Automatically Extending Named Entities coverage of Arabic WordNet using Wikipedia

Musa Alkhalifa and Horacio Rodríguez

***Abstract*— This paper focuses on the automatic extraction of Arabic Named Entities (NEs) from the Arabic Wikipedia, their automatic attachment to Arabic WordNet, and their automatic link to Princeton's English WordNet. We briefly report on the current status of Arabic WordNet, focusing on its rather limited NE coverage. Our proposal of automatic extension is then presented, applied, and evaluated.**

***Index Terms*—Arabic NLP, Arabic WordNet, Named Entities Extraction, Wikipedia.**

I. Introduction

Ontologies have become recently a core resource for many knowledge-based applications such as knowledge management, natural language processing, e-commerce, intelligent information integration, database design and integration, information retrieval, bio-informatics, etc. The Semantic Web, [7], is a prominent example on the extensive use of ontologies. The main goal of the Semantic Web is the semantic annotation of the data on the Web with the use of ontologies in order to have machine-readable and machine-understandable Web that will enable computers, autonomous software agents, and humans to work and cooperate better through sharing knowledge and resources.

In the area of Natural Language Processing, by far, the most widely used lexico-conceptual ontology is Princeton's English WordNet, [17]. Princeton's WordNet has become a de facto standard repository of lexical semantic information. The coverage of English WordNet, as shown in Table 1, is really impressive in terms of number of synsets, words, and relations.

Due to the success of Princeton's English WordNet, a lot of efforts have been devoted for building wordnets for other languages. Although most of these wordnets have been built manually, in some cases a substantial part of the work has been performed automatically using English WordNet as source ontology and bilingual resources for proposing alignments.

Okumura and Hovy [30] proposed a set of heuristics for associating Japanese words to English WordNet by means of an intermediate bilingual dictionary and taking advantage of the usual genus/differentiae structure of dictionary definitions.

Version	nouns	verbs	adj	adv	total synsets	relations
WN1.5	51,253	8,847	13,460	3,145	76,705	103,445
WN1.6	66,025	12,127	17,915	3,575	99,642	138,741
WN1.7	74,488	12,754	18,523	3,612	109,377	151,546
XWN	74,488	12,754	18,523	3,612	109,377	551,551
WN1.7.1	75,804	13,214	18,576	3,629	111,223	153,781
WN2.0	79,689	13,508	18,563	3,664	115,424	204,074
WN2.1	81,426	13,650	18,877	3,644	117,597	232,916
WN3.0	82,115	13,767	18,156	3,621	117,659	235,402

Table 1. Content of different versions of Princeton's English WordNet

Later, Khan and Hovy [20] proposed a way to reduce the hyper-production of spurious links by this method by searching common hyperonyms that could collapse several hyponyms.

The first attempt following Princeton WordNet's approach towards the construction of WordNets on a large scale was the development of EuroWordNet project [40]. Within the framework of this project Spanish WordNet, [34], were developed through a collective effort of three Spanish universities (UNED, UB and UPC). Although the different partners of EuroWordNet followed slightly different approaches for building their wordnets (according to their available lexical resources) a common approach of manual building of an initial set of Base Concepts, and a further top-down extension of this set was followed by all the partners in the first phase. In the second phase of the construction, complementary resources such as bilingual dictionaries were used.

Later on, Catalan [6] and Basque [1] WordNets were developed following the same approach.

Musa Alkhalifa is doctoral student at Universitat Pompeu Fabra (UPF), Barcelona, Spain, (e-mail: musa.alkhalifa01@campus.upf.edu).

Horacio Rodríguez is professor at Universitat Politècnica de Catalunya (UPC), Barcelona, Spain (e-mail: horacio@lsi.upc.edu).

:

A similar methodology was applied to building the Hungarian WordNet [27]. In this case, the basic bilingual-based approach was complemented with methods using a monolingual explanatory dictionary. Also Chen [10] complemented the bilingual resources with information extracted from a monolingual Chinese dictionary for building both a Chinese and a Chinese-English wordnets.

Arabic synsets	11270
Arabic words	23496

pos	DB content
a	661
n	7961
r	110
v	2538

Named entities:

Synsets that are named entities	1142
Synsets that are not named entities	10028
Words in synsets that are named entities	1656

Figure 1. Figures of Arabic WordNet database at the end of the project (February 2008)

In [4], Barbu and Barbu-Mititelu followed a similar approach for building the Romanian WordNet, but using additional knowledge sources as Magnini's WordNet domains codes [23] and WordNet glosses. They used a set of meta-rules for combining the results of the individual heuristics for achieving a 91% accuracy for a coverage of 9,610 synsets.

Another important project concerned with building wordnets was the BalkaNet project [39]. The Common Base Concepts of the resulting resource have been used in building the Arabic WordNet, as reported in section II.

Arabic WordNet ([8], [14], [35], [36]) has been built along the last years following the EuroWordNet methodology of manually encoding a set of base concepts while maximizing compatibility across wordnets (Arabic and English in this case). As a result, there is a straightforward mapping from Arabic WordNet onto Princeton WordNet 2.0 (Princeton's WordNet – [17]). In addition, the Arabic WordNet project aimed at providing a formal specification of the senses of its synsets using the Suggested Upper Merged Ontology (SUMO – [29]). This representation serves as an interlingua among all wordnets ([31], [40]) and will underlie the development of semantics-based computational tools for multilingual Natural Language Processing.

Arabic WordNet was a two years project. It was funded by the US government under the REFLEX program. The project was directed by Christiane Fellbaum, from Princeton University (USA), and the rest of partners were two universities, Manchester University (UK) and UPC (Spain) and two companies, Irion Technologies (the Netherlands) and Articulate Software (USA).

In Accordance with the objectives of the project, Arabic WordNet currently, i.e. at the end of the project[1], consists of 11,270 synsets (7,961 nominals, 2,538 verbals, 661 adjectivals, and 110 adverbials), containing 23,496 Arabic expressions (words and multiwords). This number includes 1,142 synsets that are Named Entities which have been extracted automatically and checked by the lexicographers. For the most up-to-date statistics on the content of Arabic WordNet see:

http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php.

Figure 1 shows the current figures of Arabic WordNet as presented in this Web page.

In accordance with the conditions set by Arabic WordNet project contractors, all the content of Arabic WordNet database was manually built or at least, as in the case of Named Entities, manually revised. In this later case, the coverage is rather limited and meant to be considered just as a sample of the capabilities of the resource. Our current, more important, goal which is presented in this paper is to devise a way to automatically enrich the current set of Named Entities in the database using high quality sources such as the Arabic Wikipedia.

The organization of this paper after this introduction is as follows: Section II describes briefly the methodology used in the construction of Arabic WordNet. Section III is devoted to the approaches followed for the semi-automatic extension of Arabic Wordnet. Section IV reviews the way we followed for

English form	Arabic form
(he) studied	دَرَسَ
(I) studied	دَرَسْتُ
(I) study	أَدْرُسُ
(he) studies	يَدْرُسُ
(we) study	نَدْرُسُ
...	...

Table 2. Some inflected verbal forms (of 82 possibilities) for درس (DaRaSa,, to study)

[1] Several attempts for extending Arabic WordNet are currently in progress.**********

collecting the Named Entities currently included in Arabic WordNet. Section V discuses the potential use of the Wikipedia as source for enriching the set of Named Entities. Section VI outlines our approach. In section VII a detailed example illustrating this approach is presented. Results and evaluation are discussed in section VIII. Finally, in section IX, our conclusions and further work are presented.

II BUILDING AN ARABIC WORDNET

Following EuroWordNet methodology, Arabic WordNet[2] was developed in two phases: first, building a core WordNet around the most important concepts, the so-called Base Concepts, and secondly extending this core WordNet downward to more specific concepts using certain criteria. The core WordNet was designed to be highly compatible with WordNets in other languages that have been developed according to the same approach.

For the core WordNet, the Common Base Concepts of the 12 languages in EuroWordNet (1,024 synsets) and BalkaNet (8,516 synsets) that are translatable into Arabic were encoded as synsets in Arabic WordNet. Other Arabic language-specific concepts were added and translated and manually linked to the closest synsets. The same procedure was performed on all English synsets having an equivalence relation in the SUMO ontology. Synset encoding proceeded bi-directionally: given an English synset, all corresponding Arabic variants (if any) were selected; given an Arabic word, all its senses were determined and for each of them the corresponding English synset was encoded.

For the sake of coherence and connectivity with English WordNet the set of Arabic synsets was extended with hypernym relations to form a closed semantic hierarchy. Also, in this phase, wherever possible lexical gaps in the hypernymy hierarchy were filled. SUMO was used in this phase to maximize the semantic consistency of the hyponymy links. The result represents the core WordNet, which was the semantic basis for further extension. All the work was done manually with the help of lexicographic interfaces and sets of Arabic wording suggested for each English synset. Arabic lexicographers decided on either accepting, rejecting, extending or modifying the proposed mappings. We proceeded in this way for both nouns and verbs (adjectives and adverbs were added opportunistically when derived from a verb). When a new Arabic verb was added, extensions were suggested from verbal entries, including verbal derived forms, nominalizations, verbal nouns, active and passive participles and so on.

In a second phase the database was extended from the Arabic core WordNet. We proceeded downwards adding layers of hyponyms chosen according to certain criteria: maximal connectivity, relevance, and generality.

At a final step, a set of terminological data corresponding to pre-defined domains[3] were added to the database, filling gaps when needed. See [35] for a more in depth description of the procedure for selecting these synsets.

The database structure comprises four principal entity types: item, word, form and link. Items are conceptual entities, including synsets, ontology classes, and instances. An item has a unique identifier and descriptive information such as a gloss. Items lexicalized in different languages are distinct. A word entity is a word sense, where the word's citation form is associated with an item via its identifier. A form is an entity that contains lexical information (not merely inflectional variation). The basic content of the forms are the root forms of the words but additional data (such as the irregular/broken plural form), where applicable, can be represented in this way.

Encoding root information is an important issue in Arabic WordNet. The root groups together a set of semantically related forms. For instance, the verbal basic form دَرَسَ (DaRaSa, to study/to learn) has a a root reduced to درس (DRS), from this root, lexical rules can produce derived verbal forms as دَرَّسَ (DaRRaSa, to teach), among others. From any verbal form (whether basic or derived), both nominal and adjectival forms can also be generated in a highly systematic way: the nominal verb (masdar) as well as masculine and feminine active and passive participles. Examples include the masdar دَرْسٌ (DaRSun, lesson, study), مُدَرِّسٌ (MuDaRRiSun, male teacher), or مُدَرِّسَةٌ (MuDaRRiSatun, female teacher). Note that all these forms owning the same root are semantically related, sometimes in a predictable way. Having access to this information in Arabic WordNet opens interesting possibilities in several Natural Language Processing,tasks.

A link relates two items, and has a type such as "equivalence," "subsuming," etc. Links interconnect sense Items , e.g., an English synset to an Arabic synset, a synset to a SUMO concept, etc. This data model was specified in XML as an interchange format, and was implemented in a MySQL database.

Following this approach Arabic WordNet was built and reached the overall coverage shown in Figure 1.

[2] To our knowledge the only previous attempt to build a WordNet for the Arabic language consisted of a set of experiments carried out by Mona Diab [12] for attaching Arabic words to English synsets using only English WordNet and a parallel Arabic English corpus as knowledge source.

[3] A set of domains to be covered was defined in the contract. These domains were manually mapped into Magnini's domains codes, [23]: atomic_physic, biology, economy, chemistry, commerce, doctrines, military, politics, drugs and dangerous things.

III SEMI-AUTOMATIC EXTENSIONS OF ARABIC WORDNET

Although the construction of Arabic WordNet was performed manually (in accordance with the terms of the contract), some efforts have been made to automate part of the process of extension using available bilingual (Arabic/English) and monolingual (Arabic) lexical resources. Using lexical resources for the semi-automatic building of wordnets for languages other than English is not new, as was discussed above.

For obtaining generic synsets we have investigated two general approaches which take advantage of an important characteristic of Arabic (and other Semitic languages), which is that words sharing a common root (i.e. a sequence of almost always three consonants) usually have related meanings and can be derived from a common base of verbal form by means of a reduced and very accurate set of lexical rules. Besides these two general approaches for obtaining generic synsets, two other lines of research for extending Arabic WordNet have been followed for obtaining i) domain restricted terminological synsets and ii) Named Entities. The first line is not approached in this paper, the latter is discussed in sections IV and VI.

Both approaches aim at deriving new Arabic word forms from existing Arabic verbal synsets and then producing a list of suggested English synsets for each form. The first approach, described in [35], is based on a heuristic guided application of the set of lexical rules. The second approach, described in [36], formalizes the decisions in a Bayesian framework. Both approaches can be (and have been) combined for getting more accurate results.

The central problems to be faced are, on the one hand, filtering noise caused by overgeneration of Arabic word forms (obviously, the application of the whole set of lexical rules to a given form results in a severe overgeneration of Arabic forms, for instance, for درس, out of the nine possible derived form generated by the application of the first rule set, only the six shown in Table 3 are valid according to [41]) and, on the other, mapping the newly created forms to appropriate English WordNet synsets.

To deal with the filtering problem, we implemented a set of Decision Tree classifiers using the C5.0 implementation in the Weka toolbox, [42]. Details are reported in [35].

Regarding the second problem, i.e. associating these Arabic words with Princeton's English WordNet synsets, we translated the Arabic words (layer 1, A_i in Figure 2) into English (layer 2, E_i) and identified all the synsets these translations belonged to (layer 3, S_i), thus producing a set of <Arabic word, English word, Princeton's English WordNet synset> tuples. Furthermore, we looked for semantic relations holding in Princeton's English WordNet that involve the synsets in layer 3, and this led to a new layer (layer 4, S_i). In this way an undirected graph was built.

Consider once more the example of دَرَسَ (DaRaSa, to study/to learn) presented above. From this verbal basic form, the root درس (DRS) can easily be extracted. In our case, as we try to extend Arabic WordNet semi-automatically starting in the already existing verbal entries, the root form exists in the database, thus the extraction of the root is quite straightforward. Anyway, extracting the root from the basic verbal form in a general case, when the verbal form does not occur in Arabic WordNet is not difficult (obviously extracting the root from whatever form is more challenging). Several root extractors are freely available. An example is *gendic*[4]. Interesting systems are [2], [42] and [11].

Once the root is extracted, sets of lexical and morphological rules can be used for extracting related forms. Table 2 presents some examples of the inflected verbal forms corresponding to the basic form, دَرَسَ (DaRaSa, to study/to learn) and to its corresponding root form, درس (DRS). The set of lexical rules was automatically built using as Knowledge Source the LOGOS database of Arabic verbs which contains 944 fully conjugated Arabic verbs[5].

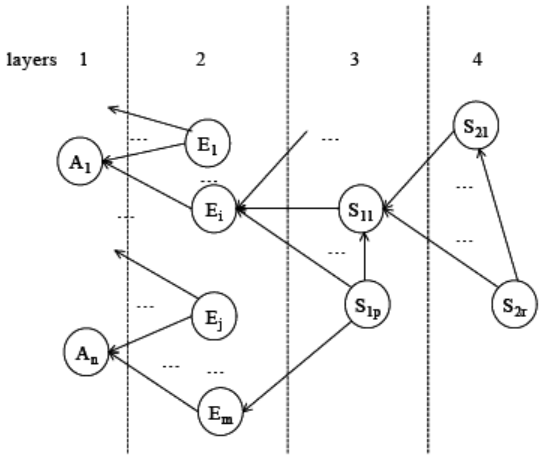

Figure 2. Example of Graph of associations

The number of different forms depends on the class of the verb (basic class and up to 10 derived classes) but it ranges from 44 to 84 different forms. Class 1, the basic class, has 82 forms, some of which are presented in Table 2, and, thus,

[4] http://www.freshmeat.net
[5] http://www.logosconjugator.org/verbi_utf8/all_verbs_index_ar.html

requires the application of 82 different morphological rules for generating them. Table 3 presents the valid derived verbal forms corresponding to the same basic form. Note that not all the possible derived forms for a basic one correspond to valid forms. Sets of lexical rules for deriving both the inflected verbal forms of the root and the derived verbal forms can be easily written. Combining the two rule sets would result in the generation of all the valid inflected forms from both the basic verbal form and all its valid derived forms. We have used for this purpose the Xerox Finite State software, [5] with no major problems.

From any verbal form (whether basic or derived by the corresponding rule set, both nominal and adjectival forms can also be generated in a highly systematic way: the nominal verb (masdar) as well as masculine and feminine active and passive participles. Examples generated from from درس (DaRaSa, to study/to learn) include the masdar form درس (DaRSun, lesson, study) and مدرّس (MuDaRRiSun, male teacher) in this case coming from درّس (DaRRaSa, to teach), second class derivative of the original basic form.

Beyond this, we aimed to extend this basic approach to the derivation of additional forms including the feminine form from any nominal masculine form (for instance, مدرّسة, MuDaRRiSatun, female teacher, from مدرّس, MuDaRRiSun,

Class	English form	Arabic form
1 (basic)	to learn, to study	د ر س
2	to teach	د رّ س
3	to study (together with someone)	د ا ر س
4	to learn with	ا د ر س
6	to study (carefully together)	تـد ا رَس
7	to vanish	انـد ر س

Table 3. Valid derived forms from درس (DaRaSa, to study)

male teacher), or the regular plural forms from any nominal singular form. For instance, the regular nominative plural form is created by adding the suffix (Una) to the singular form (e.g., مدرّسون MuDaRRiSUna, male teachers, is derived from مدرّس, MuDaRRiSun, male teacher).

As a result of this process we have built for each of the 2,538 verbal entries of Arabic WordNet a graph like the one presented in Figure 2. As said above, we have followed two ways of using this structure for proposing new <Arabic word/ English synset> associations, based, respectively, on a set of heuristics and a Bayesian model. We will now briefly describe the two approaches.

Both approaches start by building the set of association graphs described above but differ in the way of scoring the reliability of these candidates. Our scoring routine is based on the observation that in most cases the set of derivative forms have semantically related senses (because they own the same root). For instance, درس (DaRaSa, to study) belongs to Class 1 and its masdar is درس (DaRSun, lesson). درّس (DaRRaSa, to teach) belongs to Class 2 and its masculine active participle is مدرّس (MuDaRRiSun, male teacher). All these words have the same root (درس). Clearly these four words are semantically related. Therefore, if we map Arabic words to English translations and then to the corresponding English synsets, we can expect that the correct assignments will correspond to the most semantically related synsets. In other words, the most likely <Arabic word, English synset> associations are those corresponding to the most semantically related items.

Using the graph as input (see Figure 2), the first approach to compute the reliability of association between an Arabic word and an English synset consists of simply applying a set of five graph traversal heuristics. The heuristics are as follows (note that in what follows, A_i refers to an Arabic word forms, E_i to an English word form, and S_i to an English synset):

1. 1. If a unique path A-E-S exists (i.e., A is only translated as E), and E is monosemous (i.e., it is associated with a single synset), then the output tuple <A, S> is assigned a score value of 1.

2. If multiple paths A-E_1-S and A-E_2-S exist (i.e., A is translated as E_1 and E_2 and both E_1 and E_2 are associated with S among other possible associations) then the output tuple <A, S> is assigned a score value of 2.

3. If S in A-E-S has a semantic relation to one or more synsets, S_1, S_2 … that have already been associated with an Arabic word on the basis of either heuristic 1 or heuristic 2, then the output tuple <A, S> is assigned a score value of 3.

4. If S in A-E-S has some semantic relation with S_1, S_2 … where S_1, S_2 … belong to the set of synsets that have already been associated with related Arabic words, then the output tuple <A-S> is assigned a score value of 4. In this case there is only one translation E of A but more than one synset associated with E. This heuristic can be sub-classified by the number of input edges or supporting semantic relations (i.e. 4.1, 4.2, 4.3, ...).

5. This heuristic is similar to 4 except that there are multiple translations E_1, E_2, … of A and, for each translation E_i, there are possibly multiple associated synsets S_{i1}, S_{i2}, ... In this case the output tuple <A-S> is assigned a score value of 5 and again the heuristic can be sub-classified by the number of input edges or supporting semantic relations (5.1, 5.2, 5.3 ...).

Applying all the heuristics resulted in a precision score of 0.5 for a recall of 0.61. Limiting the application to heuristics 1 and 2, the recall falls to 0.18 while the precision raised to 0.65. With few exceptions the expected trend for the reliability scores are as expected (heuristics 2 and 3 perform better than heuristic 4 and the latter better than heuristic 5). It is also worth noting that heuristic 3, the first that relies on semantic relations between synsets in English WordNet, outperforms heuristic 2.

The second approach starts building a Bayesian Network from the association graph. This implies:

1. Assigning direction to edges in order to transform the undirected graph into a directed one. We followed a greedy approach to avoid cycles when inserting S nodes. S nodes were sorted by number of output edges and edges are added once at a time if no cycle is produced.

2. Computing the Conditional Probability Table, CPT, for each node in the net. Being binary all the variables, the CPT size of a node *i* is in our case, 2^n, for *n* = number of fathers of *i*. We have used a threshold (set to 10) on the maximum number of fathers for a node. The same approach used for avoiding cycles was also used for deciding which nodes will be selected as fathers. Computing the CPT depends on the type of edge. For edges EW -> AW we used probabilities coming from Statistical Translation Models, built from UN Arabic/English bilingual corpus[6] using GIZA++ (word-word probabilities) for estimating conditional priors. For edges ES -> EW we have, simply, uniformly distribute the probability mass between the variants of the synset. For edges ES -> ES the process is more complex. See [36] for details.

For each built Bayesian network, a Bayesian inference has been performed setting as evidences the nodes in AW layer and looking for the probabilities of all the synsets in S1. The set of candidates is built with tuples <X,Y> where X belongs to AW, Y belongs to S1 having a non null probability, when there is a path from X to Y. The tuple is scored with the posterior probability of Y given the evidences provided by the net.

The results of this second approach using different thresholds rank from a precision of 0.4 for a recall 1.0 until a precision of 0.6 for a recall of 0.28.

Intersecting both methods results on a clear improvement. The best recall (0.71) produced a precision of 0.59. The best precision (0.71) was obtained with a quite restrictive threshold. Although the recall in this case is low (0.38), the average number of words candidates to AWN is really high (92 words for base form in average).

IV Collecting Named Entities in Arabic WordNet project

The process of collecting Named Entities for being included in Arabic Wordnet followed, too, a semi-automatic approach that allows us to use it as a base for the automatic approach presented in this paper. The process consisted of two steps:

1. Selection of the candidates.
2. Manual validation. As for all the content of Arabic WordNet a manual revision of the set of synsets is needed.

According to the conditions of our contract, at least 1,000 Named Entities synsets should be built, covering a variety of types (locations, persons, organizations, etc.) that should be, whenever possible, linked to existing instances in Princeton's English WordNet.

A. Selecting candidates

Our goal in this step was constraining as much as possible the set of candidates in order to reduce the human effort in the second step. We started with the information contained in three resources:

1. The GEONAMES[7] database for toponym information corresponding to Arabic countries. GEOnet Names Server is a worldwide database of geographic feature names, excluding the United States and Antarctica, with 5.3 million entries. Each gazetteer entry contains a geographical name (toponym) and its geographical coordinates (latitude, longitude), language of the geographical name and other geographical features as country name, capital, main cities,, first administrative division, organizative districts, etc.) as well as non geographical such as the current head of state, the head of govern, the currency, and other. Only information involving Named Entities has been extracted in our case. See Figure 3 for some examples of the information extracted.

[6] UN (2000-2002) bilingual Arabic-English Corpus (available through LDC: catalog # LDC2004E13).

[7] http://www.geonames.org/

2. A gazetteer of Countries in the world from FAO[8]. This gazetteer contains the name of all the countries around the world in different languages indexed by ISO code. We have used the Arabic and English files.

3. The Named Entity entries contained in the NMSU (New Mexico State University) bilingual Arabic-English lexicon[9]. The candidates from these resources have, however, a non null intersection and some inconsistencies occur that need to be solved manually in the second step.

In the case of the GEONAMES and FAO databases, the procedure was quite straightforward. For GEONAMES we started by selecting the pages corresponding to Arabic countries (see an example in Figure 3), then we wrote wrappers, i.e. web page specific scripts, for extracting information from these web pages and formatting results. For

Morocco / اَلْمَغْرِب

ma - cas - rba - subdivisions

official name in English:	native name:
Kingdom of Morocco	المملكة المغربية (al-Mamlakâtu l-Ma?ribiyyâ)
adjective:	native adjective:
Moroccan	مغربي (ma?ribī)
capital:	native name:
Rabat	الرباط (ar-Ribā?)
official language:	native name:
Arabic + Tamazight	العربية (al-? arabiyyâ) + tmazi? t / ? ? ? ? ? ? ? / تمازيغت
currency:	native name:
1 dirham = 100 centimes	? درهم = ??? سنتيم (1 dirham = 100 santīm)
head of state / government:	native name:
King Mohammed VI Prime Minister Idriss Jettou	الملك محمد السادس (al-Malik Mu? ammad as-sādis) الوزير الأول إدريس جطو (al-Wazīru l-Awwal Idrīs ? a? ? ū)

Figure. 3. A fragment of GEONAMES database

FAO, we simply aligned English and Arabic Named Entities by means of the ISO code (see Figure 4).

The case of NMSU was more complex. The database had a larger coverage but the entries had no diacritics at all, including not only vowels but other marks as the "shadda" diacritic, and obviously, not only Named Entities but also normal entries are included in the dictionary. Although most Arabic texts are unvowelized, i.e. do not contain diacritics, a design decision when building Arabic WordNet was that all the entries should be fully vowelized, including Named Entities[10].

The case of shadda is specially problematic, shadda in Arabic marks a consonant duplication (a gemination) and the meaning of a word with or without the mark can be absolutely different. Consider, for instance, the Arabic word درس (DaRaSa) without shadda and درّس (DaRRaSa,) with shadda . Both entries appear as درس in the NMSU lexicon but the meaning is to study/to learn in the former and to teach in the later.

We proceed in the following steps:

1. Identifying synsets corresponding to instances in English WordNet. A known problem in WordNet is the lack of distinction between synsets corresponding to classes (e.g. country) and those corresponding to instances (e.g. Morocco). From Enrique Alfonseca's page[11] a list of PWN1.7 synsets corresponding to instances can be downloaded. These synsets were then mapped from PWN1.7 to PWN2.0 using TALP mappings[12] between different versions of Princeton's English WordNet. The mapping resulted on very small loss in accuracy.

2. Obtaining the generic types, i.e. the Princeton's WordNet synsets corresponding to the direct hyperonyms of the instance synsets. This resulted in obtaining 371 generic types from which only synsets already linked to Arabic Wordnet were collected (such as 'capitals', 'cities', 'countries', 'inhabitants' or 'politicians'). In some cases, when the generic synset was highly productive, the Arabic counterpart, if not already present in Arabic WordNet was manually added to the database.

3. Proceeding downwards for getting all the instance synsets corresponding to the hypernyms of the generic types. In most cases these sysnsets correspond to those recovered in step 1. But new ones appear.

4. Obtaining NMSU entries corresponding to the variants in the instance synsets obtained in step 3. Only nominal entries were recovered. For example, for instances of hyponyms of the generic

[8] http://www.fao.org/faoterm/

[9] http://crl.nmsu.edu/Resources/dictionaries/download.php?lang=Arabic

[10] The decision is controversial because there is no common agreement in different Arabic countries on the way of vowelizing Named Entities. In case of doubt we have used the most frequent vowelization according to our lexicographers.

[11] http://alfonseca.org/pubs/ind-conc.tgz

[12] http://www.lsi.upc.edu/~nlp/

entry 'politicians' 129 synsets were found.

Finally we unified the formats produced by the three sources and merged the results.

B. Manual validation

This step iterates on the associations proposed by the first step. For each candidate the following tasks were performed:

ISO	COUNTRY NAME	INFO
AW	أروبا	--
AZ	أذربيجان	
AM	أرمينيا	
AU	أستراليا	
AF	أفغانستان	
AL	ألبانيا	
DE	ألمانيا	
AG	أنتيغوا وبربودا	

Fig. 4. A fragment of FAO database

- Deciding the acceptance or rejection of the pair <English synset/Arabic form>.
- Modifying Arabic form if needed.
- Adding diacritics in the case the proposed Named Entity was unvowelized.
- Adding additional variants if available to the new created Arabic synset.
- Completing attachments to English WordNet if possible.

arabic	number_of_instances	english
إله	18	deity, divinity, god, immortal
عَاصِمَة	16	capital
بَلَد	100	country, state, land
دَوْلَة	17	state, nation, country, land, commonwealth, res_publica, body_politic
جَزِيرَة	12	island
مَدِينَة	458	city, metropolis, urban_center
مُقَاطَعَة	321	district, territory, territorial_dominion, dominion
نَهْر	10	river
سَاكِن	20	inhabitant, dweller, denizen, indweller

Fig. 5. Distribution of Arabic Wordnet Named Entity coverage by generic type (most frequent types).

The whole procedure resulted in obtaining 1,147 synsets that have in total 1,659 variants corresponding to 31 generic types.

Figure 5 presents the number of instances of the most frequent types. See:

http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php for more details.

V WIKIPEDIA AS SOURCE OF LEXICAL RESOURCES

Wikipedia[13] , is by far the largest encyclopedia in existence with more than 3 million articles in its English version (English Wikipedia) contributed by thousands of volunteers. Wikipedia experiments an exponential growing in size (number of articles, number of links, etc). There are versions of Wikipedia in more than 200 languages although the coverage (number of articles and average size of each article) is very irregular.

The Arabic version (Arabic Wikipedia) has over 65,000 articles[14] (about 1% of the total size of Wikipedia). Among all the different languages, Arabic has a rank of 29, just above Serbian and Slovenian. The growing of Arabic Wikipedia is, however, very high (more than 100% in last year) so it seems that in a relatively short time the size of Arabic Wikipedia could correlate with the importance (of the number of speakers) of Arabic language.

Wikipedia basic information unit is the "Article" (or "Page"). Articles are linked to other articles in the same language by means of "Article links". There are about 15 output article links (links are not bidirectional) in average in each Wikipedia article. The set of articles and their links in Wikipedia form a graph. Wikipedia articles can be assigned to Wikipedia categories (through "Category links") that are also organized as a graph (see [44] for an interesting analysis of both graphs). Besides article and category links.

Wikipedia pages can contain "External links", that point to

[13] http://www.wikipedia.org/

[14] The figures about Wikipedia coverage are taken from the version we used in the experiments reported in this paper. We downloaded the version of Arabic Wikipedia corresponding to February 2008. Currently Arabic Wikipedia has 110,000 articles. The comparison of these figures gives insights of the growing rate of the resource.

external URLs, and "Interwiki links", from an article to a presumably equivalent, article in another language. There are in Wikipedia several types of special pages relevant to our work: "Redirect pages", i.e. short pages which often provide equivalent names for an entity, and "Disambiguation pages", i.e. pages with little content that links to multiple similarly named articles.

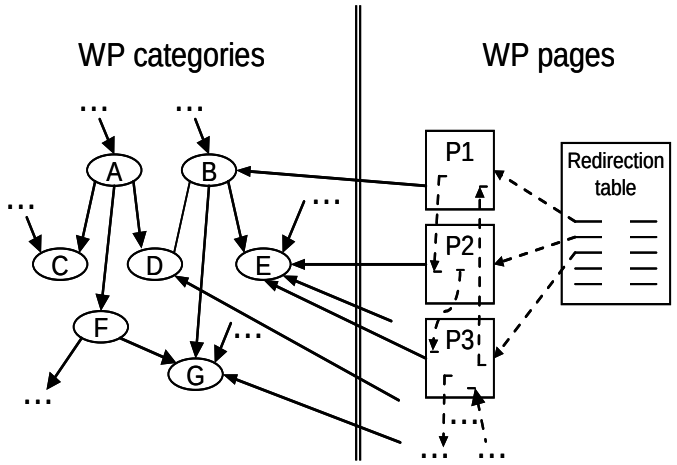

Fig. 6. The graph structure of Wikipedia

Figure 6 shows the graph structure of Wikipedia. The two subgraphs of pages and categories are shown at the right and left part of the figure. Categories can be seen as classes that are linked to pages (pages belonging to the category) and to other classes (super and sub categories). Also each page or category) has one or more categories assigned. While edges between categories usually have a clear semantics (hypernymy and hyponymy relationships), edges between pages lack tags or semantics. Some of the categories of Wikipedia are defined by WP managers for internal organization (eg. “Wikipedia stubs”, “Wikipedia cleanup”, etc.). Also some of the Wikipedia pages are built for organizational purposes as most of the list pages (e.g. "Authors by year", "Cities by country", and so).

Wikipedia has been extensively used for extracting lexical and conceptual information. [32], and [37] build or enrich ontologies from Wikipedia, [28] derive domain specific thesauri, [3] produce a semantically annotated snapshot of English Wikipedia, [24], [26], and [43] perform semantic tagging or topic indexing with Wikipedia articles. Closer to our approach are the works of [38] and [21] where they used Wikipedia, particularly the first sentence of each article, to create lists of named entities. Relatively low effort has been devoted to exploit the multilingual information of Wikipedia. [18], [33] and more recently [16] are notable exceptions.

Extracting information from Wikipedia can be done easily using a Web crawler and a simple html parser. The regular and highly structured format of Wikipedia pages allows this simple procedure. There are, however, a lot of APIs providing easy access to Wikipedia online or to the database organized data obtained from Wikipedia dumps[15]. Some interesting systems are Waikato's WikipediaMiner toolkit[16], U. Alicante's wiki db access[17], Strube and Ponzetto's set of tools[18], Iryna Gurevych' JWPL[19], etc.

In [25] there is an excellent survey of current Wikipedia issues and applications.

VI Our approach

Our purpose is getting Named Entity candidates to be attached as instances to current synsets of the Arabic Wikipedia. In the research reported in this paper we restrict ourselves to Named Entities that have English counterparts in the English WordNet. Other Named Entities for which

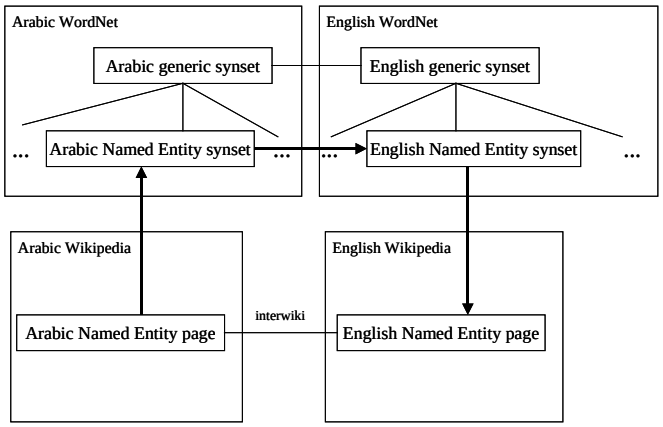

Fig. 7. Named Entities in WordNet and Wikipedia

interwiki links between Arabic and English Wikipedias exist can, however, be extracted following the same approach and attached as direct hyponyms of the corresponding generic synsets but will lack correspondence to the English Named Entities. Note that the coverage of Named Entities in English Wikipedia is at least one order of magnitude greater than the coverage of Named Entities in English WordNet.

Consider Figure 7 and the corresponding example in Figure 8. A pair of generic synsets, in this case, {city, metropolis, urban_center} in English WordNet and "مدينة" in Arabic

[15] http://en.wikipedia.org/wiki/Wikipedia_database
[16] http://wikipedia-miner.sourceforge.net/
[17] http://www.dlsi.ua.es/□atoral/
[18] http://www.eml-research.de/english/research/nlp/download/
[19] http://www.ukp.tu-darmstadt.de/software/jwpl/

WordNet are linked by an equivalence link. In English WordNet several instances of the generic synset are related to it by a direct hyponymy link. In the example, such instances are cities. One of these instances, in figure 8, is "Barcelona". Some of the instances of the generic synset exist as entries in the English Wikipedia. This is the case of "Barcelona". In this example the entries in WordNet and Wikipedia have the same name, but this is not always the case. In some cases the English Wikipedia page has an interwiki link with the Arabic Wikipedia, this is the case of "Barcelona" that is linked to "برشلونة ". A link between Arabic Wikipedia and Arabic WordNet is set for completing the loop.

At first glance, given an English Named Entity, obtaining the Arabic counterpart using Wikipedia seems to be easy: We can recover the page corresponding to the English Named Entity. If the page exists, we can look for an occurrence of an "interwiki link" to an Arabic page and just return the title of the page. Unfortunately things are not so easy. Several problems must be faced:

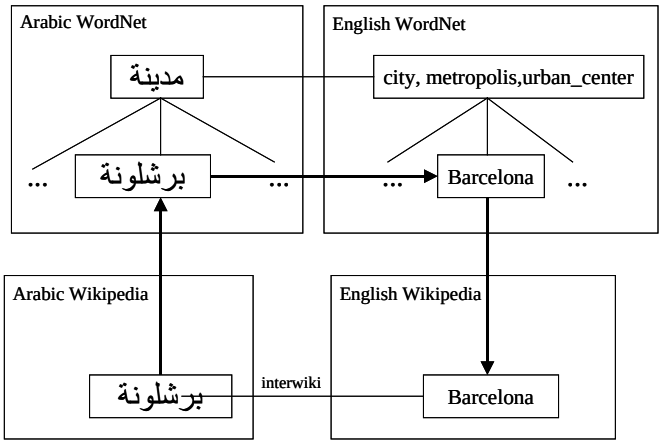

Fig. 8. Named Entities in WordNet and Wikipedia (instances)

- Which English Named Entities have to be looked for?
  - We can consider all the English Wikipedia pages but, in this case, i) we are introducing a lot of noise in the case of pages not corresponding to a Named Entity (compared to WordNet, Wikipedia contains many more Named Entities as article titles, i.e. as entries,. However, about 30% of Wikipedia content corresponds to generic, not named, entries), and ii) the Wikipedia pages have to be mapped to Princeton's WordNet synsets and thus a possible Word Sense Disambiguation, problem arises. For instance, looking at Wikipedia for "Picasso" results on a page corresponding to the painter, but also other pages are accessed, a couple of museums, other persons and some buildings. So, the correct page has to be selected. In the framework of Wikipedia, the Word Sense Disambiguation problem can be solved, or at least alleviated, using the information of Disambiguation Pages but this is not the case of WordNet.
  - We can start not from Wikipedia but from Princeton's WordNet. In this case we have to locate in Princeton's WordNet the set of initial instances (using the same procedure described in section IV) and we have to face the same problem of Word Sense Disambiguation in this case not against Princeton's WordNet synsets but against the English Wikipedia pages.

From the two possibilities we have chosen this latter approach. The reason is that we are interested not in extracting Named Entities from Wikipedia in general but in enriching the current Arabic WordNet with Named Entities that are attachable to existing Named Entities in English WordNet.

- How to deal with polysemy, i.e. when multiple pages correspond to the English Named Entity or from it to the interwiki-linked Arabic Named Entity? The existence of disambiguation pages in Wikipedia can help in solving the problem. Although not all the cases of polysemy have a disambiguation page. Moreover, the way of going to a disambiguation page is not always straightforward, sometimes getting the redirection implies some kind of linguistic processing. For instance in the Wikipedia page of "Picasso", Figure 10, the following text occurs near the title: "This article is about the artist. For other uses, see Picasso (disambiguation)". In other cases the first page returned to a query is directly a disambiguation page.
- Arabic pages in Arabic Wikipedia are unvowelized. The problem for us is that Arabic WordNet, as we have discussed above, has to be vowelized. Of course this process can be made manually but our aim is to limit, as much as possible, human intervention, so an automatic solution of this problem has to be proposed.

The global architecture of our approach is shown in Figure 9.

First the set of PWN1.7 instances is obtained from Alfonseca's Web as discussed in section IV. Then using the TALP mappings the corresponding set of PWN2.0 instances is got.

The "Extracting Candidates" step consists of obtaining the generic types, i.e. the PWN2.0 synsets corresponding to the direct hyperonyms of the instance synsets, also as described in section V. The generic types not having Arabic counterparts are removed from the list. In some cases, however, as

described above, the Arabic generic type has been manually added to Arabic WordNet.

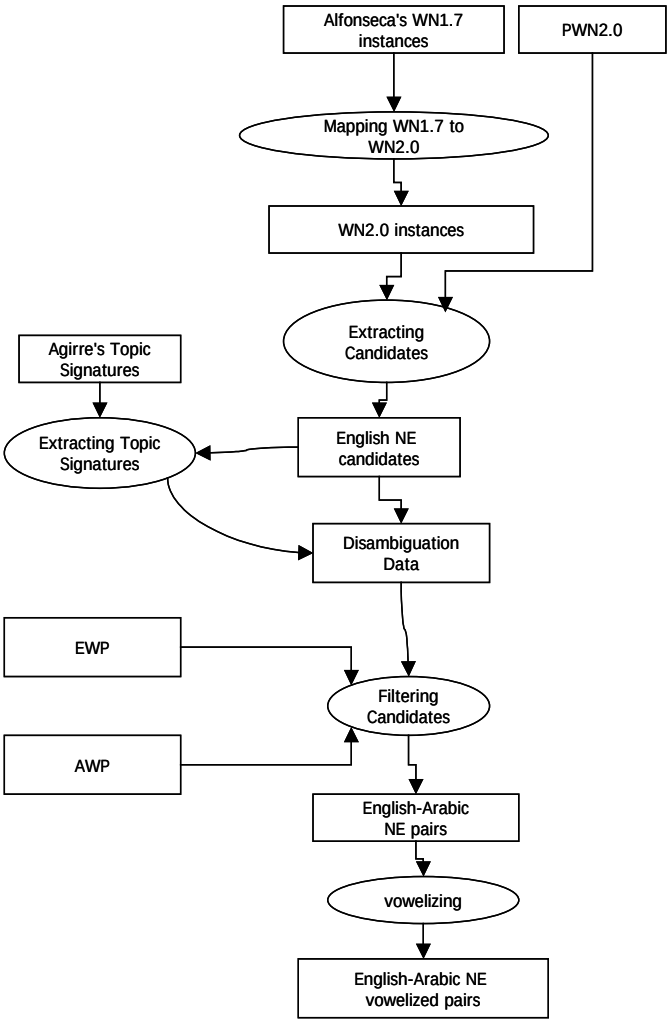

Figure. 9. Overall architecture of the system

In order to face the Word Sense Disambiguation problem, a process for adding disambiguation information to the generic types has been performed. We have used as disambiguation data three sets of words:

1. The set of variants (senses) included in each generic type synset.

2. The set of words occurring in the gloss after stopwords and example removing. The gloss is simply considered as a bag of words.

3. The Topic Signature of the English synset. The Topic Signature of a linguistic unit (in this case of a synset) is simply a list of weighted terms with high probability of occurring in the neighborhood of the unit. The technique was first introduced by Lin and Hovy [22] in the framework of Automatic Summarization. Later, Topic Signatures were used for Word Sense Disambiguation. We have used for our purposes the repository of Topic Signatures of IXA group at the University of the Basque Country[20]. This repository assigns to all ambiguous nominal English synsets their corresponding Topic Signature. In this case the words are weighted with a relevance score.

Consider, for instance, the word "painter" one of whose senses corresponds to generic synset "painter", direct hyperonym of most of the instance painters occurring in English WordNet. For this word three senses occur in WordNet 1.6 (the version of WordNet for which Topic Signatures are available). See Table 4. The terms with the highest scores from the Topic Signature for the three synsets in Table 4 are shown in Table 5.

So, the English synset for which we try to find Arabic counterpart has attached three data structures: the set of variants, the bag of words of the gloss, and the Topic Signature of the synset. With these data we have to face the Word Sense Disambiguation problem at Wikipedia level.

The core of our approach is the "Filtering Candidates" process. This process involves the use of English Wikipedia.

Sense	variants	gloss
1	painter	an artist who paints
2	painter	a worker who is employed to cover objects with paint
3	cougar, puma, catamount, mountain_lion, painter, panther, Felis_concolor	large American feline resembling a lion

Table 4. Senses corresponding to "painter" in WordNet 1.6

Among the systems described in section V for the management of Wikipedia we have chosen Iryna Gurevych's (Univ. of Darmstadt) JWPL system, [45]. This system is based on a local copy of Wikipedia loaded into a database (we have used MySQL as database management system). The local copy we have downloaded (for both Arabic and English Wikipedias) were those of February 2008. The system allows an easy recovering of all the data we need for our purposes[21]

[20] http://ixa.si.ehu.es/Ixa/resources/sensecorpus

[21] Unfortunately JWKL does not allow a direct recovery of "interwiki" links. As the system is monolingual, multilingual links are not included in the database tables and have to be extracted from text. Maintaining tables for interwiki links imply loading copies of all the Wikipedias for which

by means of APIs in Java. Using JWPL the procedure for each candidate (English Named Entity with disambiguation information attached) is the following:

1. Using the English Named Entity we look for it in English Wikipedia. If the page does not exist, the entry cannot reach hypothetical Arabic counterpart and is not taken into account. If the page corresponds to a redirection page, the link to a true content page is recovered directly from the corresponding table of the database. If the page is a disambiguation page or points to a disambiguation page, as discussed above, a disambiguation procedure is followed. In section A below we describe such procedure

2. The last step in the process of filtering candidates is looking for an occurrence of an "interwiki link" to an Arabic page. In this case the title of the page is returned as Arabic Named Entity. In the case the database contains redirection pages for this page, the alternate pages are considered too as Arabic Named Entities translation of the original English synset.

Sense	Topic Signature
1	landscapist(24.19) sculpturer(22.80) watercolourist(21.25) miniaturist(20.40) watercolorist(15.22) gauguin(14.76) utrillo(14.68) creative(14.14) colorist(13.75) dauber(13.52) abstract(09.95) oil(09.84) postimpressionist(09.73) master(09.44) constructivist(06.95)
2	funeral(33.65) bread(32.44) l ens(32.06) worker(27.36) lockmaster(23.37) tuner(20.38) harpooner(19.91) repairman(19.82) projectionist(19.35) slaughterer(18.04) lobsterman(17.95) mortician(17.57) maker(17.39) balloonist(17.39) optician(15.14)
3	felis(226.98) serval(81.67) ocelot(78.95) lynx(62.27) margay(51.38) bengal(51.04) jaguarundi(50.70) wildcat(49.00) manul(44.92) leopard(41.07) jungle(30.26) puma(25.86) jaguar(19.39) panther(10.67) f eline(10.54)

Table 5. 15 Most scored terms of the Topic Signatures of the three senses of "painter" in WordNet 1.6

A. Page Disambiguation

The procedure we have followed for page disambiguation is quite simple because we use as context for disambiguation only the text attached to the different options of the disambiguation page.

Basically what is done is deriving a unigram language model from the disambiguation information described above, i.e. variants, gloss and Topic Signature. The three language models are then merged into a unique one. From each of the options of the disambiguation page the likelihood that the text attached to it would be generated by this language model is computed. The option with the highest likelihood is considered as the correct page.

Fig. 10. English Wikipedia, Fragment of the page of "Pablo Picasso"

We have experimented with a linear combination of these three language models. The inclusion of Topic Signatures has resulted in all the cases in a drop of accuracy. This result could be a consequence of the noise present in the repository, at least for Topic Signatures attached to direct hyperonyms of Named Entities[22]. The weights assigned to the other components (variants and gloss) has been set to 2/3 and 1/3 respectively.

B. Term vowelization

The last step in our approach is vowelization. It is controversial if Named Entities have or do not have to be vowelized. In fact many Named Entity have different vowelization patterns depending on the geographic area. When designing Arabic Wordnet we decided to make the entries vowelized and this decision was applied both to normal entries and to Named Entities. So, when building Arabic Wordnet we performed a manual vowelization using the criterion of assigning the most common vowelization pattern

interwikis exist. Anyway, the procedure for extracting interwiki links is very simple.

[22] Note that Topic Signatures are available only for ambiguous, i.e. polysemous, terms. Most of generic terms direct hyperonyms of Named Entities are monosemous and thus lack Topic Signature.

to each entry. In this extension we apply the same criterion.

The Arabic alphabet consists of 28 letters that can be extended to a set of 90 by additional shapes, marks, and vowels (motions). The 28 letters represent the consonants and long vowels such as ,ا (pronounced as /a:/), ي (pronounced as /i:/), and و (pronounced as /u:/). The short vowels (Sukoon, ْ, represents no vowel at all, Fatha, َ, represents the /a/ sound, Kasra, ِ, represents the /i/ sound, and Damma, ُ, represents the /u/ sound) and certain other phonetic

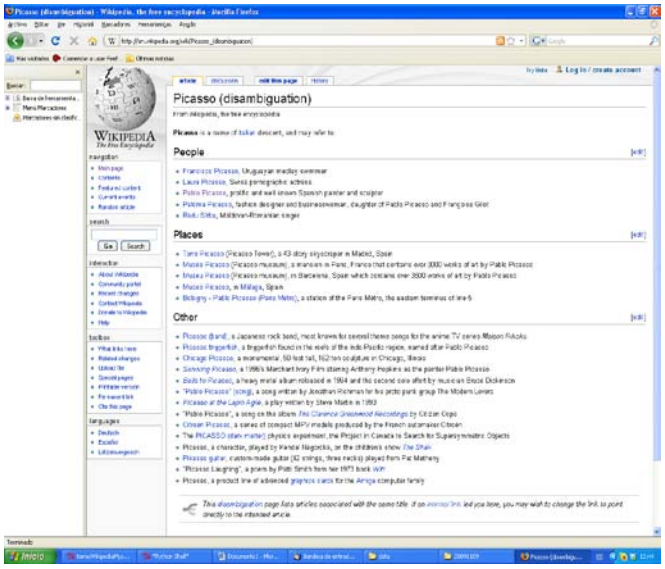

Fig. 11. English Wikipedia, Fragment of the page of "Pablo Picasso"

information such as consonant doubling, the gemination mark (shadda, ّ,) are not represented by letters, but by diacritics.

A diacritic is a short stroke placed above or below the consonant. The doubled case ending diacritics (nunation) are vowels used at the end of the words; the term ‘‘tanween” is used to express this phenomenon. Tanween marks indefiniteness and it is manifested in the form of case marking or in conjunction with case marking. Similar to short vowels, there are three different diacritics for tanween: tanween al-fath, tanween al-damm, and tanween al-kasr. They are placed on the last letter of the word and have the phonetic effect of placing an ‘‘N” at the end of the word.

The problem of automatic generation of the Arabic diacritic marks is known in the literature under various translations (such as automatic vocalization, vowelization, diacritization, accent restoration, and vowel restoration). The formal approach to the problem of restoration of the diacritical marks of Arabic text involves a complex integration of the Arabic morphological, syntactic, and semantic features.

Arabic diacritic restoration is a non-trivial task. Native speakers of Arabic are able, in most cases, to accurately vocalize words in text based on their context, the speaker’s knowledge of the grammar, and the lexicon of Arabic. The goal of diacritic recovering algorithms is to convert knowledge used by native speakers into features that could be used by the system (usually a Machine Learning algorithm) to perform the task.

There are several vowelization, in general diacritic recovering algorithms. Most of the early methods were rule-based, as in [13]. More recently statistical and Machine Learning algorithms were used. [19] and [15] use HMM approaches, [46] uses a Maximum Entropy approach.

Unfortunately none of these approaches can be applied to Named Entities. All these methods use the context of the word to be diacritized as source of features for the task. In this way the results for vowelizing normal words (nouns, verbs, etc.) are usually good. Although no statistics are provided for Named Entities there is a notable drop in the accuracy when dealing with Named Entities. The lack of context is obviously another

title	description
Francisco Picasso	Uruguayan medley swimmer
Laura Picasso	Swiss pornographic actress
Pablo Picasso	prolific and well known Spanish painter and sculptor
Paloma Picasso	fashion designer and businesswoman, daughter of Pablo Picasso and Françoise Gilot
Torre Picasso	(Picasso Tower), a 43-story skyscraper in Madrid, Spain

Table 6. Some of the entries of the disambiguation page of Pablo Picasso (from 19 possibilities)

limitation but the results with and without context do not differ significantly.

For our task only short vowel restoration is needed. Shadda diacritics are already recovered and do not need restoration. Moreover in many cases all or most of the vowels are included as long vowels and do not need restoration.

We have implemented a very simple HMM-based algorithm. We used for learning the set of 1,656 vowelized words corresponding to Named Entities in Arabic WordNet. We converted this set into a set of pairs <vowelized Named Entity/unvowelized Named Entity> simply by removing vowels of the set. We used the GHMM[23] library for managing the HMM. We tried first to vowelize Arabic Named Entities without context. Then we added context using for this purpose

[23] General Hiden Markov Model Library (GHMM), http:// ghmm.sourceforge.net

the sentences in · the Arabic GigaWord Corpus[24] where the Named Entities occur.

The results, as expected, were bad, almost 20% worse than the reported for general words in the literature (and in our experiments). No improvement was obtained when including contextual information. So we decided to follow an ad-hoc approach that took into account the characteristics of Arabic Named Entities (at least, and this is an important constraint, those connected to English WordNet synsets).

We follow here a rather conservative approach. We consider four cases for vowelization:

- Many cases correspond to direct transliteration of foreign words (i.e. they are named arabizations) and usually the Arabic term includes long vowels, for representing vowels in the original language. In such cases no vowelization is needed. For instance, the Named Entity "Pablo Picasso" is interwiki linked to "بابلو بيكاسو." In this case the vowels are included in the Arabic form as long vowels and, so, no recovery is needed.

- Some cases correspond also to direct transliteration of foreign words but some (or all) of the vowels are not long and need to be recovered. In this case we have transliterated the Arabic Named Entity into Buckwalter encoding, [9] and then compared it with English, French, Italian and Spanish translations[25] (using "interwiki links" if available) for choosing the best match. Consider the case of "Barcelona" that is interwiki linked to "برشلونة". In this case not all the vowels are long vowels and, so they have to be recovered. The interwikies of the Arabic page with English and Spanish point to "Barcelona", the corresponding to French to "Barcelone" and the corresponding to Italian to "Barcellona". The best match leads, in this case to the correct vowelization.

- Some Arabic Named Entities correspond to normal words occurring in Arabic Wordnet and can be vowelized accordingly. Some Arabic Named Entities correspond to multiwords with elementary components existing in Arabic WordNet, we proceed then in the same way. The paradigmatic example is "Casablanca". The entry is interwiki linked to " البيضاء الدار ", that is not fully vowelized. "Casablanca" can be decomposed into "casa" (house) and "blanca" (white). Both are normal words occurring in English WordNet and linked to vowelized entries in Arabic WordNet. So we can assign to the Arabic Named Entity the vowels occurring in the component words.

- The rest of cases correspond to Arabic Named Entities with no direct connection with foreign terms and corresponding to no normal words. In this case we left the vowelization unsolved in the automatic phase and delayed the solution to a posterior manual intervention. An example of this case is "Jerusalem" that is interwiki linked to "القدس". In this case no vowelization is proposed[26].

VII A DETAILED EXAMPLE

In this section we present a detailed example illustrating the approach described in section VI. Consider the case of the generic synset "painter" corresponding to the first sense of the word "painter" as shown in Table 4. The three disambiguation knowledge sources are:

1. The set of variants, in this case reduced to {painter}. So, the bag of words is simply {painter}.

2. The gloss, "an artist who paints". So, the bag of words is simply {artist, paint}, after stopwords removing and lemmatization.

3. The Topic Signatures, see Table 5. The bag of words here is weighted: {landscapist (24.19), sculpturer (22.80), watercolourist (21.25), miniaturist (20.40), watercolorist (15.22), etc.}.

We got the set of direct hyponyms of this synsets. Consider the case of one of them, the corresponding to "Pablo Picasso". In this case there is an entry in the Wikipedia (with several redirections) presented in Figure 10. From the main page we obtain the disambiguation page after processing the sentence " This article is about the artist. For other uses, see Picasso (disambiguation)." and following the link. The disambiguation page is shown in Figure 11. Some of the disambiguation items are presented in Table 6.

[24] available through LDC: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T02

[25] We thought that in the case of foreign words (whatever the direction, from or to Arabic) the languages involved should be geographically or culturally closed to an Arabic country. Including other languages does not seem to be useful.

[26] In the English page of "Jerusalem" the following redirection and disambiguation information apperars: "al-Quds" redirects here. For other uses, see al-Quds (disambiguation)". The reference here to "al-Quds" gives a clue to the vowel restoration. How to make use of this is unclear now, but will be considered in the future.

From the content of the disambiguation Knowledge Sources, the correct page is selected (Pablo Picasso, the painter).

There exists an interwiki link to the corresponding page of the Arabic Wikipedia, "بابلو بيكاسو." In this case the all the vowels are included in the Arabic form as long vowels and no recovery is needed. This is an example of the first case discussed above.

The Arabic title of the page, "بابلو بيكاسو.", is thus considered a correct translation of the Named Entity "Pablo Picasso" and, thus, incorporated to Arabic WordNet and linked as an "equivalent" of the English synset "Pablo Picasso" and as an "hyponym" of "painter".

VIII Results and evaluation

In our experiments we started with 16,873 English Named Entities occurring as instances in PWN2.0. From them, 14,904 occurs as well in English Wikipedia as article titles. This is a really nice coverage (88%). 3,854 Arabic words corresponding to 2,589 English synsets were recovered following our approach. The coverage (26%) is really high taking into account the relatively small size of Arabic Wikipedia. From the recovered synsets only 496 belonged to the set of Named Entities already included in Arabic WordNet following the manual procedure described in section IV.

The obvious way of evaluating our system consists of comparing the obtained Named Entities with the manually collected and manually incorporated Named Entities that are already in Arabic WordNet. From the 496 synsets included in both sets 464 were the same and 32 differed (and thus could be considered errors). The accuracy measured in this way was of 93.4%. As the size of the automatically evaluated set was small (only 496 synsets, i.e. 12% of the set of the recovered synsets) we decided to perform a manual validation of the set. The set of Arabic Named Entities was, thus, fully evaluated (by one of the authors[27]).

From the 3,854 proposed assignments, 3,596 (93.3%) were considered correct, 67 (1.7%) were considered wrong and 191 (5%) were not known by the reviewer. There is, so, a high coincidence between the automatic and manual validation procedures.

We can conclude, thus, that our approach is highly reliable and can be used for the task.

[27] Musa Alkhalifa

IX Conclusions and future work

We have presented an approach for automatically attaching Arabic Named Entities to English Named Entities using Arabic WordNet, Princeton's WordNet, Arabic Wikipedia and English Wikipedia as Knowledge Sources. The system is fully automatic, quite accurate, and has been applied to a substantial enrichment of the Named Entity set in Arabic WordNet.

Due to the high growing ratio of Arabic Wikipedia the approach can be applied to progressively improve Named Entity coverage of Arabic WordNet. An automatic way of incorporating to Arabic WordNet new Named Entities coming from the enrichment of Arabic Wikipedia is an obvious extension of our system.

Besides this task we will try to apply a similar procedure for building a multilingual (including Arabic, Catalan, English and Spanish)[28] geographical ontology based on GEONAMES[29] and GNIS[30] databases. Another task that could make use of our approach is the automatic extraction of transliterated pairs from bilingual (or comparable) corpora.

Acknowledgments

This work has been partially supported by the Spanish Government in the framework of the projects KNOW-2, TIN2009-14715-C04-04 . and AECID-C/026728/09.

We have to acknowledge, too, the valuable suggestions and comments of three anonymous reviewers. Their help has resulted in a clear improvement of this paper.

References

[1] E. Agirre, O. Ansa, X. Arregi, J. Arriola, A. Díaz de Ilarraza, E. Pocielo, L. Uría (2002) Methodological issues on the building of the Basque WordNet: Quantitative and qualitative análisis. In Proceedings of the first International Global WordNet Conference, Mysore, India, 21-25 January 2002.

[2] Al-Serhan, H.M. & Ayesh, A.S. (2006) A Triliteral Word Roots Extraction Using Neural Network for Arabic. In: IEEE Int. Conf. on Computer Engineering and Systems ICCES06, Cairo, Egypt 5-7 Nov., pp. 436-440 (2006)

[28] These are the languages we are currently working with in our group at UPC.

[29] http: //geonames . usgs . gov/geonames /stategaz

[30] http://geonames.usgs.gov/pls/gnispublic/

[3] Atserias, J. Zaragoza, H. Ciaramita M. and Attardi. G. (2008) Semantically Annotated Snapshot of the English Wikipedia. Proceedings of the Sixth International Language Resources and Evaluation (LREC-2008).

[4] E. Barbu, V. Barbu-Mititelu, (2006) A case study on automatic building of WordNets. In Proceedings of OntoLex, Ontologies and Lexical Resources, 2005

[5] Kenneth R. Beesley and Lauri Karttunen, (2003) Finite State Morphology, CSLI Publications, 2003

[6] L. Benitez, S. Cervell, G. Escudero, M. López, G. Rigau, M. Taulé (1998) Methods and tools for building the Catalan WordNet in Proceedings of LREC workshop on Language Resources for European Minority Languages, 1998.

[7] T. Berners-Lee, M. Fischetti (1999) Weaving the Web: The Original Design and Ultimate Destiny of the World Wide Web by its inventor, Harper Collins Publishers, New York, 1999.

[8] Black, W., Elkateb, S., Rodriguez, H, Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C., (2006). Introducing the Arabic WordNet Project. In Proceedings of the Third International WordNet Conference, Fellbaum and Vossen (eds).

[9] Buckwalter, T. (2002) Arabic Morphological Analysis, http://www.qamus.org/morphology.htm

[10] H. Chen, C. Lin, W. Lin (2002) Building a Chinese-English WordNet for translingual applications. ACM transactions on Asian Languages Information Processing, Vol. 1, Num. 2, June 2002, pages 103-122.

[11] De Roeck, A.N. & Al-Fares, W. (2000) A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots. In: Proc. of 38th Annual Meeting on Association for Computational Linguistics, Hong-Kong, Oct., pp. 199–206 (2000)

[12] Diab, Mona (2004). The Feasibility of Bootstrapping an Arabic WordNet leveraging Parallel Corpora and an English WordNet. Proceedings of the Arabic Language Technologies and Resources, NEMLAR, Cairo 2004.

[13] El-Imam, Y., (2003). Phonetization of arabic: rules and algorithms. Computer Speech and Language 18, 339–373.

[14] Elkateb, S., Black, W., Rodriguez, H., Alkhalifa, M., Vossen, P., Pease, A., and Fellbaum, C., (2006). Building a WordNet for Arabic. In Proceedings of the Fifth International Conference on Language Resources and Evaluation, Genoa, Italy.

[15] Moustafa Elshafei, Husni Al-Muhtaseb, Mansour Al-Ghamdi: Machine Generation of Arabic Diacritical Marks. MLMTA 2006: 128-133

[16] M. Erdmann, K. Nakayama, T. Hara, S. Nishio (2008): Extraction of Bilingual Terminology from a Multilingual Web-based Encyclopedia, in IPSJ Journal of Information Processing (Jul. 2008).

[17] Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge, MA: MIT Press.

[18] Ferrández, S. Toral, A. Ferrández, O. Ferrández, A. Muñoz R. (2007) Applying Wikipedia's Multilingual Knowledge to Cross-Lingual Question Answering. In Proceedings of NLDB 2007: 352-363

[19] Ya'akov Gal (2002) An HMM Approach to Vowel Restoration in Hebrew and Arabic. In Proceedings of ACL-2002 Semitic Language Workshop

[20] L.R. Khan, E. Hovy (1997) Improving the precision of lexicon-to-ontology alignment algorithms. In Proceedings of the AMTA-SIG-IL First Workshop on Interlinguas. San Diego, California, 1997.

[21] Kazama, J. and Torisawa, K. (2007) Exploiting Wikipedia as External Knowledge for Named Entity Recognition. Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning

[22] Lin, C.-Y. and E.H. Hovy. (2000) The Automated Acquisition of Topic Signatures for Text Summarization. Proceedings. of the COLING, Conference. Strasbourg, France. August, 2000.

[23] Magnini, B., and Cavaglia, G. (2000) Integrating Subject Field Codes into WordNet. In Gavrilidou M., Crayannis G., Markantonatu S., Piperidis S. and Stainhaouer G. (Eds.) Proceedings of the Second International Conference on Language Resources and Evaluation, Athens, Greece, 31 May- 2 June, 2000, pp. 1413-1418.

[24] Medelyan, O. Witten, I. H. Milne D. (2008) Topic indexing with Wikipedia. In Proceedings of Wikipedia and AI workshop at the AAAI-08 Conference. Chicago, US

[25] Olena Medelyan, David N. Milne, Catherine Legg, Ian H. Witten (2009) Mining meaning from Wikipedia. International. Journal. Human-Computer Studies. 67(9) pages 716-754 (2009)

[26] Mihalcea, R. Csomai A. (2007) Wikify!: linking documents to encyclopedic knowledge. In Proceedings of CIKM 2007: 233-242

[27] M. Mihaltz, G. Proszeky (2004) Results and evaluation of the Hungarian nominal WordNet V1.0. In Proceedings of the Second Global WordNet Conference, GWC 2004, Masaryk University, Brno, 2003, pags. 175-180.

[28] Milne, D., Medelyan, O. and Witten, I.H. (2006) Mining Domain-Specific Thesauri from Wikipedia: A case study. In Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence, WI'06, pp. 442-448, Hong Kong, China, December.

[29] Niles, I., and Pease, A., (2001) Towards a Standard Upper Ontology. In Proceedings of FOIS 2001, Ogunquit, Maine, pp. 2-9.

[30] A. Okumura, E.Hovy (1994) Buildina a Japanese-English dictionary based on ontology for Machine Translation,. In Proceedings of the ARPA Human Language Technology Conference, Princeton, NJ, 1994, pages 236-241.

[31] Pease, A. (2003) The Sigma Ontology Development Environment. In Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems, Vol. 71 of the CEUR Workshop Proceeding series.

[32] Ponzetto, P; Strube, M. (2008). WikiTaxonomy: A large scale knowledge resource. In: Proceedings of the 18th European Conference on Artificial Intelligence, Patras, Greece, 21-25 July, 2008 pp. 751-752.

[33] Richman, A.. and Schone, P. (2008) Mining Wiki Resources for Multilingual Named Entity Recognition. In Proceedings of ACL-08

[34] H. Rodríguez, S. Climent, P. Vossen, L. Bloksma, W. Peters, A. Roventini, F. Bertagna, A, Alonge (1998) The top-down strategy for building EuroWordNet, Vocabulary coverage, base concepts and top ontology.. Computers and the Humanities. Special Issue in EuroWordNet. Vol. 32 (1998). Pages 117-152.

[35] Rodríguez, H., Farwell, D., Farreres, J., Bertran, M., Alkhalifa, M., Martí, M.A., Black, W., Elkateb, S., Kirk, J., Pease, A., Vossen, P., and Fellbaum, C., (2008). Arabic WordNet: Current State and Future Extensions. In Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary. January 22-25, 2008.

[36] Rodríguez, H. Farwell, D. Farreres, J. Bertran, M. Alkhalifa, M. Martí M.A (2008) Arabic WordNet: Semi-automatic Extensions using Bayesian Inference. In Proceedings of the 6th Conference on Language Resources and Evaluation LREC-2008. Marrakech (Morocco), May 2008.

[37] Suchanek F. (2008) Automated Construction and Growth of a Large Ontology PhD-Thesis. Max-Planck-Institute for Informatics. U. Saarbrücken, Germany

[38] Toral, A. Muñoz. R. (2006) A proposal to automatically build and maintain gazetteers for Named Entity Recognition using Wikipedia. In Proceedings of the Workshop on New Text, 11th Conference of the European Chapter of the Association for Computational Linguistics. Trento (Italy). April 2006.

[39] Tufis, D. (2004 ed.) Special Issue on the BalkaNet project. Romanian Journal of Information Science and Technology, Vol. 7, nos 1-2.

[40] Vossen P. (2004) EuroWordNet: a multilingual database of autonomous and language specific wordnets connected via an Inter-Lingual-Index. International Journal of Lexicography, Vol.17 No. 2, OUP, 161-173.

[41] Wehr, H. (1976) Arabic English Dictionary. The Hans Wehr Dictionary of Modern Written Arabic. Edited by J. Milton Cowan. Spoken Languages Services Inc. Ithaca, New York.

[42] Witten, I.H. and Frank, E. (2005) Data mining: Practical machine learning tools and techniques (second edition). San Francisco, CA: Morgan Kaufmann.

[43] Wu, F. Hoffmann, R. Weld D. (2007) Autonomously Semantifying Wikipedia, In Proceedings of the Sixteenth ACM Conference on Information and Knowledge Management (CIKM-07), Lisbon, Portugal, November, 2007.

[44] Zesch, T. Gurevych, I. Analysis of the Wikipedia Category Graph for NLP Applications. In Proceedings of the (Workshop on TextGraphs: Graph-Based Algorithms for Natural Language Processing, 2007)

[45] Zesch, T. Müller C. and Gurevych I. (2008) Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the 6th Conference on Language Resources and Evaluation, LREC-2008. Marrakech (Morocco), Mai 2008.

[46] Imed Zitouni, Ruhi Sarikaya (2009) Arabic diacritic restoration approach based on maximum entropy models. In Computer Speech & Language 23(3): 257-276 (2009)