Empirical Acquisition of Word-Sense Distinctions

Tom O'Hara
Department of Computer Science
& Computing Research Laboratory
New Mexico State University
Las Cruces, NM 88003-0001
tomohara@nmsu.edu

Applications currently make use of many kinds of lexical information: categorial relations ("dog" *is-a* "canine"), synonymy ("pooch" *same-as* "dog") and word associations ("walk" *occurs-often-with* "dog"). However, an important type of information, *differentia*, is often omitted, especially in broad-coverage applications. Differentia are properties that distinguish a concept from others belonging to the same higher-level category. For instance, both beagles and wolfhounds are hounds, but the former are small, whereas the latter are quite large. Applications should incorporate differentia to provide finer word-sense distinctions and to facilitate inference of information not mentioned in the text.

Determining the differentia is a difficult task, since the available knowledge sources define these properties using natural language. For instance, WordNet (Miller 1990), a commonly used source of lexical knowledge, provides explicit information on categorial relationships but leaves the differentia implicit in the definitions. This work will investigate empirical approaches for extracting these properties from machine readable dictionaries (MRD's) and text corpora. The result will be lexical relations between the word being defined and words used in the definition. There has been some work on deriving differentia, but these have relied predominantly on manually developed heuristics. Here, corpus-derived associations will augment such heuristics for extracting information from MRD's. Furthermore, this work will investigate a novel use of Bayesian networks for representing the various types of lexical knowledge in order to model the uncertainty in the relations and support integration of statistical and analytical knowledge.

Dictionary definitions use certain fixed patterns, often with prepositional phrases, to indicate differentia. However, since prepositions are highly ambiguous, the same pattern can be used for different properties. To address this problem, syntactic pattern matching will be applied to each definition to identify potential properties. Then, statistical classification will be used to select the most plausible ones. To support this work, a representative sample of definitions will be annotated to indicate the properties that apply. This will serve as the primary training data for the classifier. To allow for a fallback mechanism, a separate classifier will be trained on the semantic role annotations in the second release of the Penn Treebank (Marcus *et al.* 1994).

Providing sufficient annotated training data for automated disambiguation of the definition text would require a large corpus of sense-tagged data for all content words, which is currently not available. Therefore, unsupervised corpus-based approaches will be used for disambiguation, such as extensions to Yarowsky's (1992) method. He uses co-occurrence statistics for the words listed under each Roget category to determine contextual support for that category. The extensions will include dynamic use of WordNet categories for defining the topics. In addition, training will be over definitions as well as over general text to allow for constraints on the potential topics.

To represent word-sense distinctions using Bayesian networks, lexical relations will be modeled by causal links among word sense nodes. We will use a representation with a clear probabilistic interpretation to enable us to take advantage of formal tools from the applied statistics literature. To support word-sense disambiguation, a Bayesian network used for modeling the interdependencies among word senses will be integrated with a statistical word-sense classifier that uses clues from the local syntactic context, following Bruce and Wiebe (1995). This work builds upon theirs by including the differentia, allowing for links between senses of words in different categories (augmenting the explicit WordNet links, which are mostly in the same category). Integrating differentia obtained from an MRD with information derived from a corpus should improve performance on word-sense disambiguation and other lexical applications as well.

References

Bruce, R., and Wiebe, J. 1995. Towards the acquisition and representation of a broad-coverage lexicon. In *Proc. AAAI Spring Symposium on Representation and Acquisition of Lexical Knowledge*, 174–179.

Marcus, M.; Kim, G.; Marcinkiewicz, M.; et al. 1994. The penn treebank: Annotating predicate argument structure. In *Human Language Technology Workshop*.

Miller, G. 1990. Wordnet: An on-line lexical database. *International Journal of Lexicography* 3(4).

Yarowsky, D. 1992. Word-sense disambiguation using statistical models of roget's categories trained on large corpora. In *Proc. COLING-92*, 454–460.