translation. Annals of Library and Information Studies. Vol.57, pp. 388-393, 2010.

Table 3: Evaluation

Sentence	My MT system results	Google MT results	Systran MT results	Human judgment
Sarah writes a letter	تكتب سارة رسالة	سارة يكتب بريد إلكتروني	سارة يكتب حرف	ترجمة الباحث أصوب
The boys write a book	يكتب الأولاد كتاب	الأولاد إرسال كتاب	يكتب الفتى كتاب	ترجمة الباحث أصوب ينقصها تنوين المفعول به
They ate the meat	هم أكلوا اللحم	أكلوا لحوم	اللحم هم أكلوا	ترجمة الباحث و ترجمة سيستران أصوب
The girls were good	كانت البنات جيدات	وكانت الفتيات جيدة	جيّد البنت كان	ترجمة الباحث أصوب
The lions eat the meat	تأكل الأسود اللحم	الأسود تأكل اللحم	اللحم الأسد يأكل	ترجمة الباحث و ترجمة جوجل أصوب
She needs help	هي تحتاج مساعدة	وهي في حاجة إلى مساعدة	مساعدة يحتاج هو	ترجمة الباحث أصوب
It eats the meat	إنها تأكل اللحم	وهو يأكل اللحوم	اللحم يأكل هو	ترجمة الباحث أصوب
They need help	هم يحتاجون مساعدة	إنهم بحاجة إلى مساعدة	مساعدة يحتاجون هم	الترجمات الثلاث صائبة
Sarah ate the apple	أكلت سارة التفاحة	سارة أكل التفاح	سارة أكل التفاح	ترجمة الباحث أصوب
They elected him president	هم انتخبوه رئيس	إنهم انتخبوه رئيسا	هم انتخبواه رئيس	ترجمة ترجمة جوجل و ترجمة الباحث ينقصها تنوين المفعول به
Sarah lives a good life	تعيش سارة حياة جيدة	سارة تعيش حياة جيدة	سارة يعيش حياة جيد	ترجمة الباحث و ترجمة جوجل أصوب
Ahmad lives a good life	يعيش احمد حياة جيدة	أحمد يعيش حياة جيدة	أحمد يعيش حياة جيد	ترجمة الباحث و ترجمة جوجل أصوب
They elected him president	هم انتخبوه رئيس	إنهم انتخبوه رئيسا	هم انتخبواه رئيس	أصوب وترجمة ترجمة جوجل الباحث ينقصها تنوين المفعول به