Semantic Lexicon Acquisition for Learning Natural Language Interfaces

Cynthia A. Thompson and Raymond J. Mooney
Department of Computer Sciences
University of Texas
Austin, TX 78712
cthomp@cs.utexas.edu, mooney@cs.utexas.edu

Abstract

This paper describes a system, WOLFIE (WOrd Learning From Interpreted Examples), that acquires a semantic lexicon from a corpus of sentences paired with representations of their meaning. The lexicon learned consists of words paired with meaning representations. WOLFIE is part of an integrated system that learns to parse novel sentences into semantic representations, such as logical database queries. Experimental results are presented demonstrating WOLFIE's ability to learn useful lexicons for a database interface in four different natural languages. The lexicons learned by WOLFIE are compared to those acquired by a competing system developed by Siskind (1996).

Content areas: corpus-based methods, inductive learning, machine learning, natural language processing

Introduction & Overview

The application of learning methods to natural-language processing (NLP) is a growing area. Using machine learning to help automate the construction of NLP systems can eliminate much of the difficulty of building such systems by hand. The semantic lexicon, or the mapping from words to meanings, is one component that is typically difficult to construct and update, and changes from one application to the next. Constructing a lexicon by hand is difficult and time consuming, as noted by Copestake et al. (1995) and Walker & Amsler (1986). Johnston, Boguraev, & Pustejovsky (1995) discuss the need for systems that can learn the meanings of novel words.

This paper describes a system, WOLFIE (WOrd Learning From Interpreted Examples), that learns a semantic lexicon of word/meaning pairs from input consisting of sentences paired with semantic representations. The goal of this research is to automate lexicon construction for an integrated NLP system that acquires semantic parsers and lexicons. A subgoal is to learn a lexicon that is as good or better than a manually-built one based on performance on a chosen task.

Although a few others (Siskind 1996; Hastings & Lytinen 1994; Brent 1991) have presented systems for semantic lexical acquisition, this work is unique in combining several features. First, interaction with a system, CHILL (Zelle 1995), that learns to parse sentences into their semantic representations, is demonstrated. Second, it uses a fairly simple batch, greedy algorithm that is quite fast and accurate. Third, it is easily extendible to new representation formalisms. Finally, it is able to bootstrap from an existing lexicon.

We tested WOLFIE on its ability to acquire a semantic lexicon for the task of answering geographical database queries, using a corpus of queries collected from human subjects and annotated by an expert with their executable logical form. In this process, it has been integrated with CHILL, which learns parsers but requires a semantic lexicon (previously built manually). Results demonstrate that the final acquired system performs nearly as accurately at answering novel questions when using a learned lexicon as compared to a hand-built lexicon. The system is also compared to an alternative lexical acquisition system developed by Siskind (1994; 1996), demonstrating superior performance on this task. Finally, we translated the corpus from English into Spanish, Japanese, and Turkish and ran experiments on learning database interfaces with these languages as well. Overall, the results demonstrate a robust ability to acquire accurate lexicons that can be directly used for semantic parsing.

Background

The output produced by WOLFIE can be used to assist a larger language acquisition system; in particular, it is currently used as part of the input to a parser acquisition system called CHILL. CHILL uses inductive logic programming (Muggleton 1992; Lavrac & Dzeroski 1994) to learn a deterministic shift-reduce parser written in Prolog. The input to CHILL is a corpus of sen-