Automatically Extending NE coverage of Arabic WordNet using Wikipedia

Musa Alkhalifa* and Horacio Rodríguez**
* Universitat de Barcelona (UB), Barcelona, Spain, musa@thera-clic.com
** Universitat Politècnica de Catalunya, Barcelona, Spain, horacio@lsi.upc.edu

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

***Keywords*—Arabic NLP, Arabic WordNet, Named Entities Extraction, Wikipedia**

I. Introduction

An Arabic WordNet (AWN – [2], [4], [16], [17]) 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 (PWN – [5]). In addition, the AWN project aimed to provide a formal specification of the senses of its synsets using the Suggested Upper Merged Ontology (SUMO – [12]). This representation serves as an interlingua among all wordnets ([13], [20]) and will underlie the development of semantics-based computational tools for multilingual NLP.

In Accordance with the objectives of the project, Arabic WordNet currently consists of 11,269 synsets (7,960 nominal, 2,538 verbal, 661 adjectival, and 110 adverbial), containing 23,481 Arabic expressions. This number includes 1,142 synsets that correspond to named entities which have been extracted automatically and were checked by the lexicographers. For the most up-to-date statistics see:

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

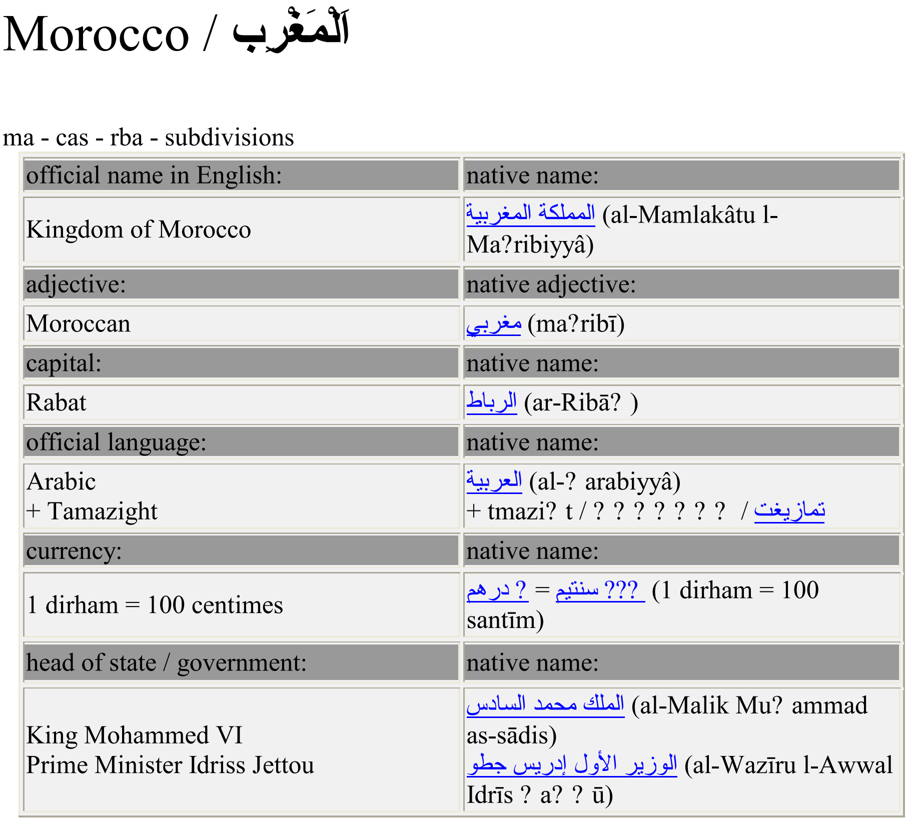
According to the conditions of AWN project, all the content of AWN was manually built or at least, in the case of NEs, manually revised. In the case of NEs the coverage is rather limited and our present goal in this paper is to automatically enrich the current set using high quality sources such as the Arabic Wikipedia (AWP).

The organization of this paper after this introduction is as follows: Section II revises the way we followed for collecting the NEs currently included in AWN. Section III discuses the potential use of the Wikipedia as source for enriching the set of NEs. Section IV outlines our approach. Result and evaluation are presented in section V. Finally in section VI our conclusions and further work are presented.

II. Collecting NEs in AWN project

The process of collecting NEs for being included in AWN followed a semi-automatic approach. The process consisted of two steps: 1) Selection of the candidates and 2) Manual validation.

According to the conditions of our contract, at least 1,000 NEs synsets should be built, covering a variety of types (locations, persons, organizations, etc.) that should be, whenever possible, linked to existing instances in PWN.



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 1. A fragment of GEONAMES database

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; the GEONAMES[1] database for toponyms[2] information corresponding to Arabic countries, a gazetteer of Countries in the world from FAO[3] and the NE entries contained in the NMSU (New Mexico State University) bilingual Arabic-English lexicon[4]. 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 1), then we wrote wrappers for extracting information from these web pages and formatting results. For FAO, we simply aligned English and Arabic NEs by means of the ISO code (see Figure 2).

The case of NMSU was more complex. The database had a larger coverage but the entries had no diacritics[5] at all (including "shadda" diacritic[6]) and obviously not only NEs but also normal entries are included in the dictionary. We proceed in the following steps:

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

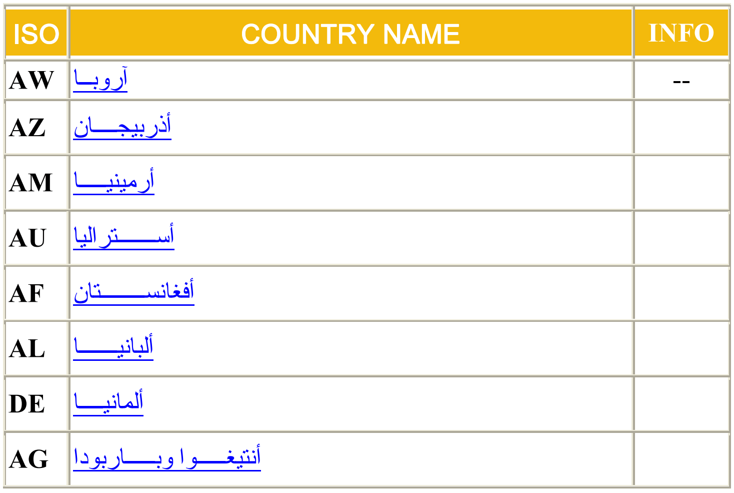
[2] Besides toponym information (country name, capital, main cities, organizative districts, etc.) this resource contains other NE information, such as the current head of state, the head of govern, the currency, and other. See Figure 1 for some examples.

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

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

[5] Most Arabic texts are unvowelized, i.e. do not contain diacritics.

[6] "shadda" is an important diacritic in Arabic, marking consonant duplication.



ISO	COUNTRY NAME	INFO
AW	آروبــا	--
AZ	أذربيجـــان	
AM	أرمينيـــا	
AU	أســـــتراليا	
AF	أفغانســــــتان	
AL	ألبانيـــــا	
DE	ألمانيـــا	
AG	أنتيغـــوا وبـــاربودا	

Figure 2. A fragment of FAO database

- Identifying synsets corresponding to instances. From Enrique Alfonseca's page[7] 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[8] between different versions of PWN.
- Obtaining the generic types, i.e. the PWN2.0 synsets corresponding to the direct hyperonyms of the instance synsets. This resulted in obtaining 371 generic types from which only synsets linked to AWN were collected (such as 'capitals', 'cities', 'countries', 'inhabitants' or 'politicians').
- Obtaining NMSU entries corresponding to the variants in instance synsets. Only nominal entries were recovered. For example, for instances of hyponyms of the generic entry 'politicians' 129 synsets were found.
- Formatting and merging the results of the three sources.

B. Manual validation

This step iterates on the associations proposed by the 1st step. For each candidate the following tasks are performed:

- Deciding the acceptance or rejection of the pair.
- Modifying Arabic form if needed.
- Adding diacritics.
- Completing attachments to PWN2.0 if possible.

The whole procedure resulted in getting 1,147 synsets that have in total 1,659 variants corresponding to 31 generic types. Figure 3 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.

III. Wikipedia as source of lexical resources

Wikipedia[9] (WP), is by far the largest encyclopedia in existence with almost 3 million articles in its English version (EWP) contributed by hundreds of thousands of volunteers. WP experiments an exponential growing. There are versions of WP in more than 200 languages although the coverage (number of articles and average size of each article) is very irregular.. The Arabic version (AWP) has over 65,000 articles (about 1% of the total size of WP). Among all the different languages, Arabic has a rank of 29, just above Serbian and Slovenian. The growing of AWP is, however, very high (more than 100% of last year) so it seems that in a relatively short time the size of AWP could correlate with the

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

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

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

importance (of the number of speakers) of Arabic language.

WP 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 WP article. The set of articles and their links in WP form a graph. WP articles can be assigned to WP categories (through "Category links") that are also organized as a graph (see [22] for an interesting analysis of both graphs). Besides article and category links, WP pages can contain "External links", that point to external URLs, and "Interwiki links", from an article to a presumably equivalent, article in another language. There are in WP 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.

WP has been extensively used for extracting lexical and conceptual information. [14], and [18] build or enrich ontologies from WP, [11] derive domain specific thesauri, [1] produce a semantically annotated snapshot of EWP, [9], [10], and [21] perform semantic tagging or topic indexing with WP articles. Closer to our approach are the works of [19] and [7] where they used WP, 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 WP. [6] and [15] are notable exceptions.

Extracting information from WP can be done easily using a Web crawler and a simple html parser. The regular and highly structured format of WP pages allows this simple procedure. There are, however, a lot of APIs providing easy access to WP online or to the database organized data obtained from WP dumps[10]. Some interesting systems are Waikato's WikipediaMiner toolkit[11], U. Alicante's wiki db access[12], Strube and Ponzetto's set of tools[13], Iryna Gurevych' JWPL[14], etc.

IV. Our approach

At first glance, given an English NE, obtaining the Arabic counterpart using WP seems to be easy: We can recover the page corresponding to the English NE. 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:

- Which English NEs have to be looked for?
    - We can consider all the EWP pages but in this case i) we are introducing a lot of noise in the case of pages not corresponding to a NE, and ii) the page has to be mapped to a PWN synset and thus a possible Word Sense Disambiguation, WSD, problem arises. For instance, looking at WP 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.
    - We can start not from WP but from PWN. In this case we have to locate in PWN the set of instances (using the same procedure described in section A) and we have to face the same problem of WSD in this case not against PWN but against the EWP. We have chosen this later approach.
- How to deal with polysemy, i.e. when multiple pages correspond to the English NE or to the interwiki-linked Arabic NE? The existence of disambiguation pages in WP can help in solving the problem.
- Arabic pages in AWP are unvowelized. The problem for us is that AWN have to be vowelized. Of course this process can be made manually but our aim is to limit, as much as possible, human intervention.

The global architecture of our approach is shown in Figure 4.

First the set of PWN1.7 instances is obtained from Alfonseca's Web as discussed in section A. 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 A.

[10] http://en.wikipedia.org/wiki/Wikipedia_database

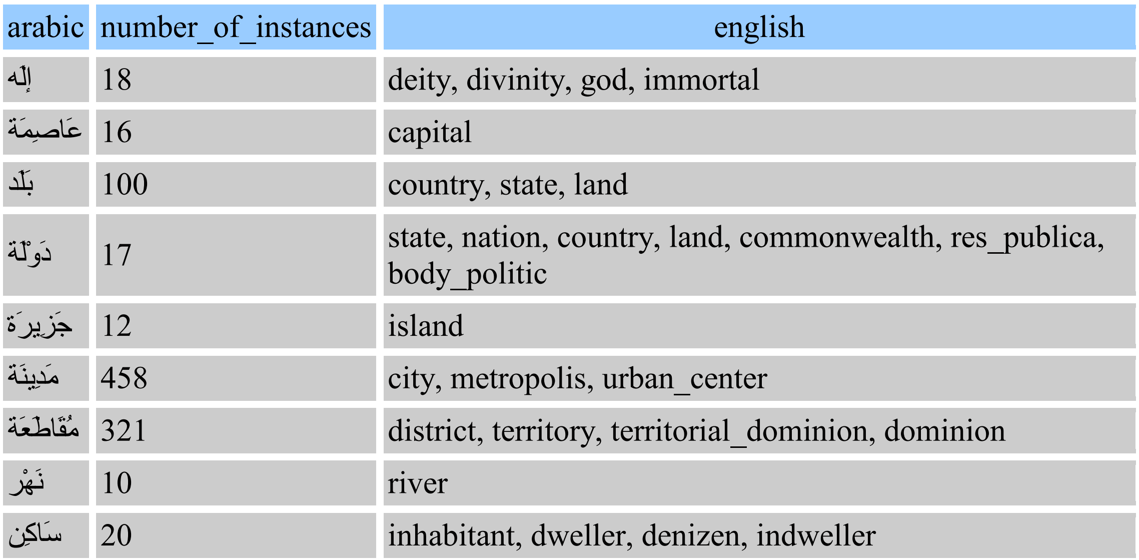
[11] http://wikipedia-miner.sourceforge.net/

[12] http://www.dlsi.ua.es/~atoral/

[13] http://www.eml-research.de/english/research/nlp/download/

[14] http://www.ukp.tu-darmstadt.de/software/jwpl/

The generic types not having Arabic counterpart are removed from the list.



arabic	number_of_instances	english
إِلَٰه	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

Figure 3. Distribution of AWN NE coverage by generic type (most frequent types).

In order to face the WSD problem, a process for adding disambiguation information to the generic types is performed. We have used as disambiguation data three sets of words: i) the set of variants (senses) of each generic type, ii) the set of words occurring in the gloss (after stopwords and example removing) and iii) the topic signature, TS, (in this case the words are weighted with a relevance score). The TS of a linguistic unit (in this case of a synset) is simply a weighted list of terms with high probability of occurring inn the neighborhood of the unit. The technique was first introduced by [8] in the framework of Automatic Summarization. We have used the repository of TS of IXA group[15].

The core of our approach is the "Filtering Candidates" process. This process involves the use of EWP. Among the systems described in section 3 we have chosen Iryna Gurevych (U. Darmstadt) JWPL system, [23]. This system is based on a local copy of WP loaded into a database (we have used MySQL). The system allows an easy recovering of all the data we need for our purposes[16] by means of APIs in Java. Using JWPL the procedure for each candidate (English NE with disambiguation information attached) is the following:

Using the English NE we look for it in EWP. If the page does not exists, the entry has no Arabic counterpart. If the page corresponds to a redirection page, the links are followed in sequence until arriving to a true content page.

If the page is a disambiguation page or points to a disambiguation page, a disambiguation procedure is followed. Basically what is done is deriving a unigram language model from the disambiguation information described above (we have experimented with a language model formed with a linear merging of the three components of the disambiguation information, variants, gloss and TS)[17] and computing the likelihood of the text attached to each of the options of the disambiguation page being generated by this language model. The option with the highest likelihood obtained is considered as a correct page. If the page is a normal one, it is chosen directly as the correct page.

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

[16] 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. Anyway the procedure for doing so is very simple.

[17] The inclusion of TSs resulted in a drop in accuracy due to the great quantity of noise existing in the data. The best results were obtained with doubling the weights of variants with respect to the gloss.

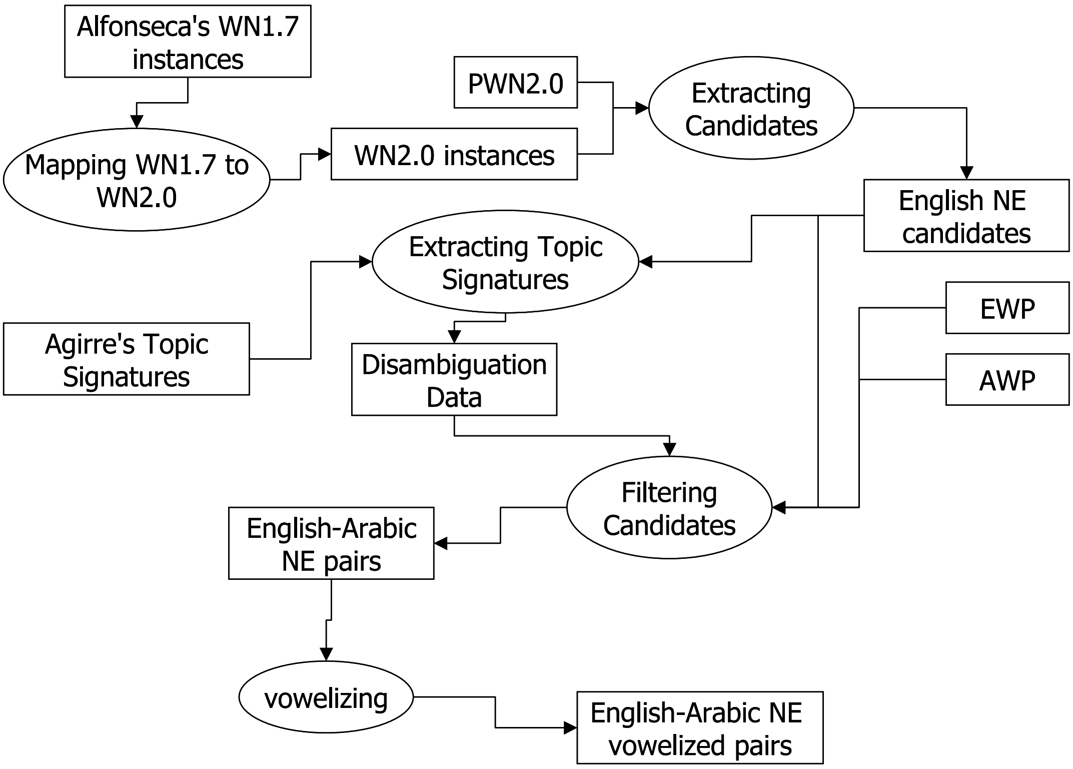


Figure 4. Overall architecture of the system

The last step 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 NE.

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

There are several vowelization (in general diacritic recovering) algorithms. Unfortunately none of them can be applied to NEs. We follow here a rather conservative approach. We consider four cases for vowelization:

Many cases correspond to direct transliteration of foreign words and usually the Arabic term includes long vowels, in such cases no vowelization is needed.

Some cases correspond also to direct transliteration of foreign words but some (or all) of the vowels are not long and have been recovered. In this case we have transliterated the Arabic NE into Buckwalter encoding, [3] and then compared it with English, French, Italian and Spanish translations[18] (using "interwiki links" if available) for choosing the best match.

Some Arabic NEs correspond to normal words occurring in AWN and can be vowelized accordingly. Some Arabic NEs correspond to multiwords with elementary components existing in AWN, we proceed then in the same way.

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

The rest of cases correspond to Arabic NEs with no direct connection with foreign terms and corresponding to no normal words. Iin this case we left the vowelization unsolved.

V. Results and evaluation

In our experiments we started with 16,873 English NE occurring as instances in PWN2.0. From them 14,904 occurs as well in EWP 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 small size of AWP. From the recovered synsets only 496 belonged to the set of NEs included in AWN.

The obvious way of evaluating our system consists on comparing the obtained NEs with the manually collected and incorporated sofar to AWN. From the 496 synsets included in both sets 464 were the same and 32 differed (and could, so, 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 NEs was, thus, fully evaluated (by one of the authors). 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. There is, so a high coincidence between the automatic and manual validation procedures.

VI. Conclusions and future work

We have presented an approach for automatically attaching Arabic NEs to English NEs using AWN, PWN, AWP and EWP as Knowledge sources. The system is fully automatic, quite accurate, and has been applied to a substantial enrichment of the NE set in AWN.

Due to the high growing ratio of AWP the approach can be applied to progressively improve NE coverage of AWN.

Besides this task we will try to apply a similar procedure for building a multilingual (including Arabic, Catalan, English and Spanish languages) geographical ontology based on GEONAMES and GNIS[19] 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 was supported by the Spanish Government in the framework of the project Text/Mess, TIN2006-15265-C06

References

[1] 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'08).

[2] 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).

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

[4] 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.

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

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

[7] 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

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

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

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

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

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

[13] 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.

[14] Ponzetto, P; Strube, M. (2008). WikiTaxonomy: A large scale knowledge resource. In: Proceedings of the 18th

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

European Conference on Artificial Intelligence, Patras, Greece, 21-25 July, 2008 pp. 751-752.

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

[16] Rodríguez, R., 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. Proceedings of The Fourth Global WordNet Conference, Szeged, Hungary. January 22-25, 2008.

[17] 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 the 6th Conference on Language Resources and Evaluation LREC2008. Marrakech (Morocco), May 2008

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

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

[20] 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.

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

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

[23] Zesch, T. Müller C. and Gurevych I. (2008) Extracting Lexical Semantic Knowledge from Wikipedia and Wiktionary. In: Proceedings of the Conference on Language Resources and Evaluation (LREC), electronic proceedings, Mai 2008.