International Journal of Current Engineering and Technology
ISSN 2277 - 4106

Available at http://inpressco.com/category/ijcet

Research Article

A Technique to Enhance the Process of Handling Gender Agreement Requirement in English-Arabic Machine Translation

Omer M. A. Abu Shqeer[a*] and Mohammed M. S. Abu Shquier[b]

[a]College of Computer Science and Engineering, Taibah University, Medinah Almonawarah, KSA
[b]Faculty of Computers and Information Technology, University of Tabuk, Tabuk, KSA

Accepted 22 Jan. 2013, Available online 1March 2013, **Vol.3, No.1 (March 2013)**

Abstract

In this paper we highlight some difficulties in machine translation especially from English to Arabic; and focus on the agreement requirements between subject and verb, as well as between noun and adjective. We proposed and designed an approach to enhance the translation quality in English-Arabic translators; the approach based on the identification of the exact gender of the nouns before the translation process; the target nouns are those whom used for both masculine and feminine in the source language such as driver, programmer, teacher, and so on. We applied and tested the proposed approach on two translators, and the results showed a significant enhancement of the translation output. The proposal is flexible and scalable, it is a rule-based approach, and can be applied on some other features and/or languages.

Keywords: *machine, translation, agreement, gender, rule-based, lexicon, syntactic, Arabic*

1. Introduction

Natural Language Processing (NLP), including Machine translation (MT) became one of the most concerning topics for the researchers worldwide. Nowadays, MT receives a significant concern from the academic as well as industry researchers. The success in MTs varies from language to another since languages differ from each other in grammar and morphology. Up to date there is no general purpose MT system satisfies all the wishes of the customers, while the specific purpose MT systems satisfy a high level of accuracy, but these systems are restricted in their domains. MT is normally taken in its restricted and precise meaning of fully automated translation; but we can define MT to include any computer-based process that transforms (or helps a user to transform) written text from one human language into another. MT can be fully automated where the machine should perform the translation without the intervention of a human being during the process, human-assisted MT in which a computer system does most of the translation, appealing in case of difficulty to a human for help, and machine-aided translation in which a human does most of the work but uses one or more of the computer systems resources for assistant such as dictionaries and spelling checkers. (Abu Shqeer Omer *et al*, 2002) (El Kholy Ahmed, 2012)

Even though MT is the oldest application of NLP, and several approaches including rule-based, example-based, hybrid, and statistical have been adopted in developing MT applications, we can say that the main problem of getting high quality output of translation remains unsolved (Elming Jakob *et al*, 2009). Such difficulties and problems in MT – but not all - that are related to this research are:

a) Word order: Order of the words in the statement/sentence varies from one language to another and sometimes in the same language depending on the text context. Therefore, a lot of analysis and processing are needed to handle this issue.
b) Agreement: Requirements of Subject-verb agreements as well as noun-adjective agreements in terms of number, gender, person and case. Some languages require all types of agreements, while some others require part of these agreements. Arabic language – the target of this research - requires all types of agreements. A huge amount of rules can help in handling those agreement requirements.
c) Ambiguity: Some words can be classified to more than one category type of speech, e.g. the word play can be marked as either a verb or a noun. Most of the words in the source language have several meanings in the target language. Therefore, the understanding of the text subject helps in selecting the proper meaning among the possible alternatives.
d) The new words/phrases in the language, the abbreviations, and the names of persons, places and things, etc. are some other difficulties. With the evolution of sciences, especially communication and information technology, every day new terms arise; in many cases there is no direct equivalent term in

*Corresponding author: Omer M. A. Abu Shqeer

Arabic for an arisen term; by the absence of centralized organization for arabization, different people may translate the same term to different meanings.

e) The correct meaning remains the most difficult problem in MT where the context affects the translation; this problem needs very complex lexicons, rules, analysis, and synthesizers to produce an acceptable translation quality.
f) Noun features: Gender of a noun may differ from one language to another, e.g. there are a lot of English words that are used for both masculine and feminine, while they have different equivalents in Arabic; one for masculine and another for feminine. Therefore, morphological analysis has to be taken into consideration in the translation into Arabic.
g) Numbers: Arabic Cardinal numbers (from 3 to 10) require anti-agreement; this means that the target takes the opposite features of the controller. "Masculine-gender numbers are used with nouns whose singular is feminine, and feminine-gender numbers are used with nouns whose singular is masculine" (Nasr T. Raja, *1967)*. In order to get the correct translation, the countable gender should be identified.

This research concerns with the agreement problem in fully automated English-Arabic MT. It proposes a technique to enhance the quality of these translators. It focuses on the enhancement of satisfying the gender agreement using semantic analysis. Nasr stated that gender in Arabic is grammatical, not natural (Nasr T. Raja, *1967)*. Hence, "Every noun in Arabic is either masculine or feminine" (Al-Jarf Reim, 2000) . Typical Gender agreement contrasts are: masculine, feminine, and neuter. In Arabic there are further contrasts between animate/inanimate and human/non-human objects; and each is given grammatical properties accordingly. (Attia Mohammad, 2002)

The next section briefly presents the agreement problem for number, gender, person and case respectively in MT from English to Arabic; in the following section we discussed how different MT approaches handle these agreements; after that we present a proposal to enhance the process of handling the gender agreement in English-Arabic MT, followed by experiments and results discussion; and finally the last section presents the conclusions of this work.

2. The Agreement Problem in MT from English to Arabic

Any natural language has its own grammar rules. For example the rule "S: NP VP" means that a sentence in English can be a noun phrase followed by a verb phrase; and hence the two sentences: "the student reads a book" and "the students reads a book" are both valid sentences according to the given rule; but the former is valid while the latter is not because the verb (reads) is invalid to be used with the plural noun (students). To a void accepting the second sentence grammatically, the system should handle what is called number agreement between the subject and the verb.

For Arabic verbs, a set of morpho-syntactic features arisen, namely: person (1^{st}, 2^{nd} and 3^{rd}), number (singular, dual and plural), gender (masculine and feminine) and tense (past, present and future) as shown in figure-1 below. Other features are also shown for nouns, adjectives as well as pronouns in the same figure.

In Arabic, variant derivations of the verb are used based on the number, gender, person and other features of the subject. Some other agreements are required between the adjective and the described noun; the adjective derivation in Arabic is also depend on the number, gender and person and other features of the noun as shown in figure-1 above. Furthermore there is an agreement requirement between the numbers and the countable nouns. The following two subsections clarify by examples the subject-verb as well as noun-adjective agreements (Abu Shqeer Omer *et al*, 2002) (Al-Jarf Reim, 2000) (Nasr T. Raja, *1967)*.

2.1Subject-Verb Agreement

In general, there are two different derivations for any English verb in the present tense based on the person feature of the subject. One derivation in the past tense, and one derivation for the future tense regardless of the person feature. In Arabic the story is different; there are many derivations for the same verb in every tense based on the features of the subject; these features are mainly the number, gender, person, alive, and humanity. For example, the present verb *write* in English takes the form *writes* when it is used with he and she, and takes the form *write* when it is used with I, you, we, and they. In the future tense it is *write* with all subjects; and it is *wrote* in the past tense with all subjects. The equivalent Arabic translation for the verb *write* in any tense takes variant derivations upon the subject features. Tables 1, 2, and 3 show the Arabic derivations of the verb *write* in present, past, and future tenses respectively. It is clear from those examples that the subject gender in the addition to the person and number features plays a crucial role in the translation process.

Animated nouns can be further classified into either human or non-human nouns (humanity feature). Human nouns preserve their gender in singular, dual, and plural forms. Non-human nouns maintain their gender in singular and dual forms only, and in the plural form they invariably take the feminine gender. Therefore humanity feature is another issue in deriving the correct verb form in Arabic. The following examples show the effect of the humanity feature on the translations.

1)	The boys are running "yarkodhoon"	الاولاد يركضون
	The dogs are running "tarkodh"	الكلاب تركض
2)	Two boys are running "Yarkodhan"	ولدان يركضان
	Two dogs are running "Yarkodhan"	كلبان يركضان
3)	The boy is running "Yarkodh"	الولد يركض
	The dog is running "Yarkodh"	الكلب يركض

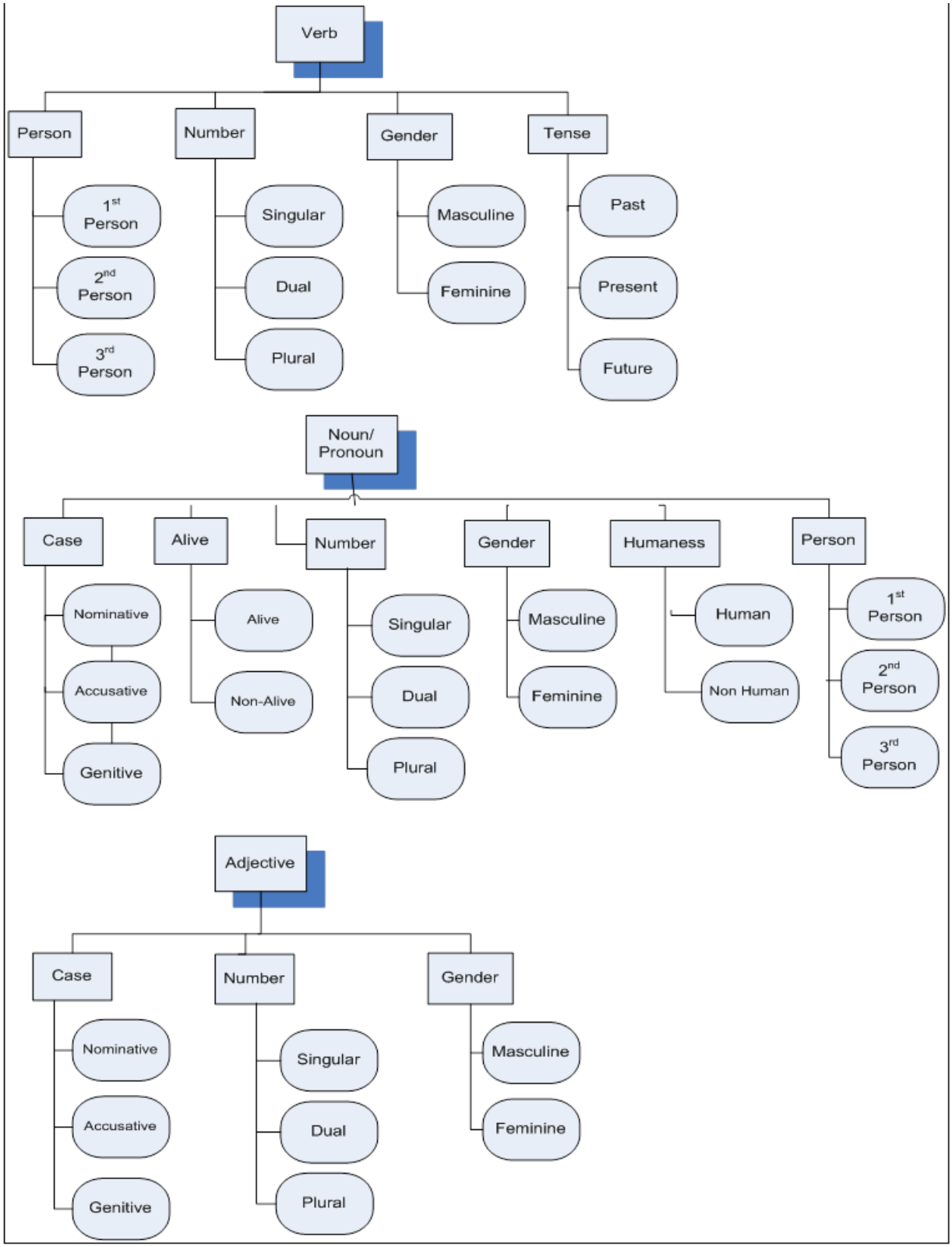

Figure-1: Arabic POS and particular morpho-syntactic features

In the second sentence of example 1, we used the verb "tarkodh تركض" which is singular feminine form with a plural masculine subject "the dogs"; while it is "yarkodhoon يركضون " with "the boys" in the first sentence of the same example. Furthermore, examples 2 and 3 preserve the gender of both nouns boy and dog as it is in the singular and dual forms respectively.

Table 1: Subject agreement markers in Arabic (present verb write)

Person	Gender	Singular	Dual	Plural
First	Mas.	أكتبُ - aktbu	نكتب - naktbu	
	Fem.			
Second	Mas.	تكتبُ - taktbu	تكتبان - taktban	تكتبون - taktbon
	Fem.	تكتبِ - taktbi		تكتبنَّ - taktbna
Third	Mas.	يكتب - yaktbu	يكتبان - yaktban	يكتبون - yaktbon
	Fem.	تكتب - taktbu	تكتبان - taktban	يكتبنَّ - yaktbna

Table2: Subject agreement markers in Arabic (past verb wrote)

Person	Gender	Singular	Dual	Plural
First	Mas.	كتبتُ -katabtu	كتبنا –katabna	
	Fem.			
Second	Mas.	كتبتَ - katabta	كتبتما -katabtuma	كتبتم - katabtum
	Fem.	كتبتِ - katabti		كتبتنّ - katabtunna
Third	Mas.	كتبَ - katabu	كتبا - katabaa	كتبوا - katabuu
	Fem.	كتبت -katabut	كتبتا - katabutaa	كتبنَ - katabna

Table3: Subject agreement markers in Arabic (future verb write)

Person	Gender	Singular	Dual	Plural
First	Mas.	سأكتبُ – sa'ktbu	سنكتب - sanaktbu	
	Fem.			
Second	Mas.	ستكتبُ - sataktbu	ستكتبان - sataktban	ستكتبون - sataktbon
	Fem.	ستكتبِ - sataktbi		ستكتبنَّ - sataktbna
Third	Mas.	سيكتب - sayaktbu	سيكتبان - sayaktban	سيكتبون - sayaktbon
	Fem.	ستكتب - sataktbu	ستكتبان - sataktban	سيكتبنَّ - sayaktbna

2.2Noun-Adjective Agreement

The adjectives in Arabic require number, gender and person agreements between the adjective and the described entity. Let us explain this by the following example: the adjective *big* in English has many equivalent derivations in Arabic based on the number, gender and person of the described entity; these derivations are shown in Table 4

Like verbs, when the adjective is used to describe non-human nouns, its grammatical gender is preserved in both singular and dual forms, but not with the plural where a singular feminine form of the adjective is used instead of the plural form. The following examples show this issue:

Table4: Noun agreement markers in Arabic with the adjective big

Person	Gender	Singular	Dual	Plural
First	Mas.	كبير - kabeer	كبيران - kabeeran	كبار – kebar
	Fem.	كبيره - kabeerah	كبيرتان - kabeertan	كبيرات - kabeerat
Second	Mas.	كبير - kabeer	كبيران - kabeeran	كبار - kebar
	Fem.	كبيره - kabeerah	كبيرتان - kabeertan	كبيرات - kabeerat
Third	Mas.	كبير - kabeer	كبيران - kabeeran	كبار – kebar
	Fem.	كبيره - kabeerah	كبيرتان - kabeeratan	كبيرات – kabeerat

1) The children are beautiful "jameloon" الاطفال جميلون
   The birds are beautiful "jamela الطيور جميلة
2) The two children are beautiful "jamela" الطفلان جميلان
   The two birds are beautiful"jamelan" الطائران جميلان
3) The child is beautiful "jamelon" الطفل جميل
   The bird is beautiful "jamelon" الطائر جميل

In the second sentence of the first example, we used the adjective "jamelah جميلة " which is singular feminine form with a plural noun "*the birds*", while it is "jameloon جميلون " with "*the children*" in the first sentence of the same example; The difference between the two sentences is the humanity feature; children are human while the birds are not. Examples 2 and 3 preserve the gender of both nouns *child and bird* as it is in the singular and dual forms respectively.

Whatever mentioned are just few examples to explain that we need a huge amount of rules and exceptions during the implementation of translation algorithms, these rules are necessary to handle the agreement problem either between subject and verb, or between noun and adjective.

3. Handling agreement in MT from English to Arabic

This section introduces in general how the translators handle the required agreements. In example-based MT the problem is solved by creating huge amount of examples containing pairs of statements in the source and target languages; when the translator process any sentence it looks for the best match example(s) and use it; the main problem here is that we need a database with huge number of examples; even though we cannot cover all possible statements of any language. On the other hand, rule-based machine translators solve the problem by building enough number of rules and lexicons; the lexicons should include besides the meaning a lot of features that affect the translation such as number, gender, person, case, humanity, and alive features. The hybrid approach use example-based whenever a good match exists in the database, otherwise it uses the available rules to create the translation.

The different derivations in Arabic for verbs and adjectives can be generated by adding suffixes (prefix and/or infix and/or postfix) to the stem verb or adjective based on the features of the related subject. In some cases there is a need for pre-derivation process before the addition of suffixes; examples are erasing, replacing or changing a character of the root verb, and these irregular cases can be handled by building a special lexicon for them.

A question is: do we have a translator from English to Arabic that handles the agreements completely? If no, is it possible to satisfy this or at least enhance the output quality?

We can say that there exist good translators, but none of them solve the agreement problem completely. (Abu shquier and Sembok) manually evaluated four different English-Arabic translators for handling agreement and words ordering problems; their methodology consists of three steps: pass each sample source text to the translator, compare the output with a human translation, and then give a grade; their results showed that these translators satisfied the percentages 92.2,84.6, 94.2, and 96.1 for Al-Kafi, Google, Tarjim Sakhr, and EA-RBMT respectively as shown in table 5. (Abu Shquier Mohammad *et al*, 2008a)

Regarding the second part of the question, we think that it is impossible to get 100% agreement in English-Arabic MT since lexicons cannot cover all phrases in any language, and cannot contain all names of persons, places, and things; furthermore if a sentence for example says "the consultant engineer said …", then the translator will assume that the engineer is male which is not always correct; Assume that the engineer in the sentence is a female person, then the sentence should be translated to "ALMOHANDESAH ALESTISHARIAH المهندسة الاستشارية" instead of "ALMOHANDES ALESTISHARI المهندس الاستشاري". We can also say, even if we can build the huge number of rules needed to handle the agreement of the standard cases, the handling of the exceptional cases still a challenge.

In the next section we proposed a technique to enhance the quality of gender agreement in English-Arabic MTs; This is because many nouns in English are used with both masculine and feminine, examples are: employee, leader, engineer, teacher, programmer, driver, passenger, writer, player, … etc. To get a correct Arabic translation we need to know the exact gender of the subject.

Table5: results of Abushguier and Sembok experiment

Cat. / Translator		ALKAFI	GOOGLE	TARJIM	EA-RBMT
Matches sentences	Number	115	50	125	141
	Total score	575	250	625	705
Mismatches or partially matches sentences	Number	56	121	46	30
	Total score	213	474	181	117
Total score out of 855		788	724	806	822
Percentage (%)		92.2	84.6	94.2	96.1

4. Proposed Technique

This proposal aims to enhance the manipulation of handling gender agreement in MT from English to Arabic. The main idea of the proposal is to identify the exact gender of the subjects that have neutral gender in the English such as engineer, teacher, employee, driver, and so on. This translation preprocess definitely will improve the output quality. The technique depends on building extra lexicons and applies such semantic analysis.

In our opinion, any rule-based product should maintain a database of:

a) A lexicon of all original words/phrases and their derivations in the source and target languages, this lexicon includes the words' meaning and their features such as number, gender, person, case, humanity, and alive.
b) A lexicon of the foreign words/phrases in the languages with their features.
c) A lexicon of the irregular words/phrases in the languages with their features.

Set of rule tables for grammar, derivation, pre-derivation, determinant, and others to be used in the translation process.

In this research we suggest maintaining extra lexicons for names of persons, places, things, and abbreviations. We know that it is impossible to build these lexicons completely, but they can contain as much as possible and to be updated periodically. To build these lexicons we can obtain help from the internet resources and from the governments departments that have such information. These lexicons will definitely help in identifying the subject gender.

We also suggest doing more analysis on the source text to get extra features about the subjects and nouns in the text through the correlation between sentences. To explain the idea, assume that the source text is "*The project engineer hold a meeting with the team members, she discussed a lot of issues regarding the project phases*"; In this example if we analyze the correlation between the two sentences in the context, we can discover that the gender of *engineer* is a feminine since the second statement starts with the pronoun "she", and hence we can get more accurate translation. Here is the analysis algorithm that we proposed to be applied as a pre-process before passing the source text to the translation process. A DFD of the process is shown in figure-2.

Step 1: For each sentence in the source text:
- 1.1: Assign a sequence number for the sentence, and fill them in the SENTENCE-DB table that described in table 6 bellow.
- 1.2: Identify the nouns and pronouns using a parser.
- 1.3: For each noun or pronoun:
  - 1.3.1: Get the gender feature (masculine, feminine, or neutral) from the lexicons.
  - 1.3.2: Add a record to the GENDER-DB table that described in table 7 bellow.

Step 2: For each sentence in the source text:
- 2.1: Find its relations with other sentences.
- 2.2: For each relation:
  - 2.2.1: Add a record to the RELATIOS-DB table that described in table 8 bellow.

Step 3: For each record in GENDER-DB with neutral gender, modify the gender value by consulting the relationships in RELATIOS-DB.

Table 6: SENTENCE-DB table

Sentence #	Sentence

Table 7: GENDER-DB table

Sentence #	Noun / Pronoun	Gender (M: masculine F: feminine N: neutral)

Table 8: RELATIOS-DB table

Sentence #	Noun/ Pronoun	Related sentence #	Noun / Pronoun in the related sentence

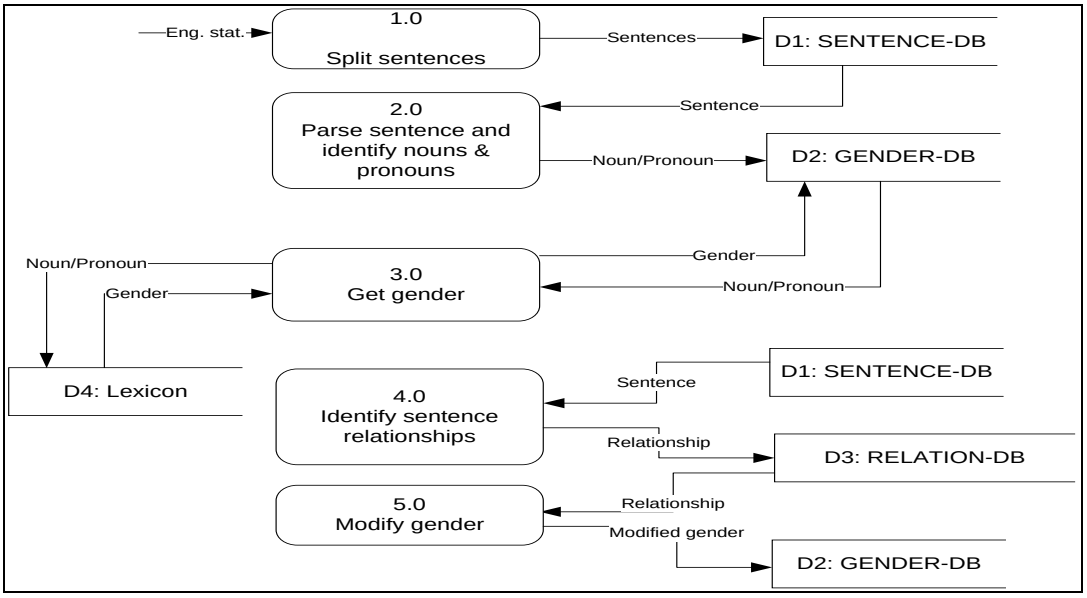

Figure- 2

5. Experiments and Results Discussion

Most of the available English – Arabic machine translators are black box commercial products, and hence it is difficult - if not impossible - to apply our proposal on them. We tested our proposed algorithm on two partial translators: the first one has been developed by Abu Shqeer and Kong as part of a MSc thesis (Abu shqeer Omer *et al*, 2002), and the second has been developed by Abu Shquier and Sembok as part of a PhD thesis (Abu Shquier Mohammad *et al*, 2008b); Furthermore we built the RELATIONS-DB table - the output of step 2 in the algorithm for the test cases – manually since it is out of scope of this research to build the semantic analyzer.

Example: Assume we have the following source text: *'The project engineer was worried about the next day's duty; she woke-up early, made a call for her friend and consulted her in a topic*'. One of the best MT could be:

'مهندس المشروع كان قلقا من مهمة اليوم التالي، استيقظت مبكرا واتصلت مع صديقها واستشارتها في موضوع'

If we can identify that the engineer is female through the pronoun *'she'* in the second sentence and we know that the friend is female through the pronoun *'her'*, then we can get the perfect translation:

'مهندسة المشروع كانت قلقة من مهمة اليوم التالي، استيقظت مبكرا واتصلت مع صديقتها واستشارتها في موضوع'

However, we can say that unless we look for clues in previous or subsequent sentences; some words like the *engineer* and *friend* in this example will take the default value of masculine. If we apply the algorithm on this example, we will get the database entries as shown in tables 6A, 7A, and 8A as a result of steps 1 and 2. By applying step 3 some entries in the table 7A will be modified as shown in table 7B; the gender of both *engineer* and *friend* have been changed to feminine instead of neutral. This change reflected positively on the translation output.

Table 6A: SENTENCE-DB table as a result of step 1

Sentence #	Sentence
1	The project engineer was worried about the next day duty
2	she woke-up early
3	she made a call for her friend and consult her in a topic

Table 7A: GENDER-DB table as a result of step 1

Sentence #	Noun / Pronoun	Gender (M: masculine F: feminine N: neutral)
1	project	M
1	engineer	N
1	day	M
1	duty	F
2	she	F
3	she	F
3	call	F
3	her	F
3	friend	N
3	her	F
3	topic	M

Table 8A: RELATIOS-DB table as a result of step 2

Sentence #	Noun / Pronoun	Related sentence #	Noun / Pronoun in the related sentence
1	engineer	2	she
1	engineer	3	she
3	friend	3	Her

Table 7B: GENDER-DB table as a result of step 3

Sentence #	Noun / Pronoun	Gender (M: masculine F: feminine N: neutral)
1	project	M
1	engineer	F
1	day	M
1	duty	F
2	she	F
3	she	F
3	call	F
3	her	F
3	friend	F
3	her	F
3	topic	M

As a result, the gender agreement in the translation output has been significantly improved. Almost 99% gender agreement satisfied for all subjects that had been gender-wise identified. This means that if we apply a robust semantic analyzer to produce the RELATIONS-DB table correctly and completely, then we can get an accurate gender agreement generation in English – Arabic MT.

Conclusions

In this paper we clarified the importance of handling agreement in MT from English to Arabic. The research showed that the gender feature should receive more concern from the researchers and developers of machine translators. We also explained that there exist English-Arabic machine translators with acceptable level of service, but none of them handles the agreement problem completely.

Finally, the research proved that more enhancements on the machine translation are possible; the output of the experiments of the proposed technique is evidence, where a significant improvement has been seen through the results. Gender agreement was the concern of this research, but other agreement features can get benefits by using similar techniques.

References

Abu Shqeer Omer, Kong Tang, (2002), Handling Agreement and Words Reordering in Machine Translation From English to Arabic; *MSc thesis, School of Computer Science, Universiti Sains Malaysia*.

Abu Shquier Mohammad, Sembok T, (2008a), Word Agreement and ordering in English-Arabic machine Translation, *Proceeding of the International Symposium on Information Technology*, pp: 1-10, IEEE Xplore Press, USA.

Abu Shquier Mohammad, Sembok T, (2008b), Word Agreement and Ordering in English-Arabic Machine Translation: Evaluational Approach, *Ph.D thesis, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, Bangi, Selangor*.

Al-Jarf Reim, (2000), Grammatical agreement errors in L1/L2 translations", IRAL: International Review of Applied Linguistics in Language Teaching, 38.1, pp 1-15.

Attia Mohammad, (2002), Implications of the Agreement Features in Machine Translation; *MSc thesis, Faculity of Languages and Translation, Al-Azhar University*.

El Kholy Ahmed, Habash Nizar, (2012), Orthographic and Morphological Processing for English-Arabic Statistical Machine Translation", Machine Translation, Vol 26, issue 1-2.

Elming Jakob, Habash Nizar, (2009), Syntactic Reordering for English-Arabic Phrase-Based Machine Translation, *Proceedings of the EACL2009 Workshop on Computational Approaches to Semitic Languages*, Athens, Greece, pp 69–77.

Nasr T. Raja, *(1967), The Structure of Arabic: From Sound to Sentence*, Librairie du Liban, Beirut.

http://ice.he.net/~hedden/intro_mt.html, Thomas D. Hedden, machine translation:a breif introduction, cited on 1/3/2012.

http://www.itl.nist.gov/iad/mig/tests/mt/doc/ngram-study.pdf, Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics, cited on 2/3/2012

http://www.sajan.com, Machine translation: Can the quality really exist? , cited on 5/3/2012