Feasibility of Bootstrapping an Arabic WordNet Leveraging Parallel Corpora and an English WordNet

Mona T. Diab*

*Linguistics Department
Margaret Jacks Hall
Stanford University
Stanford, CA 94305, USA
mdiab@stanford.edu

Abstract

In this paper, we propose the automatic bootstrapping of a Modern Standard Arabic WordNet on the lexeme level using Arabic English parallel corpora and an English WordNet. We address the feasibility of such an endeavor and present a qualitative evaluation of the meaning correspondences cross linguistically between Arabic and English. We further present an automatic means of performing this task using an unsupervised Word Sense Disambiguation System. We test the feasibility of the bootstrapping by qualitatively evaluating the meaning definition projection of English words onto their Arabic translations. We manually evaluate 447 word instances of the Arabic words that correspond to correctly sense tagged English words using English WordNet 1.7. from the SENSEVAL 3 data. The words evaluated correspond to Nouns, verbs, adjectives in English. We find that for Arabic verbs, adjectives and nouns, on average 52.3% of all the words examined, the corresponding English WordNet set of definitions are sufficient as definitions for the Arabic translation word; 39.96% of the Arabic words correspond to specific subsets of the WordNet definitions; and finally, 7.8% of the Arabic words comprise supersets of their corresponding English WordNet translation definitions. These results are very encouraging as they are similar to those obtained by researchers building EuroWordNet.

1. Introduction

Hierarchical taxonomies are proving to be the preferred forms of lexical knowledge representation utilized in natural language processing applications. An example of such taxonomies is WordNet. Due to their popularity and the existence of tools for manipulating their underlying structure, several teams from different countries are currently developing WordNets for different languages. Efforts in the domain of taxonomy creation have mostly been manual. EuroWordNet exists for several languages: Dutch, Spanish, French, Czech, Italian and Estonian; EuroWordNet interfaces these different Ontologies with the Internal Language Index (ILI). The bootstrapping method starts with monolingual dictionaries for the new language, and an ontology is created in the WordNet format. Apart from the immense time investment in the bootstrapping phase, researchers are faced with the challenge of linking the created WordNet with existing WordNets and dealing with sense granularity issues which is one of the biggest challenges facing such an endeavor.

Having a method that leverages existing resources is a big plus as the manual task of creating an ontology such as WordNet is extremely expensive and genuinely daunting. The problem becomes even more challenging when the language in question is a language with scarce automatic knowledge resources such as Arabic. To date, there exists no such resource or effort for the creation of an Arabic WordNet. Therefore, we propose a method that relies on the existence of English Arabic parallel corpora and a WordNet taxonomy for English, where the content words of the English text are annotated with their meaning definitions (senses) as listed in WordNet. We apply an approach called SALAAM (Diab, 2003), where the token correspondences are found cross linguistically and then the senses associated with the English side of the parallel corpus are projected on the Arabic tokens. Given a large and diverse enough parallel corpus with good quality token alignments, this method can help bootstrap a large ontology for Arabic from scratch.

The appeal of building a WordNet for Arabic is not only based on empirical grounds for computational linguistic applications, but also it allows for an exploration of interesting lexical semantic cross-linguistic variations — albeit at this stage exclusively paradigmatic. Like other languages, Arabic lexemes exhibit the full range of ambiguity attributes from regular polysemy to metonymy and homonymy. Lexical ambiguity in modern standard Arabic is further compounded by the writing system: written texts in Arabic typically omit the short vowels leading to more ambiguity, creating false homonyms.

In this paper, we propose the automatic bootstrapping of a Modern Standard Arabic WordNet on the lexeme level using Arabic English parallel corpora and an English WordNet. We address the feasibility of such an endeavor and present a qualitative evaluation of the meaning correspondences cross linguistically between Arabic and English. We further present an automatic means of performing this task using an unsupervised Word Sense Disambiguation System.

We test the feasibility of the bootstrapping by qualitatively evaluating the meaning definition projection of English words onto their Arabic translations. We manually evaluate 447 word instances of the Arabic words that correspond to correctly sense tagged English words using English WordNet 1.7. from the SENSEVAL 3 data.[1]. The words evaluated correspond to Nouns, verbs, adjectives in English.

[1] http://www.senseval.org/

2. WordNet Taxonomies

A WordNet taxonomy (Fellbaum, 1998) is a computational semantic lexicon for English. It is rapidly becoming the community standard lexical resource for English since it is freely available for academic research. It is an enumerative lexicon in a Quillian style semantic network that combines the knowledge found in traditional dictionaries (Quillian, 1968). Words are represented as concepts, referred to as synsets, that are connected via different types of relations such as hyponymy, hypernymy, synonymy, meronymy, antonymy, etc. Words are represented as their synsets in the lexicon. For example, the word *bank* has 10 synsets in WN17pre corresponding to 10 different senses. The concepts are organized taxonomically in a hierarchical structure with the more abstract or broader concepts at the top of the tree and the specific concepts toward the bottom of the tree. For instance, the concept *FOOD* is the hypernym of the concept *FRUIT*, for instance. WordNet taxonomies comprise four part of speech databases corresponding to nouns, verbs, adjectives, and adverbs databases. The noun database is the richest of the four databases as it comprises approximately 69,000 concepts and has a depth of 15 nodes, nearly four times the size of the verbs database and three times the size of the adjectives database.

3. Representation Granularity for Arabic

Since this paper is concerned with building a WordNet for Arabic, representation granularity becomes an issue more than it is for English due to Arabic's rich morphological nature. Arabic has a templatic syntax; roots are transformed into stems based on a templatic fit. For example, the root **ktb**[2] becomes **kitab** based on the template **fiEal**. Stems are not necessarily the same as the lexeme. For instance the stem of the word **Hasanathm** meaning *their virtue* is **Hasanat** which is not the same as the lexeme *virtue*, **Hasanap** where the **t** is transformed into the taa marboutah **p**. Most traditional Arabic lexical resources list entries by reducing words to their roots. However, in our view, the lexeme level is the appropriate sense granularity level for natural language processing applications. Roots are the underlying forms from which stems and surface forms generate. Most words in Arabic can be reduced to 3 or 4 letter roots. Roots are typically consonant based. Arabic has generative templates that lead to the creation of stems. Roots are highly generative and typically very ambiguous. For instance, the root **$Er** is the root for *hair*, *poetry* or *to feel*, simultaneously. This could be treated as a case of homonymy that is resolved by applying the appropriate template; therefore, the lemma or lexeme for *hair* is **$aEr**, for *poetry* **$iEr** and for *to feel* it is *$aErA*. Similarly, the root **Hrm** generates **Haram** as in *shrine*, *sanctuary*, *wife* or *forbidden*; it is also the root for the *clothes worn by pilgrims* as in **IHrAm**, as well as the root for thief as in **HarAmy**.

Due to the pervasive ambiguity in the root representation, one would expect a huge overlap between the different POS databases in an Arabic WordNet.

We find the option of creating an ontology based on roots theoretically elegant, especially if the templates are not ambiguous. A root based ontology will have to be generative and under specified. The main bottleneck is extracting the root from a surface level representation since words do not occur in their root form in written nor spoken Arabic. Several off-the-shelf morphological analyzers may be utilized to reduce surface forms to their corresponding roots, yet coverage remains a severe bottleneck (Darwish, 2002; Buckwalter, ; Mona Diab, 2004).

On the other hand, a lexeme based ontology is a more direct approach to building an enumerative WordNet style sense inventory. Empirically, Arabic lexemes are more accessible by computational systems. Surface form words are typically affixed with clitics, However, there exist systems for reducing surface words to lexemes with very high accuracies (99%) such as `ArabicSVM_tools`[3] (Mona Diab, 2004). Lexemes are distinguishable as different POS tags based on the templates they correspond to in Arabic. One of the problems facing a lexeme level representation is normalization; Words referring to the same concept and root maybe written in various ways. The problem is rampant with the plural form of many nouns. For example, the word for *schools* in Arabic maybe **mdArs** or **mdrsAt**. The second form is mostly predictable but the former form is not. Given good low level processors of Arabic text and large amounts of training data, this ambiguity is resolvable with very high accuracy. Hence, lexemes are directly recoverable from given text which renders them a natural candidate for WordNet entries.

4. Proof of Concept

In our conceptual approach, we opt to build an Arabic WordNet leveraging an existing English WordNet. The crux of this work is based on the fact that paradigmatic relations are deeply semantic in nature that they, by and large, tend to hold cross linguistically. The idea is that if we know the Arabic translation of every node in an English WordNet we will have provided a very good starting point for bootstrapping an Arabic WordNet. However, in order for this approach to work, it is imperative to possess some means of finding correspondences between specific English synsets in the English WordNet and their Arabic counterparts. Our goal is to evaluate the quality of the correspondences between English senses and their Arabic correspondents; Do the Arabic words exhibit similar ambiguity? how accurate are the specific sense correspondences between the English WordNet synsets and their Arabic counterparts. Therefore, we qualitatively evaluate the 447 unique senses in the training data of the SENSEVAL 3[4]English Lexical Sample task (2004) and their Arabic correspondences. The SENSEVAL exercises were created in the late 1990's as a venue for evaluating the performance of word sense disambiguation systems. All SENSEVAL tasks presume a predefined set of senses from WordNet for automatic annotation of ambiguous words in text. The training data provided by SEN-

[2] All the Arabic in this paper is transliterated using the Buckwalter transliteration scheme.

[3] http://www.stanford.edu/mdiab/ArabicTools

[4] http://www.senseval.org

SEVAL is an invaluable resource since it comprises large amounts of manually annotated data.

4.1. Evaluation Data

The data considered here comes mainly from the British National Corpus comprising several different domains and genres. In this data set, only one ambiguous English word in a sentence is manually annotated with its appropriate sense from WordNet 1.7 by the organizers of the SENSEVAL 3 organizers. The Arabic translations are automatically rendered using the off-the-shelf Arabic Machine Translation System `Tarjim`.[5]. In an Arabic translation sentence, we manually identify the Arabic words corresponding to the sense annotated English word, therefore we are guaranteeing perfect token alignment. In this data set, there are 57 word types corresponding to 32 verbs (with 204 unique sense instances in the SENSEVAL3 provided training data), 5 adjectives (with 45 unique senses) and 20 nouns (with 198 unique sense instances). In total, they correspond to 447 unique sense annotations in the training data for this SENSEVAL 3 task, i.e. we only consider one instance of a sense for a given word type.[6] The 447 English sentences are translated into MSA Arabic using the Tarjim service. We clitic tokenize the Arabic sentences, then manually align the English tokens with their corresponding counterparts in Arabic. It is worth noting that the minimum number of senses for any word type is 3 senses.

4.2. Evaluation Criteria

For this qualitative evaluation, we have two sets of evaluation criteria. One set concerned with the quality of the correspondence between the Arabic word and the SENSEVAL 3 chosen sense for its correspondent English translation. We may judge the correspondence as *Accurate, Approximate, Mistranslation,* or *No Translation*. A correspondence is judged as *Accurate* if the English sense definition is a good description of the Arabic word given the Arabic context sentence. A correspondence is judged as *Approximate* if the there exists a sense definition in the set of available English sense definitions that is more appropriate for the Arabic word given its context sentence. A *Mistranslation* is the case where the English word is mistranslated by the MT system therefore by definition there will be no correspondent in the given English sense set to the Arabic word. Finally, a *No Translation* situation arises when the English word that is annotated is not translated at all.

The other criteria set is concerned with the correspondence level between the Arabic word and the set of English senses associated with its English translation. We set three sub criteria: *Equivalence, ArabicSpec, EnglishSpec*. An *Equivalence* criterion is warranted if all the senses rendered in the English WordNet taxonomy are appropriate and *sufficient* sense definitions for the Arabic word. An *ArabicSpec* criterion arises when the Arabic word corresponds to a subset of the senses in the English WordNet set for the specific English word. This occurs when the different senses of the English ambiguous word translates into several distinct words in Arabic, especially when the English word is homonymous. Finally, *EnglishSpec* arises when the Arabic word is equivalent to all the senses of the English word but the sense definitions are not sufficient for the Arabic words. This occurs when the Arabic word is homonymous.

4.3. Qualitative Results

Table 1 illustrates the results of the manual evaluation based on the first criterion set.

POS	Acc	App	Mis	NoT
Verb	83.7	8.7	6.3	1.3
Adj	71.1	6.6	20	2.2
Nouns	89	7	3	1

Table 1: Percentages of Manual Evaluation of the tagging correspondence between Arabic words and English Synset definitions from WordNet 1.7

The Adjectives perform the worst mainly due to the fact that the Machine translation system has limited variability for these specific adjectives. The 2 adjectives *solid* and *hot* are always translated as **Slb** and **sxn**, respectively. In many instances *solid* should have been translated as **slym** and *hot* should have been translated as **HAr**. In these cases, they are counted as mistranslations. Performance for the verbs is significantly less than for Nouns since verbs, on average, are more polysemous in WordNet than nouns and they exhibit more homonymy as exemplified by the increased difference of 3% in mistranslations for verbs.

Of the instances considered Accurate and Approximate, we apply the second set of evaluation criteria in order to judge the level of equivalence in meaning representation between the Arabic words and the English sense definitions from WordNet 1.7. Table 2 illustrates this evaluation by Part Of Speech tag.

POS	Equiv	ArabSpec	EngSpec
Verb	37.7	50	12.5
Adj	57.1	42.9	0
Nouns	62	27	11

Table 2: Percentages of manual evaluation of equivalence quality between Arabic words and English Synset definitions from WordNet 1.7

We note the existence of more ArabicSpecific equivalence across verbs and adjectives than nouns which is directly related to the high fertility in the number of senses for these parts of speech relative to the nouns. On average the number of senses for adjectives is eight senses and for verbs it is ten senses, while the number of senses for nouns is six senses. More importantly however is that the average Equivalent senses across the three parts is 52.3% which is consistent with the observation of EuroWordNet builders. Vossen, Peters, and Gonzalo (1999) find that approximately 44-55% of ambiguous words in Spanish, Dutch and Italian have relatively high overlaps in concept and their sense packaging of polysemous words (Vossen et al., 1999).

[5] http://www.tarjim.com/

[6] In the training data there are several instances of the same sense for the various word instances in different contexts.

5. Large Scale Automated Approach

The SALAAM (Sense Assignment Leveraging Annotations And Multilinguality) system provides us with a means of realizing the bootstrapping process on a large scale automatically.(Diab and Resnik, 2002; Diab, 2003; Mona Diab, 2004) SALAAM exploits parallel corpora for sense annotation. The key intuition behind SALAAM is that when words in one language, *L1*, are translated into the same word in a second language, *L2*, then the *L1* words are semantically similar. For example, when the English — *L1* — words *bank, brokerage, mortgage-lender* translate into the Arabic — *L2* — word *bnk* (بنك) in a parallel corpus,[7] where the *bank* is polysemous, SALAAM discovers that the intended sense for the English word *bank* is the *financial institution* sense, not the *geological formation* sense, based on the fact that it is grouped with *brokerage* and *mortgage-lender*. Two fundamental observations are at the core of SALAAM:

- **Translation Distinction Observation (TDO)**

  Senses of ambiguous words in one language are often translated into distinct words in a second language.

  To exemplify **TDO**, we consider a sentence such as *I walked by the bank*, where the word *bank* is ambiguous with n senses. A translator may translate *bank* into *Dfp* (ضفه) corresponding to the *GEOLOGICAL FORMATION* sense or to *bnk* (بنك) corresponding to the *FINANCIAL INSTITUTION* sense depending on the surrounding context of the given sentence. Essentially, translation has distinctly differentiated two of the possible senses of *bank*.

- **Foregrounding Observation (FGO)**

  If two or more words are translated into the same word in a second language, then they often share some element of meaning.

  FGO may be expressed in quantifiable terms as follows: if several words $(w_1, w_2, \ldots, w_x)$ in *L1* are translated into the same word form in *L2*, then $(w_1, w_2, \ldots, w_x)$ share some element of meaning which brings the corresponding relevant senses for each of these words to the foreground. For example, if the word *Dfp* (ضفه), in Arabic, translates in some instances in a corpus to *shore* and other instances to *bank*, then *shore* and *bank* share some meaning component that is highlighted by the fact that the translator chooses the same Arabic word for their translation. The word *Dfp* (ضفه), in this case, is referring to the concept of *LAND BY WATER SIDE*, thereby making the corresponding senses in the English words more salient. It is important to note that the foregrounded senses of *bank* and *shore* are not necessarily identical, but they are quantifiably the closest senses to one another among the various senses of both words.

Given observations **TDO** and **FGO**, the crux of the SALAAM approach aims to quantifiably exploit the translator's implicit knowledge of sense representation cross-linguistically, in effect, reverse engineering a relevant part of the translation process.

SALAAM's algorithm is as follows:

- SALAAM expects a word aligned parallel corpus as input;
- *L1* words that translate into the same *L2* word are grouped into clusters;
- SALAAM identifies the appropriate senses for the words in those clusters based on the words senses' proximity in WordNet. The word sense proximity is measured in information theoretic terms based on an algorithm by Resnik (Resnik, 1999);
- A sense selection criterion is applied to choose the appropriate sense label or set of sense labels for each word in the cluster;
- The chosen sense tags for the words in the cluster are propagated back to their respective contexts in the parallel text. Simultaneously, SALAAM projects the propagated sense tags for *L1* words onto their *L2* corresponding translations.

In this paper we focus on the last point in the SALAAM algorithm, namely, the sense projection phase onto the *L2* words in context. In this case, the *L2* words are Arabic and the sense inventory is the English WordNet taxonomy. Using SALAAM we annotate Arabic words with their meaning definitions from the English WordNet taxonomy. We justify the usage of an English inventory on both empirical and theoretical grounds. Empirically, there are no automated sense inventories for Arabic; Furthermore, to our knowledge the existing MRDs for Arabic are mostly root based which introduces another layer of ambiguity into Arabic processing since Modern Standard Arabic text is rendered in a surface form relatively removed from the underlying root form. Theoretically, we subscribe to the premise that people share basic conceptual notions which are a consequence of shared human experience and perception regardless of their respective languages. This premise is supported by the fact that we have translations in the first place. Accordingly, basing the sense tagging of *L2* words with corresponding *L1* sense tags captures this very idea of shared meaning across languages and exploits it as a bridge to explicitly define and bootstrap sense tagging in *L2*, Arabic. If we sense tag large amounts of Arabic text automatically and we have a means of identifying which items are accurately tagged, we may use those selected items to bootstrap an Arabic WordNet taxonomy. In (Diab, 2004b), we explore ways of automatically identifying accurately sense tagged data in English based on several factors. Once we have an actual system, the process of identification should not be a severe challenge.

[7] We use the Buckwalter transliteration scheme for the Arabic words in this paper. http://www.ldc.org/aramorph

6. Quantitative Evaluation

In order to formally evaluate SALAAM for Arabic WSD, there are several intermediary steps. SALAAM requires a token aligned parallel corpus as input and a sense inventory for one of the languages of the parallel corpus. For evaluation purposes, we need a manually annotated gold standard set.

6.1. Gold Standard Set

As mentioned above, there are no systems that perform Arabic WSD, therefore there exist no Arabic gold standard sets as such. Consequently, one needs to create a gold standard. Since SALAAM depends on parallel corpora, an English gold standard with projected sense tags onto corresponding Arabic words would serve as a good start. A desirable gold standard would be generic covering several domains, and would exist in translation to Arabic. Finding an appropriate English gold standard that satisfies both attributes is a challenge. One option is to create a gold standard based on an existing parallel corpus such as the Quran, the Bible or the UN proceedings. Such corpora are single domain corpora and/or their language is stylistic and distant from everyday Arabic; Moreover, the cost of creating a manual gold standard is daunting. Alternatively, the second option is to find an existing English gold standard that is diverse in its domain coverage and is clearly documented. Fortunately, the SENSEVAL exercises afford such sets.As mentioned above, SENSEVAL evaluations are community-wide exercises that create a platform for researchers to evaluate their WSD systems on a myriad of languages using different techniques by constantly defining consistent standards and robust measures for WSD. In this portion of the paper we use the SENSEVAL 2 English All-words data set from 2001.

Accordingly, the gold standard set used here is the set of 671 Arabic words instances corresponding to the correctly sense annotated English noun instances from the SENSEVAL2 English All Words Task. SALAAM achieved a precision of 64.5% and recall of 53% on the English test set for that task. SALAAM ranks as the best unsupervised system when compared to state-of-the-art WSD systems on the same English task. The English All Words task requires the WSD system to sense tag every content word in an English language text.

6.2. Token Aligned Parallel Corpora

The gold standard set corresponds to the test set in an unsupervised setting. Therefore the test set corpus is the SENSEVAL2 English All Words test corpus which comprises three articles from the Wall Street Journal discussing religious practice, medicine and education. The test corpus does not exist in Arabic. Due to the high expense of manually creating a parallel corpus, i.e. using human translators, we opt for automatic translation systems in a fashion similar to (Diab, 2000). To our knowledge there exist two off the shelf English Arabic Machine Translation (MT) systems: Tarjim and Almisbar.[8] We use both MT systems to translate the test corpus into Arabic. We merge the outputs of both in an attempt to achieve more variability in translation as an approximation to human quality translation. The merging process is based on the assumption that the MT systems rely on different sources of knowledge, different dictionaries in the least, in their translation process.

Fortunately, the MT systems produce sentence aligned parallel corpora.[9] However, SALAAM expects token aligned parallel corpora. There are several token alignment programs available. We use the GIZA++ package which is based on the IBM Statistical MT models.[10] Like most stochastic NLP applications, GIZA++ requires large amounts of data to produce reliable quality alignments. The test corpus is small comprising 242 lines only; Consequently, we augment the test corpus with several other corpora. The augmented corpora need to have similar attributes to the test corpus in genre and style. The chosen corpora and their relative sizes are listed in Table 3.

Corpora	Lines	Tokens
BC-SV1	101841	2498405
SV2-LS	74552	1760522
WSJ	49679	1290297
SV2AW	**242**	**5815**
Total	*226314*	*5555039*

Table 3: Relative sizes of corpora used for evaluating SALAAM

BC-SV1 is the Brown Corpus and SENSEVAL1 trial, training and test data. SV2-LS is the SENSEVAL2 English Lexical Sample trial, training and test data. WSJ is the Wall Street Journal. Finally SV2AW is SENSEVAL2 English All Words test corpus.

The three augmenting corpora, BC-SV1, SV2LS and WSJ are translated into Arabic using both MT systems, AlMisbar and Tarjim. All the Arabic corpora are transliterated using the Buckwalter transliteration scheme and then tokenized. The corpora are finally token aligned using GIZA++.

6.3. Sense Inventory

The gold standard set is annotated using the WordNet taxonomy, WN1.7pre, for English. This portion of the paper is exclusively on nouns.[11]

6.4. Experiment and Metrics

We conducted two human rating experiments which are described in detail in (Diab, 2004a). Briefly, the first experiment, the rater is asked to choose the appropriate sense definition for the Arabic word from the set of English WordNet definitions. The agreement rate between the human chosen senses and the SALAAM chosen senses is 90.1%. In the second experiment, three human subjects are asked to rate the quality of the SALAAM annotation. They deemed 90% of the SALAAM annotated data to be accurately tagged. It

[8] http://www.Tarjim.com, http://www.almisbar.com

[9] This is not a trivial problem with naturally occurring parallel corpora.

[10] http://www.isi.edu/och/GIZA++.html

[11] SALAAM, however, has no inherent restriction on part of speech.

is worth noting that the interannotater agreement is a high 96%.

7. General Discussion

It is worth noting the high agreement level between the rating judgments of the three raters in experiment 2 and the human manual annotations of experiment 1. The obtained results are very encouraging indeed but it makes the implicit assumption that the English WordNet taxonomy is sufficient for meaning representation of the Arabic words used in this text. In a similar vein to the qualitative evaluation in section 4.2., we qualitatively evaluate the quality of correspondence between the Arabic words and the English WordNet definitions. We use the three criteria described above.

With that intent in mind, we evaluate the 600 word instances of Arabic that are deemed correctly tagged using the English WN17pre.[12]

- **Arabic and English words are equivalent**

  We observe that a majority of the ambiguous words in Arabic are also ambiguous in English in this test set; they preserve ambiguity in the same manner. In Arabic, 422 word tokens corresponding to 190 word types, are at the closest granularity level with their English correspondent;[13] For instance, all the senses of *care* apply to its Arabic translation *EnAyA* (عنايه); the sense definitions are listed as follows:

  - care, attention, aid, tending: the work of caring for or attending to someone or something; "no medical care was required"; "the old car needed constant attention"
  - caution, precaution, care, forethought: judiciousness in avoiding harm or danger; "he exercised caution in opening the door"; "he handled the vase with care"
  - concern, care, fear: an anxious feeling; "care had aged him"; "they hushed it up out of fear of public reaction"
  - care: a cause for feeling concern; "his major care was the illness of his wife"
  - care, charge, tutelage, guardianship: attention and management implying responsibility for safety; "he is under the care of a physician"
  - care, maintenance, upkeep: activity involved in maintaining something in good working order; "he wrote the manual on car care"

  It is worth noting that the cases where ambiguity is preserved in English and Arabic are all cases where the polysemous word exhibits regular polysemy and/or metonymy. The instances where homonymy is preserved are borrowings from English. Metonymy is more pragmatic than regular polysemy (Cruse, 1986); for example, *tea* in English has the following metonymic sense from WN1.7pre:

  - a reception or party at which tea is served; "we met at the Dean's tea for newcomers"

  This sense of *tea* does not have a correspondent in the Arabic *$Ay* (شاى). Yet, the English *lamb* has the metonymic sense of *MEAT* which exists in Arabic. Researchers building EuroWordNet have been able to devise a number of consistent metonymic relations that hold cross linguistically such as *fabric/material*, *animal/food*, *building/organization* (Vossen et al., 1999; Wim Peters and Wilks, 2001). In general, in Arabic, these defined classes seem to hold, however, the specific case of *tea* and *party* does not exist. In Arabic, the English sense would be expressed as a compound *tea party* or *Hflp $Ay* (حفلة شاى).

- **Arabic word equivalent to specific English sense(s)**

  In this evaluation set, there are 138 instances where the Arabic word is equivalent to a subsense(s) of the corresponding English word. The 138 instances correspond to 87 word types. An example is illustrated by the noun *ceiling* in English.

  - ceiling: the overhead upper surface of a room; "he hated painting the ceiling"
  - ceiling: (meteorology) altitude of the lowest layer of clouds
  - ceiling, cap: an upper limit on what is allowed: "they established a cap for prices"
  - ceiling: maximum altitude at which a plane can fly (under specified conditions)

  The correct sense tag assigned by SALAAM to *ceiling* in English is the first sense, which is correct for the Arabic translation *sqf* (سقف). Yet, the other 3 senses are not correct translations for the Arabic word. For instance, the second sense definition would be translated as {*rtfAE* (إرتفاع) and the last sense definition would be rendered in Arabic as *Elw* (علو). This phenomenon of Arabic words corresponding to specific English senses and not others is particularly dominant

[12] The overlapping number of Arabic words rated ACCURATE by the three annotators of experiment 1 and those accurate items from experiment 1.

[13] This means that all the English senses listed for WN17pre are also senses for the Arabic word.

where the English word is homonymic. By definition, homonymy is when two independent concepts share the same orthographic form, in most cases, by historical accident. Homonymy is typically preserved between languages that share common origins or in cases of cross-linguistic borrowings. Owing to the family distance between English and Arabic, polysemous words in Arabic rarely preserve homonymy.

- **English word equivalent to specific Arabic sense**

  40 instances, corresponding to 20 type words in Arabic, are manually classified as more generic concepts than their English counterparts. For these cases, the Arabic word is more polysemous than the English word. For example, the English noun *experience* possesses three senses in WN17pre as listed below.

  - `experience: the accumulation of knowledge or skill that results from direct participation in events or activities; "a man of experience"; "experience is the best teacher"`
  - `experience: the content of direct observation or participation in an event; "he had a religious experience"; "he recalled the experience vividly"`
  - `experience: an event as apprehended; "a surprising experience"; "that painful experience certainly got our attention"`

  All three senses are appropriate meanings of the equivalent Arabic word *tjrbp* (تجربة) but they do not include the *SCIENTIFIC EXPERIMENT* sense covered by the Arabic word.

From the above points, we find that 63.9% of the ambiguous Arabic word types evaluated are conceptually equivalent to their ambiguous English noun translations. 29.3% of the ambiguous Arabic words correspond to specific senses of their English translations and 6.7% of the Arabic words are more generic than their English correspondents. It is noteworthy that the results of this evaluation are consistent with the results obtained from the SENSEVAL 3 data. This indicates that the obtained results are robust.

8. Conclusions

From both the manual evaluation of the SENSEVAL 3 and SENSEVAL 2 data, we conclude that it is feasible to automatically bootstrap an Arabic WordNet taxonomy given less than perfect translations and alignments leveraging off existing English resources. We find that for Arabic verbs, adjectives and nouns, on average 52.3% of all the words examined, the corresponding English WordNet set of definitions are sufficient as definitions for the Arabic translation word; 39.96% of the Arabic words correspond to specific subsets of the WordNet definitions; and finally, 7.8% of the Arabic words comprise supersets of their corresponding English WordNet translation definitions. These results are very encouraging as they are similar to those obtained by researchers building EuroWordNet. Moreover, SALAAM serves as a very good launching point for such an endeavor. Certainly, the whole process of building an Arabic WordNet taxonomy will not be fully automated, however the approach as presented here significantly reduces the start up time for the creation of such a needed resource for Arabic language processing.

9. Acknowledgements

I would like to thank Mr. Ihab Ragaa from Tarjim Software for facilitating the Arabic MT process. This work is supported, in part, by NSF Award #IIS-0325646.

10. References

Buckwalter, Tim. Buckwalter Arabic Morphological Analyzer Version 1.0., LDC Catalog No.: LDC2002L49. Linguistic Data Consortium, University of Pennsylvania, 2000.

Cruse, D., 1986. *Lexical Semantics*. Cambridge University Press.

Darwish, Kareem, 2002. Building a shallow arabic morphological analyzer in one day. In *Proceedings of ACL Workshop on Semitic languages*. Pennsylvania, USA.

Diab, Mona, 2000. *Exploiting Translations for Semantic Annotation*. Ph.D. thesis, University of Maryland.

Diab, Mona, 2003. Word sense disambiguation within a multilingual framework. In *PhD Thesis*. University of Maryland, College Park.

Diab, Mona, 2004a. Relieving the data acquisition bottleneck for wsd. In *Proceedings of 42th ACL Conference*. Barcelona, Spain.

Diab, Mona, 2004b. Unsupervised approach to botstrapping arabic word sense tagging. In *Arabic Script based language Processing Workshop, Coling 2004*. Geneva, Switzerland.

Diab, Mona and Philip Resnik, 2002. Word sense tagging using parallel corpora. In *Proceedings of 40th ACL Conference*. Pennsylvania, USA.

Fellbaum, Christiane, 1998. *WordNet: An Electronic Lexical Database*. MIT Press. Http://www.cogsci.princeton.edu/~wn [2000, September 7].

Mona Diab, Daniel Jurafsky, Kadri Hacioglu, 2004. Automatic tagging of arabic text: From raw text to base phrase chunks. In *NAACL Conference*. Boston, USA.

Quillian, M.R., 1968. Semantic Memory. In M. Minsky (ed.), *Semantic Information Processing*. Cambridge, MA: The MIT Press.

Resnik, Philip, 1999. Disambiguating Noun Groupings with Respect to WordNet Senses. In S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann, and D. Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora*. Dordrecht: Kluwer Academic, pages 77–98.

Vossen, P., W. Peters, and J. Gonzalo, 1999. Towards a Universal Index of Meaning.

Wim Peters, Louise Guthrie and Yorick Wilks, 2001. Cross-linguistic discovery of semantic regularity.