Arabic WordNet: Semi-automatic Extensions using Bayesian Inference

Horacio Rodríguez, David Farwell, Javi Farreres, Manuel Bertran
Universitat Politècnica de Catalunya
Jordi Girona, 1-3; 08034 Barcelona; Spain
E-mail: horacio, farwell, farreres, mbertran@lsi.upc.edu

Musa Alkhalifa, M. Antonia Martí
Universitat de Barcelona
Gran Via 585; 08007 Barcelona; Spain
E-mail: musa@thera-clic.com, amarti@ub.edu

Abstract

This presentation focuses on the semi-automatic extension of Arabic WordNet (AWN) using lexical and morphological rules and applying Bayesian inference. We briefly report on the current status of AWN and propose a way of extending its coverage by taking advantage of a limited set of highly productive Arabic morphological rules for deriving a range of semantically related word forms from verb entries. The application of this set of rules, combined with the use of bilingual Arabic-English resources and Princeton's WordNet, allows the generation of a graph representing the semantic neighbourhood of the original word. In previous work, a set of associations between the hypothesized Arabic words and English synsets was proposed on the basis of this graph. Here, a novel approach to extending AWN is presented whereby a Bayesian Network is automatically built from the graph and then the net is used as an inferencing mechanism for scoring the set of candidate associations. Both on its own and in combination with the previous technique, this new approach has led to improved results.

1. Introduction

An Arabic WordNet (AWN – Black, et al., 2006; Elkateb, et al., 2006; Rodríguez et al., 2008) has been built along the last two 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 – Fellbaum, 1998). In addition, the AWN project aimed to provide a formal specification of the senses of its synsets using the Suggested Upper Merged Ontology (SUMO – Niles & Pease, 2001). This representation serves as an interlingua among all wordnets (Pease, 2003; Vossen, 2004) and will underlie the development of semantics-based computational tools for multilingual NLP.

Accordingly with the objectives of the project, Arabic WordNet currently consists of 11,270 synsets (7,961 nominal, 2,538 verbal, 661 adjectival, and 110 adverbial), containing 23,496 Arabic expressions. This number includes 1,142 synsets that correspond to named entities which have been extracted automatically and are being checked by the lexicographers. For the most up-to-date statistics see:
http://www.lsi.upc.edu/~mbertran/arabic/awn/query/sug_statistics.php

2. Prior Work on the Semi-automatic Extension of AWN

Although AWN has been constructed manually, efforts are underway to partially automate the process using bilingual lexical resources and applying morphological rules. Here we are especially interested in reducing the development effort by reducing the number of decisions made in regard to accepting or rejecting a proposed extension. Using lexical resources for extending wordnets for languages other than English is not new. In some cases a substantial part of the work has been performed automatically, using PWN as source ontology in combination with bilingual and monolingual resources for proposing correlates (e.g., Benítez, et al., 1998; Agirre, et al., 2002; Chen, et al., 2002; Miháltz & Prószéky, 2003; Barbu & Barbu-Mititelu, 2005; and Farreres, 2005).

We have investigated two general approaches which take advantage of an important characteristic of Arabic (and other Semitic languages), the fact that words sharing a common root (i.e. sequence of almost always three consonants) usually have related meanings and can be derived from a common base verbal form by means of a reduced set of lexical rules.

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 مُدَرِّسَة (MuDaRRiSatun, female teacher).

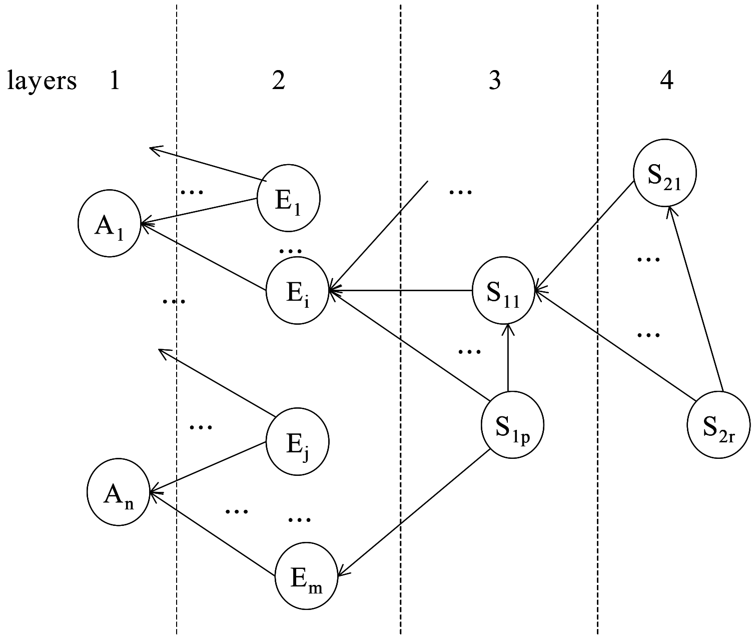

Figure 1: Topology of the BN

Both approaches coincide on deriving new Arabic word forms from existing Arabic verbal synsets and then producing a list of suggested English synsets for each form. To generate the Arabic word forms we use a very limited but highly productive set of lexical rules which produce regular verbal, nominal and adjectival derivative forms as well as inflected forms. This process is presented in (Rodríguez, et al., 2008).

The central problems to be faced are, on the one hand, filtering noise caused by over generation of Arabic word forms (obviously, the application of the whole set of lexical rules to a given form results on a severe over generation of Arabic forms) and, on the other, mapping the newly created forms to appropriate PWN synsets. To deal with the filtering problem, we implemented decision tree classifiers using the C5.0 implementation in Weka toolbox (Whitten & Frank, 2005). Details are reported in (Rodríguez, et al., 2008). To associate these words with PWN synsets, we translated the Arabic words (layer 1, A_i in Figure 1) into English words (layer 2, E_i) and identified all the synsets these translations belonged to (layer 3, S_{1i}), thus producing a set of <Arabic word, English word, PWN synset> tuples. Further more we looked for an additional layer of PWN synsets (layer 4, S_{1i}) which were associated with the synsets in layer 3 by way of one or another semantic relation. In this way an undirected graph (having the same topology than the Bayesian network in Figure 1) was build.

The first approach is based on a set of heuristics that use the graph structure directly for measuring the reliability of each <Arabic word, English synset> association, while the second, more complex, maps the graph onto a Bayesian Network and applies a learning algorithm. Details on the heuristics, which range from precision oriented to recall oriented, can be found in (Rodríguez, et al., 2008). For the sake of clarity, we include the first two:

- 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 tagged as 1.
- If multiple paths A-E1-S and A-E2-S exist (i.e., A is translated as E1 or E2 and both E1 and E2 are associated with S among other possible associations) then the output tuple <A, S> is tagged as 2.

The results of a preliminary evaluation on 10 randomly selected AWN verbs were encouraging. Overall, of 272 Arabic word-PWN synset associations proposed by all the heuristics, 135 (49.6%) were judged correct by the lexicographers, and of the 61 Arabic word-PWN synset associations proposed by the two most reliable heuristics, 40 (65.6%) were judged to correct.

3. The semi-automatic extension of AWN using Bayesian Nets

The second approach to proposing likely Arabic word-English synset associations starts by mapping the same generated graph into a Bayesian Network (BN). Then the candidates are scored using the network. The goal of this approach is increasing the coverage of the proposed associations.

In addition to AWN and PWN, these procedures make use of:

- the LOGOS database of conjugated Arabic verbs[1],

[1] See: http://www.logosconjugator.org/verbi_utf8/all_verbs_index_ar.html

- the NMSU bilingual Arabic-English lexicon[2],
- the Arabic Gigaword Corpus (available through LDC)[3],
- the UN (2000-2002) bilingual Arabic-English Corpus (also available through LDC: catalog # LDC2004E13)[4].

Each BN is organized into four layers (see Figure 1):

1) Arabic words (AW),
2) English words (EW),
3) PWN synsets linked to layer 2 (S_1),
4) other PWN synsets linked to layer 3 (S_2).

The process of building a BN starts with the undirected graph used for the first approach. Directionality is added to the edges of the graph, cycles are avoided and conditional probabilities are attached to the nodes through the corresponding Conditional Probability Tables (CPTs).

One BN is built for each of the 2296 base verb forms in AWN. Nodes in AW are the Arabic words generated by the lexical rules and filtered by the decision tree. Nodes in EW correspond to their English translations (using the NMSU lexicon). The nodes in S_1 correspond to the PWN English synsets containing the English translations. The nodes in S_2, in turn, are connected to the synsets in S_1 on the basis of PWN semantic relations. The creation of edges uses a greedy approach for avoiding cycles. Synset nodes are sorted by number of output edges and edges are added one at a time so long as no cycle is produced. In order to limit the combinatorial cost, we have applied a threshold of 10 on the maximum number of parents for any given node.

The CPT is computed as follows. For edges EW -> AW we use probabilities from statistical translation models built from the UN corpus using GIZA++ (word-word probabilities). The original translation models are filtered to avoid pairs having Arabic expressions with invalid Buckwalter encodings.

For a node in EW the following associations are possible:

- AL+BN (aligned by the model and in the BN),
- AL-BN (aligned by the model but not in the BN),
- BN-AL (in the BN but not aligned by the model).

However, we only consider associations in BN (i.e., AL+BN and BN-AL), and thus distribute the mass probability between these two cases, summing to 1. The conditional priors of the EW -> AW edges are computed as (1):

$$p_{BN}(a \mid e) = \begin{cases} \mathit{if}\ \mathrm{a} \in AL + BN & p_{AL}(a \mid e) \cdot \beta / \alpha \\ \mathit{if}\ \mathrm{a} \notin AL + BN & (1 - \beta) / \left| BN - AL \right| \end{cases}$$

where parameters α and β are defined in (2) and (3).

$$\alpha = \frac{\sum\limits_{a \in AL+BN} p(a \mid e)}{\sum\limits_{a \in AL+BN} p(a \mid e) + \sum\limits_{a \in AL} p(a \mid e)} \qquad (2)$$

$$\beta = \frac{\left| AL + BN \right|}{\left| AL + BN \right| + \left| BN - AL \right|} \qquad (3)$$

From these priors we computed the CPT using a noisy-or approach. For the S_1 -> EW and S_i -> S_j conditional probabilities, we used a linear distribution of the mass probability of each node between its descendents. The CPT for each node in EW, S_1 and S_2 is computed in the same way. Although the way of computing the CPT for nodes in layers 3 and 4 is quite straightforward, the intended use, improving the score of highly connected synsets, is achieved.

For each BN, a Bayesian inference was performed, using the nodes in AW as evidence (i.e, assigning probability 1 to nodes in layer 1) and looking for the probabilities of all the synsets in S_1. The set of candidates is built with tuples <X,Y> where X belongs to AW, Y belongs to S_1 and has a non null probability, and there is a path from X to Y. The tuple is scored with the posterior probability of Y given the evidence provided by the net. Only the tuples scored over a threshold are selected for inclusion in the final set of candidates.

4. Evaluation

For the evaluation we used the same set of 11 Arabic verbs as in previous experiments (1st approach). The sizes of the BNs are shown in Table 1.

Arabic verb	# English Words	# Synsets ($S_1 \cup S_2$)
عَامَلَ	107	190
أَعْقَبَ	71	77
صَقَلَ	31	21
رَتَّبَ	62	102
أَخَّرَ	19	9
أَخْبَرَ	80	105
رَشَّحَ	40	22
غَامَرَ	56	49
أَشْبَعَ	38	34
أَخْرَجَ	85	140
دَرَّسَ	57	51

Table 1: Size of the BN

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

[3] See: http://www.ldc.upenn.edu/Catalog/CatalogEntry.jsp?catalogId=LDC2006T02

[4] See: http://projects.ldc.upenn.edu/TIDES/index.html

Selection	Threshold	candidates	accept	reject	precision	recall	F_1
HEU	all heuristics	272	135	137	0.50	0.61	0.55
HEU	heuristics 1,2	61	40	21	0.65	0.18	0.28
BN	0	554	223	331	0.40	**1**	0.57
BN	0.01	243	125	118	0.51	0.56	0.53
BN	0.02	214	116	98	0.54	0.52	0.53
BN	0.07	112	65	47	0.58	0.29	0.39
BN	0.1	100	60	40	0.60	0.27	0.37
BN + HEU	0	272	154	118	0.56	**0.69**	**0.62**
BN + HEU	0.01	212	121	91	0.57	0.54	0.55
BN + HEU	0.02	201	115	86	0.57	0.41	0.48
BN + HEU	0.07	92	65	27	**0.71**	0.38	0.5
BN + HEU	0.1	83	59	24	0.71	0.12	0,21

Table 2: Results

Results are presented in Table 2. Figures are included on the number of candidates proposed by each method, the number of accepted and rejected ones by human validation and the usual precision, recall and F_1 measures. As can be seen, the BN approach doubles the number of candidates of the previous HEU approach (554 vs. 272). The number of accepted candidates for the first case (223) is used as upper bound for recall calculation. We have used a balanced F_1 although we are more interested on a getting a higher recall. Rows 1 and 2 show the results of our previous heuristics-based approach. The next 5 rows present the BN approach using different thresholds. Finally we present the results of intersecting both methods. The highest precision is obtained using the intersection method and restrictive threshold. Although recall is low, the number of candidates for AWN is high (92 words from the original 11 base forms). An analysis of the errors shows a substantial number were due to the lack of the shadda diacritic or the feminine form. Fixing these would increase the accepted forms to 81 from 60 for a threshold of 0.1 and the precision to 0.67 from 0.6, a 10% improvement.

5. Conclusion

A novel approach for semi-automatically extending AWN's coverage using Bayesian Networks has been presented. The approach takes profit of some characteristics of Arabic language that allow an easy development of a limited set of highly productive lexical rules for deriving from a verbal entry a set of semantically related word forms and extends a previous approach based on the performance of a set of heuristics. Initial experiments using both procedures and their combination show promising results.

6. Acknowledgements

This work was supported by the US Central Intelligence Agency.

7. References

Agirre, E., Ansa, O., Arregi, X., Arriola, J., Diáz de Ilarraza, A., Pociello, E., Uria, L., (2002). Methodological issues in the building of the Basque WordNet: Quantitative and qualitative analysis. In *Proceedings of the First International WordNet Conference*, 21-25 January 2002. Mysore, India.

Barbu, E., and Barbu-Mititelu, V., (2005). A case study in automatic building of wordnets. In *Proceedings of OntoLex 2005 - Ontologies and Lexical Resources*.

Benítez, L., Cervell, S., Escudero, G., López, M., Rigau, G., Taulé, M., (1998). Methods and tools for building the Catalan WordNet. In *Proceedings of LREC Workshop on Language Resources for European Minority Languages*.

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).

Chen, H., Lin, C., and Lin, W., (2002). Building a Chinese-English WordNet for translingual applications. *ACM Transactions on Asian Language Information Processing* 1 (2), 103–122.

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.

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

Farreres, J. (2005) *Creation of wide-coverage domain-independent ontologies*. PhD Thesis, Departament de Llenguatges i Sistemes Informàtics, Univertitat Politècnica de Catalunya, Barcelona, Spain.

Miháltz, M., and Prószéky, G., (2003). Results and evaluation of Hungarian nominal WordNet. In *Proceedings of the Second International WordNet Conference*. Masaryk University, Brno, pp. 175–180.

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

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.

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.

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.

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