Merging LDOCE and WordNet

Kevin Knight
USC/ISI
4676 Admiralty Way
Marina del Rey CA 90292
knight@isi.edu

(Position Paper)

1 Goal

One of the current goals of the Pangloss knowledge-based MT project is the construction of a large (35000-node) ontology and lexicon. The upper region of the ontology (called the Ontology Base) is a 400-node synthesis of the PENMAN Upper Model and ONTOS. The rest is a network of commonly-encountered objects, processes, qualities, relations, etc., found in the application domain. Manually constructing a network of this scale is time-consuming, so we must turn to automatic methods where feasible.

The first stage of our ontology building consists of taxonomizing tens of thousands of English word senses, and subordinating them to the Ontology Base. This we describe below. Later stages will address the insertion of additional semantic information such as restrictions on actors in events, domain/range constraints on relations, and so forth.

2 Resources

Two online knowledge sources currently being used in various NLP projects are the Longman Dictionary of Contemporary English (LDOCE) [Longman, 1978] and WordNet [Miller et. al. 1990]. LDOCE is a learner's dictionary of English. Each word sense is annotated with:

- a short definition and examples of usage
- one or more of 81 syntactic codes
- one of 33 semantic codes (PERSON, LIQUID, etc.) for nouns
- one of 124 pragmatic codes (MEDICINE, MILITARY, etc.)

The semantic codes induce a flat hierarchy of noun senses. In addition to this, researchers at New Mexico State University have built an algorithm that attempts to locate and disambiguate genus words in sense definitions. It decides that an "aisle_0_2" is a type of "passage_0_7," e.g., based on the definition "aisle. n. 2. a narrow passage between two ..." This yields a second fairly flat hierarchy, but one with many errors.

The WordNet dictionary groups synonymous word senses into single units ("synsets"). This is highly desirable from the viewpoint of ontological modeling. WordNet also contains:

- brief English descriptions of most synsets
- a deep, well-developed hierarchy
- part-of, antonym, and other relations

However, WordNet has no syntactic or pragmatic information. What we want is the best of both resources.

3 Merging

Merging two knowledge bases involves finding out which pairs of object descriptions correspond to one another. In general, this is a difficult analogy problem. Merging LDOCE and WordNet is tractable because all "concepts" (senses and synsets) are annotated with English words and definitions. The words strongly constrain the possibilities for correspondence. Unambiguous words provide immediate matchups—e.g., "gazelle_0_0" in LDOCE corresponds to (GAZELLE) in WordNet. Ambiguous words like "ball" are much harder: with 7 senses in LDOCE and 9 in WordNet, there are over 57,000,000 ways to pair up the senses.

This is the problem. Solving it yields several benefits:

1. It provides a syntactic lexicon for WordNet.
2. It groups LDOCE senses into synonymous sets.
3. It organizes LDOCE senses into a deep hierarchy.
4. It allows us to correct errors in the LDOCE and WordNet hierarchies.

However, the primary benefit is to allow us to automatically taxonomize tens of thousands of LDOCE word senses and subordinate them to the Ontology Base.

4 Algorithms

We have developed two merging algorithms. Algorithm A is based on the idea that corresponding senses often have similar definitions. For example, two corresponding senses of "foil" both mention "metal," while two others mention "sword." The algorithm computes a similarity measure for each possible pair of senses, based roughly on the number of open-class, content words they share. Senses are then matched, starting with the most similar pairs. Empirical results are as follows. For words with exactly two senses in both LDOCE and WordNet:

similarity measure	correct match	incorrect: wrong match	incorrect: no match
>= 0.0	71%	17%	12%
>= 1.0	80%	15%	5%
>= 2.0	88%	9%	3%

With highly ambiguous words, overall correctness drops to 43%, but remains high (77% and 85%) at the higher confidence levels. Of course, at those levels, fewer matches are proposed.

Algorithm B ignores sense definitions, and instead uses the isa-hierarchies in LDOCE and WordNet. First, all senses referred to by unambiguous words (e.g., "gazelle") are matched. For each match entered, two searches for further matches are performed. The first search looks upward in the hierarchies. Senses annotated with the same words are then matched. E.g., even though "animal" is ambiguous in LDOCE, only one sense dominates "gazelle." The second search proceeds downwards. Once two senses are matched, any word that is unambiguous inside the subtrees rooted at those senses provides a new candidate match. E.g., "seal" is highly ambiguous, but it is unambiguous inside the two animal hierarchies.

Forthcoming reports will detail refinements, combinations, and results from these algorithms, and will describe the initial ontology and lexicon produced with them.