Machine Learning Applied to Natural Language Processing: The Acquisition of the Lexicon

Cynthia A. Thompson
Department of Computer Sciences
University of Texas
Austin, TX 78712
cthomp@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).

Introduction

My research is in artificial intelligence, and uses insights from several of its subfields, including machine learning, natural language processing, and inductive logic programming. My dissertation research has demonstrated the application of symbolic machine learning techniques to the area of natural language processing (NLP).

NLP systems are a crucial component in many intelligent systems. However, the most successful systems to date for parsing natural language into semantic representations have relied on human engineering efforts. Building NLP systems is a long, arduous process, often requiring the expertise of a trained linguist, and typically resulting in a system that is brittle and narrowly applicable. Therefore, many researchers have begun to explore empirical, or corpus-based, methods for natural language processing. This paradigm combines the use of corpora, statistics, and machine learning methods to automatically build NLP systems.

My dissertation research extends the work in this field by addressing the automated acquisition of semantic lexicons. For this work, I helped develop an integrated natural language learning system. Given a corpus of annotated sentences, the system acquires a lexicon and parser to process the sentences. The system's goal is to learn to process novel sentences from the same domain as the training corpus. My focus has been on the development of the lexical acquisition component. I have performed experiments comparing this component to the only other previous system of its kind (Siskind 1996), with my system yielding superior results.

Semantic Lexicon Acquisition

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.

In our definition of the semantic lexicon acquisition task, we are given a set of sentences, each consisting of an ordered list of words and annotated with a single semantic representation, and we assume that each representation can be `fracture`d into all of its components (Siskind 1992). Given a valid set of components, they can be constructed into a valid sentence meaning using a relation we will call `compose`.

The goal is to find a semantic lexicon that will assist parsing. Such a lexicon consists of *(phrase, meaning)* pairs, where the phrases and their meanings are extracted from the input sentences and their representations, respectively, such that each sentence's representation can be `compose`d from a set of components each chosen from the potential meanings of a (unique) phrase appearing in the sentence. If such a lexicon is found, we say that the lexicon *covers* the corpus. Ideally, we would like to minimize the size and ambiguity of the learned lexicon, since this should ease the parser acquisition task.

Note that we allow phrases to have multiple meanings (homonymy) and for multiple phrases to have the same meaning (synonymy). Also, some phrases in the sentences may have a null meaning. We make only a few fairly straightforward assumptions about the input. First is *compositionality*, i.e. the meaning of a sentence is `compose`d from the meanings of phrases in that sentence. Second, we assume each component of

Derive possible phrase/meaning pairs by sampling the input sentence/representation pairs that have phrases in common, and deriving the common substructure in their representations.
Until the input is covered, or there are no remaining possible pairs do:
1) Add the best phrase/meaning pair to the lexicon.
2) Constrain the remaining possible phrase/meaning pairs to reflect the pair just learned.
Return the lexicon of learned phrase/meaning pairs.

Figure 1: WOLFIE Algorithm Overview

the representation is due to the meaning of a word or phrase in the sentence, not to an external source such as noise. Third, we assume the meaning for each word in a sentence appears only once in the sentence's representation. The second and third assumptions are preliminary, and we are exploring methods for relaxing them. If any of these assumptions are violated, we do not guarantee coverage of the training corpus; however, the system can still be run and learn a potentially useful lexicon.

Overview of the WOLFIE Algorithm

In order to limit search, a greedy algorithm is used to learn phrase meanings. At each step, the best phrase/meaning pair is chosen, according to a heuristic described below, and added to the lexicon. The initial list of potential meanings for a phrase is formed by finding the common substructure between sampled pairs of representations of sentences in which the phrase appears. In the current implementation, phrases are limited to at most two words.

The WOLFIE algorithm, outlined in Figure 1, has been implemented to handle two kinds of semantic representations. One is a case-role meaning representation based on *conceptual dependency* (Schank 1975). For example, the sentence "The man ate the cheese" is represented by: `[ingest, agent:[person, sex:male, age:adult], patient:[food, type:cheese]]`. Experiments in this domain were presented in Thompson (1995). The second representation handled is a logical query domain, where natural language questions are mapped directly into Prolog queries that can be executed to produce an answer. For example, "What is the capital of the state with the biggest population?" is mapped into the query: `answer(C, (capital(S,C), largest(P, (state(S), population(S,P))))).`

We now briefly describe the algorithm. We first select a random sample of the sentences that each one and two word phrase appears in, and derive an initial set of possible meanings for each phrase. This is done by deriving common substructure between pairs of representations of sentences that contain these phrases. After deriving these initial meanings, the greedy search begins. The heuristic used to evaluate candidate pairs has five weighted components:

1. Ratio of the number of times the phrase appears with the meaning to the number of times the phrase appears, or $P(meaning|phrase)$.
2. Ratio of the number of times the phrase appears with the meaning to the number of times the meaning appears, or $P(phrase|meaning)$.
3. Frequency of the phrase, or $P(phrase)$.
4. Percent of orthographic overlap between the phrase and its meaning.
5. The generality of the meaning.

At each step, the candidate word/meaning pairs are ranked according to this heuristic, and the best pair is added to the lexicon. Step two of the loop constrains the possible meanings of the remaining unlearned phrases to take into account the meaning just learned. Such constraints exist because of the assumption that each portion of the representation is due to at most one phrase in the sentence. Therefore, once part of a sentence's representation is covered by the meaning of one of its phrases, no other phrase in the sentence can be paired with that meaning. The greedy search continues until the lexicon covers the training corpus.

Experimental Results

This section describes our experimental results on a database query application. The corpus contains 250 questions about U.S. geography paired with logical representations. To evaluate the learned lexicons, we measured their utility as background knowledge for CHILL. This is performed by choosing a random set of 25 test examples and then creating lexicons and parsers using increasingly larger subsets of the remaining examples. The test examples are parsed using the learned parser, the resulting queries submitted to the database, the answers compared to those generated by the correct representation, and the percentage of correct answers recorded. We repeated the above steps for ten different random splits of the data. We compared our system to that developed by Siskind (1996). Siskind's system is an incremental learner, while ours is batch. To make a closer comparison between the two, we ran his in a "simulated" batch mode, by repeatedly presenting the corpus 500 times. Finally, since Siskind has no measure of orthographic overlap, and it could arguably give our system an unfair advantage on this data, we ran WOLFIE with a weight of zero for this component.

Figure 2 shows learning curves for CHILL when using the lexicons learned by WOLFIE (CHILL+WOLFIE) and by Siskind's system (CHILL+Siskind). The uppermost curve (CHILL+corrlex) is CHILL's performance when given a hand-built lexicon. Finally, the horizontal line shows the performance of a benchmark, *Geobase*. *Geobase* is a hand-build natural language interface to a simple geography database containing about 800 facts,

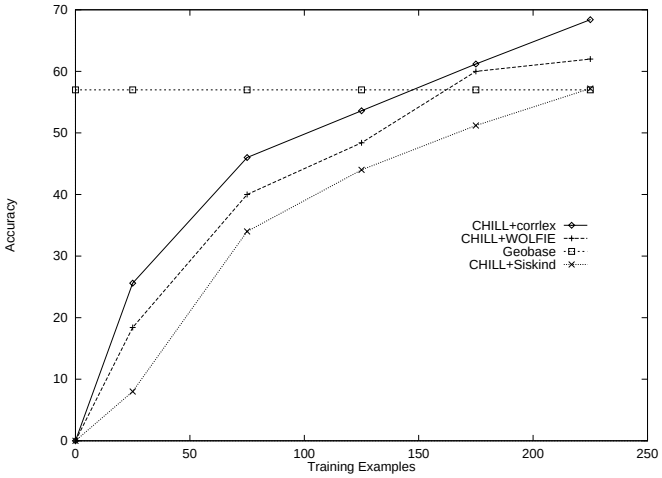

Figure 2: Accuracy on English Geography Corpus

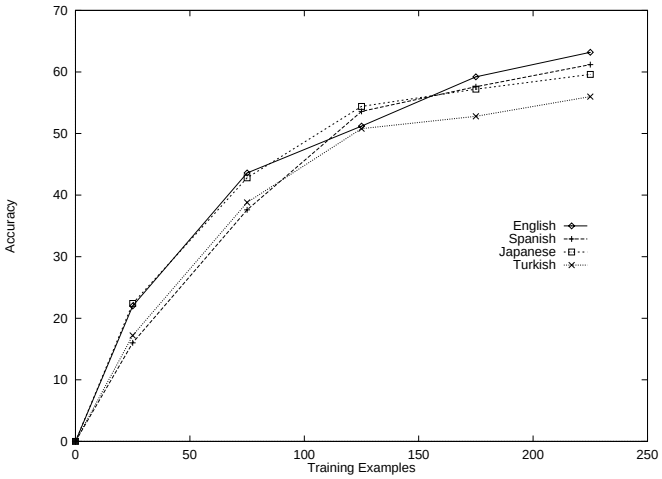

Figure 3: Accuracy on All Four Languages

and is supplied with Turbo Prolog 2.0 (Borland International 1988).

The results show that a lexicon learned by WOLFIE led to parsers that were almost as accurate as those generated using a hand-built lexicon. The best accuracy is achieved by the hand-built lexicon, followed by WOLFIE, followed by Siskind's system. All the systems do as well or better than *Geobase* by 225 training examples.

We also had the geography query sentences translated into Spanish, Japanese and Turkish, and ran similar tests to determine how well WOLFIE could learn lexicons for these languages, and how well CHILL could learn to parse them. Figure 3 shows the results. The performance differences among the four languages are quite small, demonstrating that our methods are not language dependent.

Conclusions

Acquiring a semantic lexicon from a corpus of sentences labeled with representations of their meaning is an important problem that has not been widely studied. WOLFIE demonstrates that a fairly simple greedy symbolic learning algorithm performs fairly well on this task and obtains performance superior to a previous lexicon acquisition system on a corpus of geography queries. Our results also demonstrate that our methods extend to a variety of natural languages besides English.

Most experiments in corpus-based natural language have presented results on some subtask of natural language, and there are few results on whether the learned subsystems can be successfully integrated to build a complete NLP system. The experiments presented in this paper demonstrated how two learning systems, WOLFIE and CHILL were successfully integrated to learn a complete NLP system for parsing database queries into executable logical form given only a single corpus of annotated queries.

Acknowledgements

We would like to thank Jeff Siskind for providing us with his software, and for all his help in adapting it for use with our corpus. This research was supported by the National Science Foundation under grants IRI-9310819 and IRI-9704943. Thanks also to Agapito Sustaita, Esra Erdem, and Marshall Mayberry for their translation efforts.

References

Borland International. 1988. *Turbo Prolog 2.0 Reference Guide*. Scotts Valley, CA: Borland International.

Brent, M. 1991. Automatic acquisition of subcategorization frames from untagged text. In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, 209–214.

Hastings, P., and Lytinen, S. 1994. The ups and downs of lexical acquisition. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, 754–759.

Schank, R. C. 1975. *Conceptual Information Processing*. Oxford: North-Holland.

Siskind, J. M. 1992. *Naive Physics, Event Perception, Lexical Semantics and Language Acquisition*. Ph.D. Dissertation, Department of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, MA.

Siskind, J. M. 1996. A computational study of cross-situational techniques for learning word-to-meaning mappings. *Cognition* 61(1):39–91.

Thompson, C. A. 1995. Acquisition of a lexicon from semantic representations of sentences. In *Proceedings of the 33rd Annual Meeting of the Association for Computational Linguistics*, 335–337.

Zelle, J. M. 1995. *Using Inductive Logic Programming to Automate the Construction of Natural Language Parsers*. Ph.D. Dissertation, University of Texas, Austin, TX. Also appears as Artificial Intelligence Laboratory Technical Report AI 96-249.