Introducing the Arabic WordNet Project

William BLACK, Sabri ELKATEB,
School of Informatics
University of Manchester
Sackville Street, Manchester, M60 1QD,
w.black@manchester.ac.uk, sabrikom@hotmail.com
Horacio RODRIGUEZ, Musa ALKHALIFA
Politechnical University of Catalonia
University of Barcelona
horacio@lsi.upc.edu, musa@thera-clic.com

Piek VOSSEN
Irion Technologies
piek.vossen@irion.nl
Adam PEASE
Articulate Software
apease@articulatesoftware.com
Christiane FELLBAUM
Princeton University
fellbaum@clarity.princeton.edu

Abstract

Arabic is the official language of hundreds of millions of people in twenty Middle East and northern African countries, and is the religious language of all Muslims of various ethnicities around the world. Surprisingly little has been done in the field of computerised language and lexical resources. It is therefore motivating to develop an Arabic (WordNet) lexical resource that discovers the richness of Arabic as described in Elkateb (2005). This paper describes our approach towards building a lexical resource in Standard Arabic. Arabic WordNet (AWN) will be based on the design and contents of the universally accepted Princeton WordNet (PWN) and will be mappable straightforwardly onto PWN 2.0 and EuroWordNet (EWN), enabling translation on the lexical level to English and dozens of other languages. Several tools specific to this task will be developed. AWN will be a linguistic resource with a deep formal semantic foundation. Besides the standard wordnet representation of senses, word meanings are defined with a machine understandable semantics in first order logic. The basis for this semantics is the Suggested Upper Merged Ontology (SUMO) and its associated domain ontologies. We will greatly extend the ontology and its set of mappings to provide formal terms and definitions equivalent to each synset.

Introduction

AWN will be constructed according to the methods developed for EuroWordNet (EWN; Vossen 1998) and since applied to dozens of languages around the world. The EuroWordNet approach maximizes compatibility across wordnets and focuses on manual encoding of the most complicated and important concepts.

Arabic word		English word	Category
walada	وَلَدَ	give birth	v
wila:dah	وِلاَدَة	birth	n
wallada	وَلَّدَ	generate	v
tawli:di	تَوْلِيدِي	generative	a
tawa:lada	تَوَالَدَ	breed	v
tawa:lud	تَوَالُد	reproduction	n
wa:lidah	وَالِدَة	female parent	n
wa:lid	وَالِد	male parent	n
walad	وَلَد	infant	n
wali:d	وَلِيد	born baby	n
wila:dah	وِلاَدَة	delivery	n
wallada	وَلَّدَ	deliver/assist in birth	v
mawlid	مَوْلِد	Prophet's birthday	n
mila:d	مِيلاَد	The Nativity	n
mi:la:di	مِيلاَدِي	A.D., (anno domini)	n

Table 1 Word forms and semantic relations generated by the Arabic root 'w l d'

Language-specific concepts and relations are encoded as needed or desired. This results in a so-called core wordnet for Arabic with the most important synsets, embedded in a solid semantic framework. From this core wordnet, it is possible to automatically extend the

coverage with high precision. Specific concepts can be linked and translated with great accuracy because the base building blocks are manually defined and translated. The approach follows a top-down procedure. Arabic Base Concepts are defined and extended via hyponymic relations to derive a core wordnet. The set of Common Base Concepts (CBCs) from the 12 languages in EWN and BalkaNet (Tufis 2004) are encoded as synsets; other language-specific concepts are added and translated manually to the closest synset(s) in Arabic. The same step is performed for all English synsets that currently have an equivalence relation in SUMO.

Next , the first layers of hyponyms are chosen on the basis of linguistic and applications-based criteria; the final phase completes the target set of concepts/synsets, including specific domains and named entities. Each synset construction step is followed by a validation phase, where formal consistency is checked and the coverage is evaluated in terms of frequency of occurrence and domain distribution.

Tools to be developed for AWN include a lexicographer's interface modeled on that used for EWN, with added facilites for Arabic script, following Black and Elkateb's earlier work (2004). A large ontology to provide a semantic foundation for AWN will be built on the basis of the present SUMO (Niles and Pease, 2001).

Structure and organization of AWN

Because AWN is to be aligned not just to PWN (Fellbaum 1998) but to every wordnet aligned to PWN--either directly or indirectly through an (interlingual index (ILI) or ontology--the database design supports multiple languages, and the user interface will be explicitly multilingual rather than bilingual as was the one described in Black and Elkateb (2004).

The database structure comprises four principal entity types, *item*, *word*, *form* and *link*.

Items are conceptual entities, including synsets, ontology classes and instances. Besides a unique identifier, an item has descriptive information such as a gloss. Items lexicalized in different languages are distinct.

A *word* entity is a word sense, where the citation form of the word is associated with an item via its identifier.

A *form* is a special form that is considered dictionary information (not merely an inflectional variant). The forms of Arabic words that go in this table are the root and/or the broken plural form, where applicable.

A *link* relates two items, and has a *type* such as "equivalence," "subsuming," etc. Links connect sense items to other sense items, e.g. a PWN synset to an AWN synset, a synset to a SUMO concept, etc.

This data model has been specified in XML as an interchange format, but is also implemented in a MySQL database hosted by one of the partners. The database will be the primary deliverable of the project, and will be distributed freely to the community.

Constructing AWN

The basic criteria for selecting synsets to be covered in AWN are:

- Connectivity: AWN should be as densely connected as possible by hyperonymy/ hyponymy chains, etc. Most of the synsets of AWN should correspond to English WN counterparts and the overall topology of both wordnets should be similar.
- Relevance: Frequent and salient concepts have priority. Criteria will include the frequency of lexical items (both in Arabic and English) and the frequency of Arabic roots in their respective reference corpora.
- Generality: Synsets on the highest levels of WN are preferred.

These criteria suggest two ways for proceeding:

- From English to Arabic: Given an English

synset, all corresponding Arabic variants (if any) will be selected.

- From Arabic to English: Given an Arabic word, all its senses have to be found, and for each of these senses the corresponding English synsets have to be selected.

Both steps have to be followed throughout the construction of AWN.

All AWN synsets must be manual validated (and eventually locked, when all their variants have been found) but advantage should be taken as much as possible of the available resources for guiding the construction and validation process.

Once a new Arabic verb is added toAWN, several possibilities for extension arise: extensions from verbal entries, including verbal derivates, nominalizations, verbal nouns, etc. We also consider the most productive forms of deriving broken plurals. This can be done using a set of lexical and morphological rules. To take full advantage of these extensions short iterations will be performed.

As stated in the introduction, the starting point of AWN is the manual construction of its Base Concept (BC) set from EWN and BalkaNet's CBCs. We concentrate on the most relevant terms for obtaining about 1,000 nominal and 500 verbal synsets.

The second step consists of the top-down vertical extension of BC, following Farreres 2005, Diab 2004). Some pre-processing is required for this and the next phase. We mention two tasks, preparation and extension.

Preparation includes the processing of the available bilingual resources and the compilation of a set of lexical and morphological rules. From the set of available bilingual dictionaries we construct a homogeneous bilingual dictionary (HBIL) that contains for each entry information on the Arabic/English word pair, the Arabic root (added manually), POS, relative frequencies and sources supporting the pairing.

The set of 17 heuristic methods used in the development of EWN will be applied to HBIL (following Farreres 2005) to derive candidate Arabic words/English synsets mappings. For each mapping the information attached includes the Arabic word and root, the English synset, POS, relative frequencies, mapping score, absolute depth in WN, number of gaps between the snset and the top iof the WN hierarchy, and sources containing the pair.

Arabic words in bilingual resources must be normalized and lemmatized (Diab et al. 2004, Habash and Rambow 2005) but vowels and diacritics must be maintained. Arabic roots are not vowelized.

Following pre-processing, the set of scored Arabic word/English synset pairs becomes the input to the manual validation step. We proceed by chunks of related units (sets of related WN synsets, e.g. hyponymy chains and sets of related Arabic words, i.e., words having the same root) instead of individual units (synsets, senses, words).

Finally, AWN will be completed by filling gaps in its structure, covering specific domains, adding terminology and named entities, etc.

The User Interface

In addition to search and browsing facilities for the end users of the completed database, lexicographers require an editing interface. A variety of legacy components are available, each with their relative advantages. We chose to adapt the one described in Black and Elkateb (2004), because it can handle Arabic script. However, it assumed an entirely different data model, in which the Arabic words were directly linked to offsets representing PWN synsets. It was also organized to support browsing and searching in the synset space entirely in English and merely required word-synset mappings for Arabic to be added. The new interface will

attempt to put both languages on an equal footing and indeed to be indifferent as to tje direction of alignment between the two languages' conceptual structures.

The editor's interface will moreover communicate with the database server using SOAP (Simple Object Access Protocol). This is to allow multiple lexicographers at different sites to maintain a common database.

Ontology

The AWN project will provide a deep semantic underpinning for each concept. We take the approach that was previously used in mapping all of PWN to a formal ontology (Niles & Pease, 2003), the Suggested Upper Merged Ontology (Niles & Pease, 2001).

Synsets map to a general SUMO term or a term that is directly equivalent to the given synset (Figure 1). New formal terms will be defined to cover a greater number of equivalence mappings, and the definitions of the new terms will in turn depend upon existing fundamental concepts in SUMO. The process of formalizing definitions will generate feedback as to whether word senses in WN need to be divided or combined and how the glosses may be clarified. Since many wordnets in other languages are already linked by synset number, this work will benefit wordnets in other languages as well.

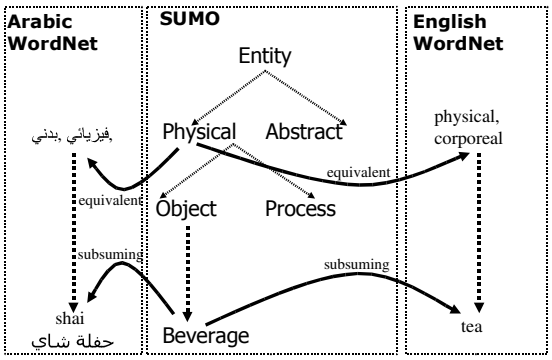

Figure 1: SUMO mapping to wordnets

The Suggested Upper Merged Ontology (SUMO) (Pease&Niles 2002, Niles&Pease 2001) is a freely available, formal ontology of about 1000 terms and 4000 definitional statements. It is provided in a first order logic language called Standard Upper Ontology Knowledge Interchange format (SUO-KIF) (Pease, 2000), and also translated into the OWL semantic web language. It is now in its 73rd version; having undergone four years of development, review by a community of hundreds of people, and application in expert reasoning and linguistics. SUMO has been subjected to formal verification with an automated theorem prover. SUMO has been extended with a number of domain ontologies, which are also public, that together number some 20,000 terms and 60,000 axioms. SUMO has been mapped by hand to the WN lexicon of 100,000 noun, verb, adjective and adverb senses, which not only acts as a check on coverage and completeness, but also provides a basis for application to natural language understanding tasks. SUMO covers areas of knowledge such as temporal and spatial representation, units and measures, processes, events, actions, and obligations. Domain specific ontologies extend and reuse SUMO in the areas of finance and investment, country almanac information, terrain modeling, distributed computing, endangered languages description, biological viruses, engineering devices, weather and a number of military applications. It is important to note that each of these ontologies employs rules. These formal descriptions make explicit the meaning of each of the terms in the ontology, unlike a simple taxonomy, or controlled keyword list. SUMO is the only formal ontology that has been mapped to all of WN, and the only formal upper ontology that has been extended with a number of domain ontologies that are also open source. SUMO has natural language generation templates and a multi-lingual lexicon that allows statements in SUO-KIF and SUMO to be expressed in multiple natural languages. These include English, German, Czech, Italian, Hindi (Western character set) and Chinese (traditional characters and pinyin). Automatic translations can be viewed on line at http://virtual.cvut.cz/kifb/en/. An additional

part of our work will be updating the Sigma ontology development environment (Pease, 2003) to handle a similar presentation of Unicode-based character sets, including Arabic.

The ontology as a structured ILI

The comprehensive mapping and definition of synsets in AWN to SUMO concepts opens a new perspective on the role of the Interlingual Index (ILI) in connecting wordnets. As discussed in Vossen et al. (1999) and Vossen (2004), it is not necessary that the ILI be the superset of all concepts that occur in all wordnets. In fact, it is argued that the ILI could be a condensed set of more or less universal concepts. Such a condensed set of concepts can still precisely link synsets across languages through multiple equivalence relations that are exhaustive. For example, the Spanish synsets *alevín* ('young fish') and *cajera* ('female cashier') do not have a direct equivalent in the English WN that is currently used as an ILI. This is solved by mapping these synsets to both *fish* and *young* in one case and to both *cashier* and *female* in the other case by means of hyponym y and a property relation, respectively. As long as we indicate that the Spanish synsets are exhaustively defined by these relations, we can find equivalent sin another language, such as the Dutch synset containing *caissière* ("female cashier"), assuming that it is also exhaustively linked to the same concepts with the same relations.

In the AWN project, we want to take this idea a step further. If both AWN and English WN synsets are exhaustively defined in terms of SUMO concepts, SUMO can in effect become the ILI for wordnets. This means that SUMO not only maps word meanings and synonyms across languages but also provides a formal semantic framework for all these languages.

If we return the example of *shai* discussed above, we can say that an exhaustive definition of the concept with a number or relations to SUMO concepts (sr1, sr2, ..., srn), can function as an ILI relation when the English synset for *tea* that currently acts as the ILI concept is also exhaustively linked to the same SUMO concepts with the same relations. Corresponding synsets in other wordnets, such as Dutch *thee* and Spanish *té* that are presently linked to the English ILI can then inherit the SUMO relations functioning as an ILI as well as a formal definition.

The development of AWN will include a transition phase where AWN synsets are both linked to the English WN serving as an ILI and exhaustively defined with SUMO. This is shown in Figure-2 below.

Here English *tea* only has a subsumption relation (sr1) with SUMO Beverage.

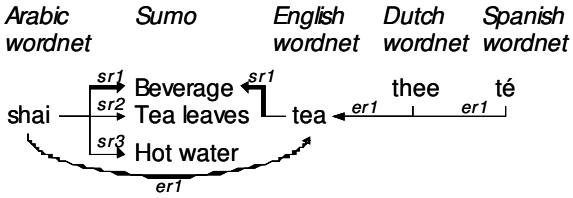

Figure 2 : SUMO and ILI

However, the equivalence relation (er1) between "shai" and "tea" can be used to upgrade the SUMO definition for English. Consequently, SUMO can replace the ILI altogether and be applied to the other languages. Obviously, it remains a topic for future research to determine to which extent this process can be completed.

Conclusion

Constructing AWN presents challenges not encountered by established wordnets. These include the script on the one hand and the morphological properties of Semitic languages, centered around roots, on the other hand. The foundations for meeting these challenges have been laid. An innovation with significant consequences for wordnet development is the proposal to substitute English WN as the ILI with SUMO.

Acknowledgements

This work was supported by the United States

Central Intelligence Agency. We are grateful to Mona Diab (Columbia University) for helpful comments.

References

Black, W. J. & Elkateb, S. (2004) A Prototype English-Arabic Dictionary Based on WordNet, Proceedings of 2nd Global WordNet Conference, GWC2004, Czech Republic: 67-74

De Roeck, A. and Al-Fares, W. (2000) A Morphologically Sensitive Clustering Algorithm for Identifying Arabic Roots, Proceedings of the 38th Annual Meeting of the ACL, Hong Kong. 199-206

Diab, M. (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.

Diab, M., Hacioglu, K. and D. Jurafsky (2004). Automatic Tagging of Arabic Text: From raw text to Base Phrase Chunks. Proceedings of HLT-NAACL .

Elkateb, S. (2005) Design and implementation of an English Arabic dictionary/editor, PhD thesis, Manchester University.

Farreres, J. (2005) Creation of wide-coverage domain-independent ontologies. PhD thesis, Univertitat Politècnicade Catalunya.

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

Habash, N. and O. Rambow (2005.) Arabic Tokenization, Part-of-Speech Tagging and Morphological Disambiguation in One Fell Swoop. In Proceedings of the 43rd Meeting of the Association for Computational Linguistics (ACL).

Niles, I & Pease A. (2001). "Towards A Standard Upper Ontology." In Proceedings of FOIS 2001, October 17-19, 2001, Ogunquit, Maine, USA. See also http://www.ontologyportal.org

Niles, I., and Pease, A., (2003). Linking Lexicons and Ontologies: Mapping WordNet to the Suggested Upper Merged Ontology, Proceedings of the IEEE International Conference on Information and Knowledge Engineering, pp 412-416.

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

Pease, A., (2000). Standard Upper Ontology Knowledge Interchange Format. Web document http://suo.ieee.org/suo-kif.html.

Pease, A., (2003). The Sigma Ontology Development Environment, in Working Notes of the IJCAI-2003 Workshop on Ontology and Distributed Systems. Volume 71 of CEUR Workshop Proceeding series.

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

Vossen, P. (ed) (1998) EuroWordNet: A Multilingual Database with Lexical Semantic Networks, Kluwer Academic Publishers, Dordrecht.

Vossen, P. Peters, W., J. Gonzalo. (1999). 'Towards a Universal Index of Meaning'. Proceedings of the ACL-99 Siglex workshop, University of Maryland, 81-90

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, pp 161-173

Vossen, P. (1998, ed.). EuroWordNet: A Multilingual Database with Lexical Semantic Networks. Dordrecht, Holland: Kluwer.