[*AMTA/SIG-IL First Workshop on Interlinguas*, San Diego, CA, October 28, 1997]

Improving the Precision of Lexicon-to-Ontology Alignment Algorithms

Latifur R. Khan **Eduard H. Hovy**

Department of Computer Science **Information Sciences Institute (ISI)**

University of Southern California **University of Southern California**

lkhan@usc.edu **hovy@isi.edu**

This paper describes an automatic technique for improving the association of two combined machine-readable Arabic-English lexicons with a semantic word thesaurus called WordNet (version 1.5). Our main goal is to attach individual Arabic word meanings to appropriate WN concepts. This paper describes the automatic methods used to achieve this goal for nouns. When searching WordNet for a concept, we performed name matching and definition matching simultaneously. Unfortunately, these matches provided too many suggested alignments. Our main contribution is the reduction in number of matches by applying a new filtering heuristic.

1. Introduction

This paper describes an automatic technique for associating words in a machine-readable Arabic-English lexicon with appropriate synsets from the concept thesaurus WordNet 1.5. This association enables knowledge-based machine translation that uses WordNet synsets as pivot terms between languages (that is, as approximations to interlingua symbols).

The major problem is that associating WordNet items with lexical items is not very easily automated and generally requires human supervision. In order to decrease the amount of human effort required, we implemented variations of two algorithms known as name match and definition match. When applied to nouns, the algorithms did successfully suggest good matches for the Arabic words; unfortunately, however, they tended to yield too many suggestions. To reduce the number of matches, we devised a new algorithm that combines both algorithms together with a new pruning heuristic.

This paper is organized as follows: Section 2 describes the linguistic resources, Section 3 describes the merging algorithm, and Section 4 contains a discussion.

2. Linguistic Resources

2.1 Arabic Lexicon

We bought an Arabic-English dictionary, available commercially on CD-ROM. We extracted a portion of the contents of this dictionary to create a machine-readable lexicon. We then merged this lexicon with ALPNET, another machine-readable lexicon developed under US Government funding and later extended by MT researchers at the Computer Science Department of the University of Maryland, College Park.

The merged A-E lexicon contains 36,637 noun citation forms, 23,266 verb citation forms, and 14,200 other (preposition, adjective, conjunction, etc.) citation forms. Each citation form includes at least one Arabic romanized entry. Each romanized entry contains at least one English translation (a word or phrase), one domain indication (e.g., BUSINESS), and one part of speech. In addition, each entry may contain a vocalized form (since Arabic is commonly written without vowels) and a root. Our Arabic nouns are categorized into one of the following parts of speech: simple noun, derived noun, proper noun, and plural noun. No Arabic or English definitions were available for the Arabic word entries.

2.2.WordNet

WordNet (Miller 1990) is a concept thesaurus based on psycholinguistic principles. WordNet contains only nouns, verbs, adjectives, and adverbs. For our work we considered only nouns. WordNet contains 60,557 nominal concepts organized into distinct senses called *synsets.*

Concepts support the superordinate relation (called *hyponymy)* which generates a hierarchical semantic organization. We used the WordNet notation in which the symbol *'x>y'* means *x* is a superconcept of *y* and the symbol *'x<y'* means *x* is a subconcept of *y.* For example, *vicereine<wife* means the concept *vicereine* is a specialization of the concept *wife.* The noun hierarchy has average depth of about 12. Although other interconcept relations are defined in WordNet and could potentially be used to suggest matches, we employed only hypernymy-hyponymy. Approx. 50% of WordNet concepts have brief informal English definitions.

3. Merging Algorithm

Our work had two major purposes, both in support of Arabic-to-English machine translation by the GAZELLE system being developed at USC/ISI (Knight et al. 95):

1. to construct a single Arabic lexicon out of various sources;

2. to construct links from Arabic words to the symbols that act as pivot terms in the system's Interlingua-based approach.

For the latter purpose, we had to associate Arabic word entries to appropriate WordNet synsets, which are in turn linked to associated English words, and whose taxonomic organization has been altered to support linguistic processing.

Manual association is time-consuming and expensive. So we turned to automated alignment algorithms, looking first at the groundbreaking work of (Knight and Luk 1994), and at the similar work of (Rigau and Agirre 1995, Ageno et al. 1994, Okumura and Hovy 1994, Agirre et al. 1994). Unfortunately, since the Arabic entries did not contain definitions, we could not employ the definition match commonly used. This match relies on the principle that two wordsenses should be matched if their definitions share enough words. However, we were able to use a weaker form of this match, considering simple the lexicon's English translations (words and/or phrases) instead of definitions. Each Arabic entry contained at least one such translation.

Still, since about 50% of the WordNet synsets do not have definitions, and our variant of the definition match algorithm produced a large number of matches, many of the proposed matches were not accurate. We therefore implemented a combined mergin+pruning method. This method involves the following three steps:

Step 1: (Name match): English translations of each Arabic word are used to retrieve synsets of WordNet. These synsets are added into the *candidate list,* which is initially empty. Given wordsense ambiguity, there were too many such candidates. To reduce the number of matches, we applied step 2.

Step 2: (Definition match): Remove a concept from the candidate list. If that concept contains a definition, we determine whether the definition shares at least one of the translation words of the Arabic entry. If so, that concept is removed from the candidate list and added into the *drop list.* Otherwise, or if the concept of the candidate list did not have a definition, then it is replaced back into the candidate list. Our belief is that it is a good candidate. (This step involves demorphing each word of the translation and of the WordNet definitions, and of removing common and closed-class words—a time-consuming operation that we performed once, initially.)

Step 3: (Concept generalization): In the drop list, we have collected sets of concepts that all contained the concept's name itself. For these candidates, we attempted to find a single WordNet parent (generalization concept). If successful, we placed this generalization back onto the candidate list. We motivate and describe this step at the end of this section.

The pseudo-code in Figure 1 illustrates the whole process.

```
Input: Arabic lexicon, WordNet 1.5
Output: a list of pairs: (Arabic entry (list of WordNet synsets))
Algorithm:
    for each English translation $x$ of an Arabic entry
            for each concept $y$ of WordNet,
                if $x$ is matched with $y$ by the name match
                    then add $y$ into the candidate list.
            for each concept $z$ of the candidate list,
                if $z$ contains a definition and if the name $x$ is contained in $z$
                    then add $z$ into drop list and remove it from the candidate list.
            for each concept $w$ in the drop list,
                if the drop list is not empty
                    collect all concepts $p,q,..$ from the drop list with the same parent as w
```

```
if this hypernym fulfills the name and definition match,
    place it on the candidate list
    remove p, q,... from the drop list.
```

Figure 1. Pseudo-code for merging algorithm.

We now provide a concrete example to illustrate how the algorithm associates Arabic nouns with WordNet synsets. The first 4 lines below are part of the Arabic lexicon, and the remaining 7 are generated by the name match and definition match algorithms (Steps 1 and 2):

```
(|!'=hl !lrjl|
  (vocalized-head  "!'a=h#lu !lr:ajul")
  (translations  "wife")
  (domain  "Relations")
  ("battle-ax<wife" "NOUN.PERSON" "a sharp-tongued domineering wife"
          "wife" 1 1)
  ("crown princess<wife" "NOUN.PERSON" "the wife of a crown prince"
          "wife" 1 1)
  ("first lady<wife" "NOUN.PERSON" "the wife of a chief executive"
          "wife" 1 1)
  ("wife" "NOUN.PERSON" "a married woman; a man's partner in marriage"
          "adult female" 1 0)
  ("matron<wife" "NOUN.PERSON" "a married woman (usually middle-aged
          with children) who is staid and dignified " "wife" 1 0)
  ("vicereine<wife" "NOUN.PERSON" "wife of a viceroy " "wife" 1 1)
  ("viscountess<wife" "NOUN.PERSON" "a wife or widow of a viscount"
          "wife" 11))
```

In the first line, the entry head is given (written in ISI's internal romanization of Arabic). The second line contain its vocalized form (Arabic is usually written without vowels. The resulting increase in ambiguity makes this entry a necessity in the lexicon). The third line contains the lexicon's English translation of the head word, and the fourth contains its semantic domain.

Each remaining line has the following structure: the first field denotes the proposed WordNet synset (battle-ax<wife), the second field (NOUN.PERSON) contains the WordNet type, the third field contains the WordNet definition, if any ("a sharp-tongued domineering wife"), the fourth field contains the WordNet superconcept (wife), the fifth field contains the name match status (1 or 0), and the sixth field contains the definition match status (also 1 or 0). Here 1 means the name or definition match was successful, and 0 means no match for that type was found.

In WordNet, wife is the generalized concept of the synsets battle-ax, crown-princess, first lady, matron, vicereine and viscountess, as found in Step 3. In this case the subconcepts all passed both matches (scoring 1,1). So they are moved to the drop list. On the other hand, the definition of the generalized concept wife does not contain the word *"wife",* avoiding a circular definition, and hence has the score (1,0). So the synset *wife* is the best candidate for a match after applying Step 3. It is replaced onto the candidate list.

4. Discussion

The pruning enhancement (step 3) helped a great deal. The Arabic entries for which WordNet matches were found tended to have around two or three suggestions. Of these, we found that the algorithm yielded approximately 69% correct associations for the first 200 Arabic lexicon entries. We applied no further pruning heuristics, although several more are possible.

The main contribution of our algorithm is that it prunes well. In this sense, it is an extension of the class of algorithms introduced by (Knight and Luk 1994).

After applying our algorithm, we found 1:1, l:many, and 1:0 associations between the Arabic lexicon and WordNet. We found only a small number of 1:0 associations, usually when an Arabic entry's translation contained a long (phrase) translation. In such cases one can associate with WordNet only by using the definition match over all remaining unattached WordNet synsets. It is an open issue how useful this match will be.

One possible improvement is to use the domain of the Arabic entry in an additional pruning step. The lexicon contains 69 domain names. By applying the name and definition matches to domain names, we can find their candidate concepts in WordNet. We can then manually associate correct synsets(s) with each domain and use this together with some kind of taxonomy-inheritance match.

Unfortunately, associating verbs and other parts of speech is more difficult than nouns. This is due to the much shallower structure of WordNet for verbs (on average, only about 3 levels) and the non-hierarchical star-like structure for modifiers.

Acknowledgments

We would like to thank Yaser Al-Onaizan, Graham Phillips, and Bonnie Glover Stalls for their assistance with this work.

References

Ageno, A. I. Castellón, F. Ribas, G. Rigau, H. Rodríguez, A. Samiotou. 1994. TGE: Tlink Generation Environment. In *Proceedings of the 15th COLING.* Kyoto.

Agirre, E., X. Arregi, X. Artola, A. Díaz de Ilarazza, K. Sarasola. 1994. Conceptual Distance and Automatic Spelling Correction. In *Proceedings of the Workshop on Computational Linguistics for Speech and Handwriting Recognition.* Leeds.

Knight, K. and S. Luk. 1994. Building a Large-Scale Knowledge Base for Machine Translation. In *Proceedings of the AAAI-94 Conference.*

Knight, K. I. Chander, M. Haines, V. Hatzivassiloglou, E.H. Hovy, M. Iida, S.K. Luk, R.A. Whitney, and K. Yamada. 1995. Filling Knowledge Gaps in a Broad-Coverage MT System. In Pro*ceedings of the 14th IJCAI Conference.* Montreal, Canada.

Miller, G. 1990. Wordnet: An On-line Lexical Database. *International Journal of Lexicography* 3(4) (special issue).

Okumura, A. and E.H. Hovy. 1994. Lexicon-to-Ontology Concept Association Using a Bilingual Dictionary. In *Proceedings of the First AMTA Conference.* Columbia, MD.

Rigau, G. and E. Agirre. 1995. Disambiguating Bilingual Nominal Entries against WordNet. In *Proceedings of the 7th ESSLI Conference.* Barcelona.