****

Topic: HOW CAN INTERLINGUAL SYSTEMS BE SCALED UP?
Title: Use of Syntax-Semantics Relation for Automatic Construction of Interlingual Lexicons
Bonnie J. Dorr
University of Maryland

bonnie@umiacs.umd.edu

Our research at the University of Maryland has focused on the construction of dictionaries for interlingual applications. One of the central questions we have addressed is that of how to build automatic procedures for scaling up these dictionaries. We believe that answering this question is the first step toward building serious, large-scale (and completed) systems for use in tasks such as machine translation, foreign language tutoring, and other multilingual information processing tasks. We take the components of meaning in our dictionary representations to be interlingual and have used these as the basis of dictionaries for languages such as Arabic, Spanish, French, and Korean.

While our emphasis appears to lie on the "supply" side of the equation (i.e., construction of large dictionaries), we are well aware that these representations must be applicable to the "demand" side of the equation (i.e., large, working systems). Nirenburg (1996) describes these two sides and categorizes the work of lexicon researchers accordingly. We view our position in this categorization to be much more fuzzy than described, falling across the supply-demand boundary.

Our approach to building large dictionaries relies on a number of techniques based on the notion that there exists a basic relation between the semantics of a verb and its corresponding syntactic behavior. Of course, we need to provide convincing evidence for his underlying assumption—the central thesis of Levin (1993)—i.e., we need to show that the semantics of a verb and its syntactic behavior are predictably related. A large part of our work has focused on demonstrating the validity of this hypothesis (Dorr and Jones, 1996a). In our experiments, we provided theoretical justification for the bases upon which we proceeded for our lexical-acquisition work, i.e., we have demonstrated that 98% of Levin's semantic classes have uniquely identifying syntactic signatures (i.e., clusters of syntactic behaviors). We view these experiments as a necessary step for proceeding with further experimentation for construction of verb classes, i.e., we want to ensure that our starting point is solid before undertaking large-scale acquisition based on the syntax-semantics relation.

Upon completion of these experiments, we have begun a long process of verb categorization of "novel" (previously unseen) verbs in English. This work has resulted in a database of verbs, classified semantically based on a system similar to that of Levin (1993). We are currently developing syntax-semantics tests for other languages, e.g., Arabic, Spanish, French, and Korean, so that we can similarly classify verbs

for those languages.

A common point of confusion (e.g., during the presentation of this work at ACL, COLING, and related workshops in summer of 1996) concerns the nature of the semantic classes upon which we have built our dictionaries. Several researchers (Saint-Dizier, 1996, among others) have pointed out that these classes are not universal, and thus cannot serve as the basis of an interlingua. What should be kept in mind is that it is not the CLASSES that are intended to be universal, but the COMPONENTS OF MEANING that underlie these classes. By their very nature (i.e., that they are based on English-specific syntactic "alternations") the English semantic classes do not hold cross-linguistically. However, it was not the intention to classify translation equivalents identically, but to isolate the meaning components associated with semantic classes, and to then find a relation between these meaning components. The meaning components, not the syntactic behaviors, are expected to be language-independent.

For example, the "Motion/Impact" verbs, but not the "Change-of-state" verbs participate in the conative:

Motion/Contact
  She tapped at the window
  She banged at the door

Change of State
  She broke at the window
  She smashed at the door

Although the conative does not exist in other languages (e.g., French), there is clearly some meaning component associated with "tap" and "bang" (contact, but no no change in structural integrity) that is not associated with break and smash (contact and change in structural integrity). While we wouldn't use the conative in French, clearly the notions of contact and structural integrity can be expressed in French, and so the isolation of these meaning components (through application of the conative test in English) is clearly of cross-linguistic value. These meaning components are what should then be included as part of the interlingual structure.

While we have justified the use of the syntax-semantics relation as the basis for building an interlingua, we are still faced with the problem of scaling up our database of lexical representations. We have addressed this problem by using automatic procedures based on syntactic tests (such as the ones above) for mapping verbs onto lexical-semantic representations. One of the major difficulties we

were faced with in automatic classification of unknown verbs is that of "polysemy" (word sense ambiguity), which, in previous work (Dorr et al. 1995) resulted in very low "precision" (i.e., a high percentage of verbs assigned incorrectly to semantic classes)—13%. (We use precision as our primary metric for judging the effectiveness of our acquisition technique. Details are given in (Dorr and Jones, 1996b).) As an attempt to address the polysemy problem, we used a WordNet based filter for classification of unknown words. We tested the filter on three different proportions of the original 2813 Levin verbs: (a) 50%, (b) 70%, and (c) 90%, chosen randomly. We then checked whether the "unknown" verbs (those not used to construct the semantic filter) were assigned to their correct classes. The result was a drastic improvement in precision—64% (for the 90% case) in contrast to the 13% precision of Dorr et al. (1995).

Our experiments indicate that, not surprisingly, but not insignificantly, the syntax-semantics relationship is very clear, particularly in our later experiment where we accounted for word sense ambiguity. These experiments served to validate Levin's claim that verb semantics and syntactic behavior are predictably related and also demonstrated that a significant component of any lexical acquisition program is the ability to perform word-sense disambiguation. We have used the results of our experiments to aid in the construction and augmentation of online dictionaries for novel verb senses and we are currently porting these results to new languages using online bilingual lexicons.

REFERENCES

Dorr, Bonnie J. and Douglas Jones (1996a). "Role of Word Sense Disambiguation in Lexical Acquisition: Predicting Semantics from Syntactic Cues," Proceedings of the International Conference on Computational Linguistics, Copenhagen, Denmark.

Dorr, Bonnie J. and Douglas Jones (1996b). "Acquisition of Semantic Lexicons: Using Word Sense Disambiguation to Improve Precision," Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics, Santa Cruz, CA.

Dorr, Bonnie J., Joseph Garman, and Amy Weinberg (1995). "From Syntactic Encodings to Thematic Roles: Building Lexical Entries for Interlingual MT," Machine Translation, 9:3-4, pp.~71-100.

Nirenburg, Sergei (1996). "All the Lexical-Semantic Flowers Bloom, Each by Itself", Proceedings of the Workshop on Breadth and Depth of Semantic Lexicons, 34th Annual Conference of the Association for Computational Linguistics, Santa Cruz, CA, 1996.

Levin, Beth (1993). "English Verb Classes and Alternations: A Preliminary Investigation", University of Chicago Press, Chicago, IL.

Saint-Dizier, Patrick (1996). "Semantic Verb Classes based on 'Alternations' and on WordNet-like Semantic Criteria: a Powerful Convergence", Proceedings of the Workshop on Predicative Forms in Natural Language and Lexical Knowledge Bases, Toulouse, France, 1996.

****