

# Design a Corpus Based Approach for Bilingual Ontology Arabic-English

Ahmed R. Elmahalawy <sup>[1]</sup>, Mostafa M. Aref <sup>[2]</sup>

Department of Mathematics <sup>[1]</sup>, Faculty of Science  
Benha University, and Benha

Department of Computer Science <sup>[2]</sup>, Faculty of Computer and Information Sciences  
Ain Shams University, and Cairo  
Egypt

## ABSTRACT

This paper proposes a description of the bilingual ontology (Arabic-English) by using a class of object oriented programming to define a concept of noun and verb. Describe the bilingual hierarchy of noun and verb concepts. We have designed a three algorithms of corpus based for bilingual ontology such as: preprocessing, (matching & alignment) and update. Make a two cases to obtain the noun and verb concepts of the bilingual ontology.

**Keywords :-** Ontology, Bilingual Ontology (Arabic - English), Corpus based approach, concepts by using object oriented programming, noun and verb concepts.

## I. INTRODUCTION

This paper presents a Design a Corpus Based Approach for Bilingual Ontology that can be used to describe the concepts by using classes. The paper is organized as follows: Section (II) gives a background on ontology, bilingual ontology and Machine Translation. Section (III) gives a more related work of bilingual ontology and machine translation. Section (IV) describe the bilingual ontology by using class of object oriented programming to define a noun and verb concepts. From the noun and verb concepts we build the hierarchy. Section (V) make a design of corpus base approach for bilingual ontology by using three algorithms such as pre-processing, matching & alignment and updating to build a new hierarchy. Section (VI) gives a two cases to apply the three algorithms. Section (VII) gives a conclusion and a future work.

## II. BACKGROUND

Machine translation is a made automated translation. This system implemented by utilizing a computer software to transform a text from a naturalistic language (such as Arabic) to another language (such as English) without any human involution. The machine translation process is shown in Fig. 1 **Error! Reference source not found.** [1].

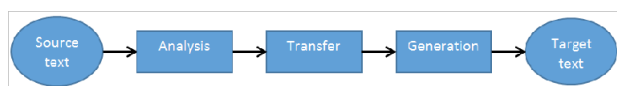


Fig. 1 Machine Translation Process

Ontology is a debate here in the applied context of software and database engineering, yet it has a theoretical grounding as well. An ontology gives details of a range of words with which to make statements, which may be inputs or outputs of knowledge agents (such as a software program) [2].

Bilingual is the most general expressions that are utilized when we speak about people who speak two languages. For example, a bilingual person might talk Arabic and English or any other two languages. How we make ability to speak two languages mostly depends on the person who works to find information and his make observations in the form of questions, or the policy maker and his statutory policy [3].

The word of Bilingual is divided into two parts: the first part is "Bi" which means (having two) and the second part is lingual which means (language), thus bilingual which means (having two languages). Bilingual is as well a noun, and a person can be called a bilingual, such as in the South American country like Canada, where the official languages are French and English, and where many of the citizens are bilingual [4].

## III. RELATED WORK

There are many related work depend on machine translation, ontology, corpus based and bilingual ontology.

In [5], the authors give a detail about the semi-automatic process of associating a Japanese word list with a semantic concept taxonomy are called an ontology, utilizing an English-Japanese bilingual dictionary. This problem focuses on how to connect the Japanese lexical things with the concepts in the ontology by automatic ways, so it is also hard to know many concepts manually. We have prepared a three algorithms to connect the Japanese lexical things with the concepts such as: the equivalent-word match, the argument match, and the example match.

In [6], the researcher describes an alignment system that aligns Malayalam - English texts at word level in parallel sentences. A parallel corpus is a combination of texts in two various languages, one of whom language is translated to tantamount of the second language. So, the prime objective of

this method is to construct word-aligned parallel corpus to be utilized in Malayalam and English machine translation (MT).

In [7], the authors developed the paper in [6]. Parallel corpus are assist in to create the statistical bilingual dictionary, in backing statistical machine translation and also in supporting as traineeship data for word meaning and translation disambiguation. Furthermore, the presentation of this approach can too be progressed by utilizing a listing of equations and morphological analysis.

In [8], the researchers describe the methodology to know the parallel Hindi-English sentences by utilizing a word alignment. This methodology is basis to improve the parallel Hindi-English word dictionary after syntactically and semantic analysis of the original text from Hindi-English. Develop this methodology is depend on two ways to solve this problem. The first way: is normalization of tagged Hindi-English sentences. The second way: is a mapping of Hindi-English sentence by utilizing parallel Hindi-English word dictionary.

#### IV. DESCRIPTION OF BILINGUAL ONTOLOGY

In this section the description of bilingual ontology is going to be focusing on a part of speech (POS) from the concept of noun and verb. The noun concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Hyponyms in the concepts of English and Arabic. The verb concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Troponyms in the concepts of English and Arabic.

The bilingual ontology is a set of concepts in two languages (Arabic - English), one of which is the translation equivalent of the other. The bilingual ontology described by using the concepts. The concept is defined by using a class from object oriented programming.

The related concepts consist of two words in two different languages. The concept is divided into two concepts noun and verb. Each of them split into several concepts. By using a class of object oriented programming to describe the concept of English and Arabic.

The general description of noun concept is defined by using a class, as illustrated in Fig. 2. From the class definition we define the symbol and characters as the following:

- The symbol (#) means the number of
- $N_S$  = number of Synonyms
- $N_E$  = number of Hypernyms.
- $N_O$  = number of Hyponyms.

N-concept		
Semantic relations	#	Concepts
# Synonyms:	$(N_S)$	concept 1 - .... - concept $(N_S)$
# Hypernyms:	$(N_E)$	concept 1 - .... - concept $(N_E)$
# Hyponyms:	$(N_O)$	concept 1 - .... - concept $(N_O)$
# المرادفات:	$(N_S)$	١ مفهوم - .... - $(N_S)$ مفهوم
# الاشتغال:	$(N_E)$	١ مفهوم - .... - $(N_E)$ مفهوم
# التضمين:	$(N_O)$	١ مفهوم - .... - $(N_O)$ مفهوم

Fig. 2 A General description of the noun concept

An example of a noun concept is "person". It has three senses in English and ten senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Hyponyms of the concept of person is described, as illustrated in Fig. 3.

N-person		
Semantic relations	#	Concepts
# Synonyms:	3	individual - human body - grammatical category
# Hypernyms:	6	organism - living - unit - object - physical entity - entity
# Hyponyms:	1	Adult
# المرادفات:	10	ذات - جسد - بشر - نفر - أفتوم - الذات - النفس - إنسان - فرد - شخص
# الاشتغال:	6	كيان - كيان مادي - شئ - وحده - شئ - كائن حي
# التضمين:	1	بالغ

Fig. 3 A Description of the noun person

Another example of a noun concept is "dinner". It has two senses in English and four senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Hyponyms of the concept of dinner is described, as illustrated in Fig. 4.

N-dinner		
Semantic relations	#	Concepts
# Synonyms:	2	main meal - dinner party
# Hypernyms:	7	meal - nutriment - food - substance - matter - physical entity - entity
# Hyponyms:	1	high tea
# المرادفات:	4	وجبة الطعام الرئيسية - حفظ - غداء - عشاء
# الاشتغال:	7	كيان - كيان مادي - شئ - مادة - طعام - غذاء - وجبة
# التضمين:	1	شاي عالي

Fig. 4 : A Description of the noun dinner

Another example of a noun concept is "car". It has five senses in English and four senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Hyponyms of the concept of car is described, as illustrated in Fig. 5.

N-car		
Semantic relations	#	Concepts
# Synonyms:	5	automobile - railcar - gondola - elevator car - cable car
# Hypernyms:	7	motor vehicle - self-propelled vehicle - wheeled vehicle - vehicle - transport instrumentation - artifact - unit - object - physical entity - entity
# Hyponyms:	1	ambulance
# المرادفات:	4	عربة قطار - مركبة - عربة - سيارة
# الاشتغال:	7	المواصلات - مركبة - عربة بعجلات - مركبة ذاتية الدفع - السيارات - الأجهزة - كيان - كيان مادي - شئ - وحدة - الأداة - الأجهزة
# التضمين:	1	سيارة إسعاف

Fig. 5 A Description of the noun car

Another example of a noun concept is "teacher". It has two senses in English and four senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Hyponyms of the concept of teacher is described, as illustrated in Fig. 6.

N-teacher		
Semantic relations	#	Concepts
# Synonyms:	2	instructor - teacher
# Hypernyms:	10	educator - professional - adult - person - organism - living thing unit - object - physical entity - entity
# Hyponyms:	1	instrumentation - artifact - unit - object - physical entity - entity art teacher
# المرادفات:	4	مُهذَّب - مدرسه - الدرس - معلم
# الاشتغال:	10	ثِي - وحده - ثِي حِي - كائن حِي - شخص - بالغ - محترف - مرب
# التضمين:	1	كيان - كيانه مادي معلم الفن

Fig. 6 A Description of the noun teacher

The hierarchy are built of the bilingual ontology by using all the previous of noun concepts such as (person, dinner, car and teacher) as shown in Fig. 7.

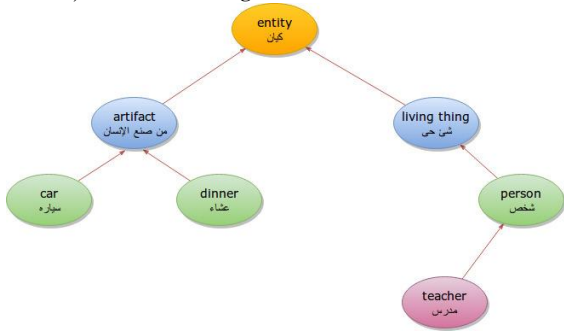


Fig. 7 A Description of the noun hierarchy

The general description of verb concept is defined by using a class, as illustrated in Fig. 8. From the class definition we define the symbol and characters as the following:

- The symbol (#) means the number of
- $N_S$  = number of Synonyms
- $N_E$  = number of Hypernyms.
- $N_T$  = number of Troponyms.

V-concept		
Semantic relations	#	Concepts
# Synonyms:	( $N_S$ )	concept 1 - .... - concept ( $N_S$ )
# Hypernyms:	( $N_E$ )	concept 1 - .... - concept ( $N_S$ )
# Troponyms:	( $N_T$ )	concept 1 - .... - concept ( $N_T$ )
# المرادفات:	( $N_S$ )	مفهوم - مفهوم - .... - ( $N_S$ ) مفهوم
# الاشتغال:	( $N_E$ )	مفهوم - مفهوم - .... - ( $N_E$ ) مفهوم
# المجاز:	( $N_T$ )	مفهوم - مفهوم - .... - ( $N_T$ ) مفهوم

Fig. 8 A General description of the verb concept

An example of a verb concept is "eat". It has six senses in English and seven senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Troponyms of the concept of eat is described, as illustrated in Fig. 9.

V-eat		
Semantic relations	#	Concepts
# Synonyms:	6	take in - eat - feed - eat on - consume - rust
# Hypernyms:	1	consume
# Troponyms:	1	wash down
# المرادفات:	7	ذهب للأكل - أكرهه على - تغدى - أكل الوجبه - تناول الفطار - التهم - أكل
# الاشتغال:	1	استهلك
# المجاز:	1	غسيل للأسفل

Fig. 9 A Description of the verb eat

Another example of a verb concept is "go". It has 26 senses in English and 17 senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Troponyms of the concept of went is described, as illustrated in Fig. 10.

V-go		
Semantic relations	#	Concepts
# Synonyms:	26	move - proceed - depart - become - awarded - run - lead - went - discarded go - sound - work - run low - run - survive - elapse - die - belong - start blend - lead - fit - rifle - spent - plump - give way
# Hypernyms:	1	move
# Troponyms:	1	swap
# المرادفات:	17	حدث - استهلك - غادر - مضى - أصبح - قال - خرج - سافر - مضى - انطلق - ذهب
# الاشتغال:	1	أدى - اعترم - ساعد على - عرف ب - دار - لجا
# المجاز:	1	تحرك
# المجاز:	1	تبادل

Fig. 10 A Description of the verb go

Another example of a verb concept is "become". It has four senses in English and three senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Troponyms of the concept of become is described, as illustrated in Fig. 11.

V-become		
Semantic relations	#	Concepts
# Synonyms:	4	get - turn - come into existence - suit
# Hypernyms:	2	change state - change
# Troponyms:	1	break
# المرادفات:	3	لاق ب - محمل - أصبح
# الاشتغال:	2	تحرك
# المجاز:	1	تغيير - تغيير الدوله

Fig. 11 A Description of the verb become

Another example of a verb concept is "do". It has 13 senses in English and ten senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Troponyms of the concept of be is described, as illustrated in Fig. 12.

V-do		
Semantic relations	#	Concepts
# Synonyms:	13	make - execute - perform - fare - cause - practice - answer - create behave - serve - manage - arrange - do
# Hypernyms:	1	create
# Troponyms:	1	overdo
# المرادفات:	10	سبق إلى تمثيل - عين - تالف - ابتدع - اتج - أحدث - ابداع - خلق
# الاشتغال:	1	كان أول من مثل كذا - لق
# المجاز:	1	خلق
# المجاز:	1	تطرف

Fig. 12 A Description of the verb do

The hierarchy are built of the bilingual ontology by using all the previous of verb concepts such as (eat, went, become and do) as shown in Fig. 13.

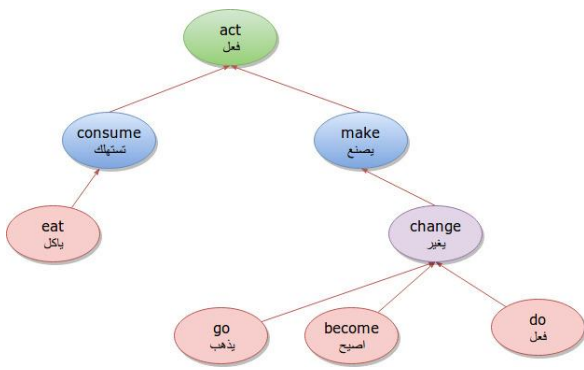


Fig. 13 A Description of the verb hierarchy

## V. DESIGN OF CORPUS BASED APPROACH FOR BILINGUAL ONTOLOGY

In this new division of page we will present the different approaches utilized in each step. There are three different steps to this part as make obvious by picture in Fig. 14 and we will describe the three distinct steps in the subsections A, B and C.

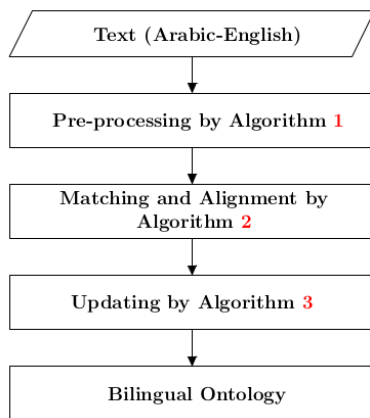


Fig. 14 Fig: Architecture of (Arabic-English) sentences

### A. Pre-processing Algorithm

Pre-processing Algorithm is an important step in the design of corpus based approach for bilingual ontology to remove all Arabic and English stop words from the sentences as show in Algorithm. 1. We will describe the work of pre-processing Algorithm as: Given an Arabic language (A) and English language (E). The Arabic sentence  $A = A_1, A_2, \dots, A_r, \dots, A_{LA}$  for length  $LA$  and the English sentence  $E = E_1, E_2, \dots, E_k, \dots, E_{LE}$  for length  $LE$ .

The sentences of Arabic and English contain a number of stop words and after removing all stop words from a list of stop words, then we find a new sentence of two languages (Arabic and English).

```

1: Start with a list of stop words Arabic ( $L_1$ ) and English ( $L_2$ ).
2: Accept the Arabic sentence ( $A$ ) and the English sentence ( $E$ );
3: for each Arabic sentence  $A$  do
4:   Separate the sentence into words
5:   if word is found in a list ( $L_1$ ) then
6:     Removing the stop word
7:   else
8:     Store the word in a new list
9:   end if
10: end for
11: for each English sentence  $E$  do
12:   Separate the sentence into words
13:   if word is found in a list ( $L_2$ ) then
14:     Removing the stop word
15:   else
16:     Store the word in a new list
17:   end if
18: end for
    
```

Algorithm. 1 Preprocessing Algorithm

### B. Matching and Alignment Algorithm

After the pre-processing algorithm a matching and alignment algorithm is used to make matching between two words in Arabic words and English words. The algorithm as illustrated in Algorithm. 2.

```

1: Accept the Arabic sentence  $A$  and the English sentence  $E$ ;
2: Let  $A_E$  = set of English concepts based on Arabic word;
3: for each Arabic word do
4:   if search in  $A_E$  is found then
5:     Match the meaning with English word in  $A_E$ 
6:   else
7:     Update and make a new concept in  $A_E$ 
8:   end if
9: end for
    
```

Algorithm. 2 Matching and Alignment Algorithm

### C. Updating Algorithm

After the matching and alignment algorithm a updating algorithm is utilized to build the bilingual ontology (Arabic - English) which contain words to translate into other words. The algorithm of the corpus based approach for bilingual ontology as illustrated in Algorithm. 3.

```

1: Accept the Arabic sentence  $A$  and the English sentence  $E$ ;
2: for each Arabic word do
3:   if Match is found then
4:     Corpus based Arabic word with English
5:   else
6:     Update and add in a hierarchy of the bilingual ontology
7:   end if
8: end for
    
```

Algorithm. 3 Updating Algorithm

## VI. CASE STUDIES

In this approach based on bilingual ontology the problems of concepts are divided into two various problems in the subsections D and E:

### D. Case 1

The first problem: is to find a new concept as a noun or a verb in Arabic and English languages. This new concept is not defined in the previous hierarchy of the bilingual ontology.

**Solution:** This new concept as a noun or a verb is defined by using a class. This concept is added in the previous hierarchy of the bilingual ontology, to get a new hierarchy of the bilingual ontology.


**Example:** We have two sentences input of Arabic (A) and English (E) languages as the following:

E Sentence: Ahmed went to the college by car  
 A Sentence: ذهب أحمد إلى الكلية بالسيارة

Applying the first step to remove all English and Arabic stop words from the list by using a pre-processing algorithm in Algorithm. 1, then we get the new sentences as:

E Sentence: Ahmed went college car  
 A Sentence: ذهب أحمد الكلية السيارة

Applying the second step to make alignment between the two new sentences by using an alignment algorithm in Algorithm. 2, then we find:

E Sentence: Ahmed went college car  
  
 A Sentence: ذهب أحمد الكلية السيارة

From the alignment algorithm we get a new noun concept (college - الكلية). The concept of "college" which has three senses in English and four senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Hyponyms of the concept of college is described, as illustrated in Fig. 15.

N-college		
Semantic relations	#	Concepts
# Synonyms:	3	body - institution - building complex
# Hypernyms:	7	building complex - construction - artifact - unit - object physical entity - entity
# Hyponyms:	2	junior college - normal school
# المرادفات:	4	مبنى الكلية - حشد - مجمع - كلية
# الاشتغال:	7	شئ - وحده - من صنع الانسان - اعمال بناء - مجمع المباني كيان - كيان مادي
# التضمين:	2	المدرسة العادية - كلية البتدئين

Fig. 15 A Description of the noun college

Applying the third step to make an updating in the hierarchy of the bilingual ontology by using an updating algorithm in Algorithm. 3. To get the new hierarchy of the bilingual ontology by using all the previous of noun concepts such as (person, dinner, car and teacher) and add a new concept "college" as shown in Fig. 16.

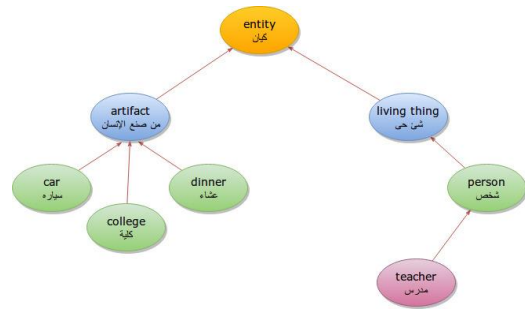


Fig. 16 A Description of the noun hierarchy after adding a college


**Example:** We have two sentences input of Arabic (A) and English (E) languages as the following:

E Sentence: Thank Ahmed to go with me  
 A Sentence: شكرا أحمد للذهاب معي

Applying the first step to remove all English and Arabic stop words from the list by using a pre-processing algorithm in Algorithm. 1, then we get the new sentences as:

E Sentence: Thank Ahmed go  
 A Sentence: شكرا أحمد ذهاب

Applying the second step to make alignment between the two new sentences by using an alignment algorithm in Algorithm. 2, then we find:

E Sentence: Thank Ahmed go  
  
 A Sentence: شكرا أحمد ذهاب

From the alignment algorithm we get a new verb concept (thank - شكرا). The concept of "thank" which has one senses in English and two senses in Arabic and every word has one or more senses. Also, the Synonyms, Hypernyms and Troponyms of the concept of college is described, as illustrated in Fig. 17.

V-thank		
Semantic relations	#	Concepts
# Synonyms:	1	thank
# Hypernyms:	7	convey - impart - tell - inform - communicate - interact - act
# Troponyms:	2	acknowledge - appreciate
# المرادفات:	2	حمد الله - شكر
# الاشتغال:	7	فعل - تفاعل-نقل - إعلام - يخبر - عرف - نقل
# المجاز:	2	نقدر - اعترف

Fig. 17 A Description of the verb thank

Applying the third step to make an updating in the hierarchy of the bilingual ontology by using an updating algorithm in Algorithm. 3. To get the new hierarchy of the bilingual ontology by using all the previous of verb concepts such as (eat, go, become and do) and add a new concept "thank" as shown in Fig. 18.

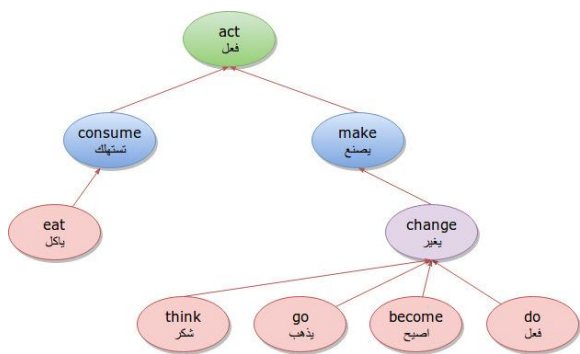


Fig. 18 A Description of the verb hierarchy after adding a thank

**E. Case 2**

The second problem: is to find a new ambiguous concept as a noun or a verb in Arabic and English languages. This new concept is not defined in the previous hierarchy of the bilingual ontology.

**Example:** We have two sentences input of Arabic (A) and English (E) languages as the following:

**E Sentence:** Ahmed went to the bank  
**A Sentence:** احمد يذهب الى البنك

Applying the first step to remove all English and Arabic stop words from the list by using a pre-processing algorithm in **Algorithm. 1**, then we get the new sentences as:

**E Sentence:** Ahmed went bank  
**A Sentence:** احمد يذهب البنك

Applying the second step to make alignment between the two new sentences by using an alignment algorithm in **Algorithm. 3**, then we find:

**E Sentence:** Ahmed went bank  
  
**A Sentence:** احمد يذهب البنك

From the alignment algorithm we get a new ambiguous concept called a "bank". If a concept is ambiguous, it can have one or more than meaning. The concept of "bank" which has two meaning in English the edge of a river, or a financial bank. In Arabic has two different meanings such as "البنك - حافه النهر".

**VII. CONCLUSIONS**

In this paper, we proposed a description of the bilingual (Arabic-English) ontology by using a class of object oriented programming to define a new concept of noun and verb. The noun concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Hyponyms in the concepts of English and Arabic. The verb concepts are going to discuss some semantic relations as the Synonyms, Hypernyms and Troponyms in the concepts of English and Arabic. Describe the bilingual hierarchy of noun and verb concepts.

We have applied a three algorithms of corpus based for bilingual (Arabic-English) ontology such as:

- 1) Pre-processing
- 2) Matching & Alignment
- 3) Update

From the description of bilingual ontology and design of corpus based approach for bilingual ontology to show the case studies. For a future work to make a big bilingual (Arabic-English) ontology by using a free open source called a protégé.

**ACKNOWLEDGMENT**

First, I would love to thank Allah. My honest gratitude goes to my family for their encouragement and support. I would like to thank my supervisor Prof. Mostafa Aref, who gave me some information and helping with my research. I would also like to thank my second supervisor Prof. Abdelkareem Abdelhaleem Soliman for helping me. My deep gratitude to all staff members of the Department of Mathematics, especially the Head of Department of Mathematics.

**REFERENCES**

- [1] C. Stern and A. Dufournet. What is machine translation? systran translation technologies, Machine translation, <http://www.systransoft.com/systran/translation-technology/what-is-machine-translation/>, August 2011. (Accessed 22-October-2015).
- [2] T. Gruber. Ontology (computer science) - definition in encyclopedia of database systems. Ontology, <http://tomgruber.org/writing/ontology-definition-2007.htm>, September 2007. (Accessed 14-October-2015).
- [3] N. Takaya. What do we mean when we say bilingual? — psychology in action. Bilingual, <http://www.psychologyinaction.org/2012/01/17/what-do-we-mean-when-we-say-bilingual/>, January 2012. (Accessed 31-October-2015).
- [4] I. Thinkmap. bilingual - dictionary definition : Vocabulary.com. Bilingual, <http://www.vocabulary.com/dictionary/bilingual>, June 2013. (Accessed 31-October-2015).
- [5] A. Okumura, E. Hovy. Building Japanese-English Dictionary based on Ontology for Machine Translation. In proceedings of ARPA Workshop on Human Language Technology, pages 236-241, 1994.
- [6] K. T. Nwet. Building Bilingual Corpus based on Hybrid Approach for Myanmar - English Machine Translation. International Journal of Scientific & Engineering Research, 2(9), 2011.
- [7] K. T. Nwet, K. M. Soe, and N. L. Thein. Developing Word-aligned Myanmar-English Parallel Corpus based on the IBM Models. International Journal of Computer Applications, 27(8), 2011.
- [8] S. Dubey, T. D. Diwan. Supporting Large English-Hindi Parallel Corpus using Word Alignment. International Journal of Computer Applications, 49, No.6,(7), 2012.