Semantic Lexicon Acquisition for Learning Natural Language Interfaces

Cynthia Ann Thompson

Report AI99-278 December 1998

cthomp@csli.stanford.edu
http://www-csli.stanford.edu/~cthomp/
Artificial Intelligence Laboratory
The University of Texas at Austin
Austin, TX 78712
Copyright

by

Cynthia Ann Thompson

1998
Semantic Lexicon Acquisition for Learning Natural Language Interfaces

by

Cynthia Ann Thompson, B.S., M.A.

Dissertation

Presented to the Faculty of the Graduate School of

The University of Texas at Austin

in Partial Fulfillment

of the Requirements

for the Degree of

Doctor of Philosophy

The University of Texas at Austin

December 1998
Semantic Lexicon Acquisition for
Learning Natural Language Interfaces

Approved by
Dissertation Committee:

RAYMOND MOONEY
CLAIRE CARDIE
BENJAMIN KUIPERS
RISTO MIIKKULAINEN
ELAINE RICH
To my parents who gave me my wings, and to Bill who gave me the wind
I did not complete this dissertation in isolation. Many people directly and indirectly contributed to the ideas contained within; many also contributed to the sanity and financial well-being of the author. If I have forgotten anyone, please do not be offended, for you are all in my heart, even if you are not in my head at the moment that I am writing this. First, I would like to thank my advisor, Ray Mooney, who taught me much about how to be a good academic researcher. His advice and input were invaluable in assisting me with the research for this dissertation. I must express my gratitude to Ben Kuipers, who acted much as an unofficial backup advisor. I enjoyed our conversations about life, the universe, and everything. Thanks also to my other committee members: Claire Cardie, Risto Miikkulainen, and Elaine Rich.

Robert Rodman was instrumental in encouraging me to go to graduate school in the first place, for which I thank him. Anita Borg and her work with sisters helped keep me going once I got here. Many others had a positive influence on this research and on my life while I was at the University of Texas. Special mention goes to my office mates who helped me survive: Paul Baffes, Mary Elaine Califf, Harold Chaput, Dan Clancy, Tara Estlin, Bert Kay, Jeff Mahoney, Emilio Remolina, Rupert Tang, Tal Tversky, and especially John Zelle for discussions on our research, Mike Hewett for the chocolate, and Soumya Ramachandran for being my best (female) friend.

Thanks to my translators: Agapito Sustaita, Esra Erdem, and Marty Mayberry. I am grateful to many others, including Chris Edmondson-Yurkanan, Vicki Almstrum, Ted Pedersen, and Mike Cline. Thanks to the Austin RC community, for reminding me of my complete brilliance, especially Ruth. Then there are the many friends who reminded me to also have fun: Lyn Pierce (for lunches and fun), Debbie Miller (for thoughts and games), Dave Pierce (for being a great brother-in-law), Michael Bogomolny (for haiku and lunches), and Todd and Mary West (for showing me the way).

I would like to thank my family, without whom nothing is possible. My parents were always a great support, even though they worried too much. My brothers didn't quite get why I was doing this to myself, but thought it was cool anyway. My family-in-law welcomed me with open arms and tons of confidence. Finally, no amount of thanks is enough for Bill, for being my best (male) friend, for helping me keep touch with reality, and just for being you.

Thanks go to Jeff Siskind for permission to use his software and for his help in getting it working with our data. Portions of this research were supported by the National Science Foundation under grant IRI-9310819 and through a grant from the Deimler-Benz Palo Alto Research Center. The author was also partially supported by a scholarship from Trident.
The University of Texas at Austin
December 1998
A long-standing goal for the field of artificial intelligence is to enable computer understanding of human languages. A core requirement in reaching this goal is the ability to transform individual sentences into a form better suited for computer manipulation. This ability, called semantic parsing, requires several knowledge sources, such as a grammar, lexicon, and parsing mechanism.

Building natural language parsing systems by hand is a tedious, error-prone undertaking. We build on previous research in automating the construction of such systems using machine learning techniques. The result is a combined system that learns semantic lexicons and semantic parsers from one common set of training examples. The input required is a corpus of sentence/representation pairs, where the representations are in the output format desired. A new system, WOLFIE, learns semantic lexicons to be used as background knowledge by a previously developed parser acquisition system, CHILL. The combined system is tested on a real world domain of answering database queries. We also compare this combination to a combination of CHILL with a previously developed lexicon learner, demonstrating superior performance with our system. In addition, we show the ability of the system to learn to process natural languages other than English. Finally, we test the system on an alternate sentence representation, and on a set of large, artificial corpora with varying levels of ambiguity and synonymy.

One difficulty in using machine learning methods for building natural language interfaces is building the required annotated corpus. Therefore, we also address this issue by using active learning to reduce the number of training examples required by both WOLFIE and CHILL. Experimental results show that the number of examples needed to reach a given level of performance can be significantly reduced with this method.
Contents

Acknowledgments vi

Abstract viii

List of Tables xii

List of Figures xiii

Chapter 1 Introduction 1

Chapter 2 Background 6
  2.1 CHILL 6
  2.2 The Work of Jeff Siskind 9

Chapter 3 Lexicon Learning and WOLFIE 11
  3.1 Problem Definition 11
  3.2 Potential Approaches 16
  3.3 The WOLFIE Algorithm and an Example 18
  3.4 Generating Candidate Lexicon Items 22
  3.5 Finding the Best Pair 23
  3.6 Constraining Remaining Candidates 25

Chapter 4 WOLFIE Experimental Results: Learning for Database-Query Parsing 30
  4.1 Methodology and Data 30
  4.2 Comparisons using English 31
  4.3 Comparisons using Spanish 34
  4.4 Accuracy on Other Languages 35
  4.5 LICS versus Fracturing 35
  4.6 A Larger Corpus 36
  4.7 Discussion 38

Chapter 5 WOLFIE Experimental Results: Artificial Corpora 40
  5.1 Case-Role Plus Conceptual Dependency 40
  5.2 A Second Set of Artificial Corpora 43
Chapter 6  Active Learning  
  6.1  Background ................................................. 48
  6.2  Application to Parser Learning  ......................... 49
  6.3  Experimental Results for Parser Acquisition ........... 51
  6.4  Application to Lexicon Acquisition  .................... 52
  6.5  Experimental Results for Lexicon Acquisition ......... 54
  6.6  Active Learning for the Integrated System ............. 55
  6.7  Experimental Results: Full System Active Learning .... 55
  6.8  Discussion .................................................. 56
  6.9  Committee-based Active Learning ....................... 57

Chapter 7  Related Work  
  7.1  Lexicon Learning ............................................ 60
     7.1.1  Learning to Execute Actions  ...................... 61
     7.1.2  Learning Mappings to Visual Input ............... 61
     7.1.3  Learning for Semantic Parsing .................... 62
     7.1.4  Learning Syntax and Semantics ................... 62
  7.2  Other Related Work ....................................... 63
     7.2.1  Translation Lexicons ............................... 63
     7.2.2  Acquisition from MRDs ............................. 63
  7.3  Active Learning ........................................... 63

Chapter 8  Future Directions  
  8.1  Lexicon Learning ........................................... 65
     8.1.1  Improving the Algorithm  ......................... 65
     8.1.2  Generalizing the Problem Definition ............. 67
     8.1.3  Handling Unknown Words ......................... 68
     8.1.4  Additional Evaluation ............................. 69
  8.2  Active Learning ........................................... 69

Chapter 9  Conclusion  

Appendix A  Sample Training Input  
  A.1  Original Geography Corpus  ............................ 73
  A.2  New Geography Queries  ................................. 74

Appendix B  Sample Learned Lexicons  
  B.1  English ..................................................... 75
  B.2  Spanish ...................................................... 77
  B.3  Turkish ....................................................... 78
  B.4  Japanese ...................................................... 79

Bibliography  80
List of Tables

5.1 Number of Examples to Reach 75% Precision .................. 45
List of Figures

1.1 Sample Annotated Corpus .................................................. 2
1.2 Sample Semantic Lexicon .................................................. 2

2.1 The ChilI Architecture ................................................... 7
2.2 Shift-Reduce Case-Role Parsing of “The man ate the pasta.” ................. 8
2.3 The Integrated System .................................................... 8

3.1 Labeled Graphs and Interpretations ..................................... 12
3.2 Meanings ................................................................. 13
3.3 A Second Tree ............................................................ 13
3.4 The Size of $F$ ............................................................ 14
3.5 Covering Example .......................................................... 17
3.6 WOLFIE Algorithm Overview ........................................... 19
3.7 Database Queries as Trees ............................................... 20
3.8 A Second Database Query ............................................... 21
3.9 LICS Plus LGGs .......................................................... 21
3.10 The Candidate Generalization Phase ................................... 26
3.11 Unsuccessful Parse of “The girl ate the pasta with the cheese.” .......... 29

4.1 Accuracy on English Geography Corpus ................................ 32
4.2 Accuracy Given Closed Class Words ...................................... 33
4.3 Train Set Parsing Accuracy ............................................... 34
4.4 Accuracy on Spanish ...................................................... 35
4.5 Accuracy on All Four Languages ........................................ 36
4.6 Fracturing vs. LICS: Accuracy .......................................... 37
4.7 Fracturing vs. LICS: Number of Candidates ............................. 37
4.8 Fracturing vs. LICS: Learning Time ..................................... 38
4.9 Accuracy on the Larger Geography Corpus ................................ 39

5.1 Intersection Accuracy for the Modified M & K Corpus .................. 42
5.2 CHILL Accuracy on the M & K Corpus .................................. 43
5.3 Baseline Artificial Corpus ................................................ 44
5.4 A More Ambiguous Artificial Corpus .................................... 45
5.5 Increasing the Level of Ambiguity ....................................... 46
5.6 Increasing the Level of Synonymy ....................................... 46
6.1 Selective Sampling Algorithm ........................................ 49
6.2 Active Learning for CHILL ........................................... 50
6.3 Active Learning on the Geography Corpus .......................... 52
6.4 Active Learning on Larger Geography Corpus .............. 53
6.5 Error Rate on Larger Geography Corpus ......................... 53
6.6 Active Learning for WOLFIE ....................................... 54
6.7 Using Lexicon Certainty for Active Learning ..................... 55
6.8 Active Learning for Lexicon and Parser Learning .......... 56
Chapter 1

Introduction

A long-standing goal for the field of artificial intelligence is to enable computer understanding of human languages. Much progress has been made in reaching this goal, but much also remains to be done. Before artificial intelligence systems can meet this goal, they first need the ability to parse sentences, or transform them into a representation that can be more easily manipulated by computers. Several knowledge sources are required for parsing, such as a grammar, lexicon, and parsing mechanism.

Natural language processing (NLP) researchers have traditionally attempted to build these knowledge sources by hand, often resulting in brittle, inefficient systems that take many hundreds of hours to build. Overcoming this “knowledge acquisition bottleneck” by applying methods from machine learning is the main goal of this dissertation. We apply methods from empirical or corpus-based NLP to learn a semantic lexicon, and from active learning to reduce the annotation effort required to learn both semantic parsers and lexicons.

The “knowledge acquisition bottleneck” is particularly difficult in the case of knowledge needed for natural language processing, where the amount of information to encode is particularly large, and it is particularly difficult to make such knowledge complete and correct. To attack this problem, recent years have seen the explosion of various computational techniques applied towards the goal of automatically acquiring the knowledge and tools needed to process language (see Brill and Mooney (1997) for an overview). These techniques have drawn from various fields including statistics, machine learning, connectionism, and information-theoretic methods. Within this larger context of empirical NLP, our research focuses on machine learning and statistical methods to automate the construction of semantic lexicons and semantic parsers, from input consisting of sentences annotated with their appropriate semantic representations.

In the empirical approach to language acquisition, two tasks are performed. First, a training corpus is built. The type of corpus depends on the type of analysis desired. Typical corpora consist of sentences paired with parse trees, database queries, or semantic forms such as case-role representations. Another possible input form is syntactically tagged text. Both semantic and syntactic knowledge is often present in the sentence representations. Alternatively, unannotated corpora may be used in the case of systems using unsupervised learning. This dissertation focuses on supervised learning. Figure 1.1 shows an example of part of a corpus of sentences paired with database queries, represented in a logical form that can be executed in Prolog.

The second task required in the empirical approach to language acquisition is to design
What is the capital of the state with the biggest population?
answer(C, (capital(S,C), largest(P, (state(S), population(S,P))))).

What is the highest point of the state with the biggest area?
answer(P, (high_point(S,P), largest(A, (state(S), area(A,S))))).

What state is Texarkana located in?
answer(S, (state(S), eq(C, cityid(texarkana,), loc(C,S))).

What capital is the biggest?
answer(A, largest(A, capital(A))).

What is the area of the United States?
answer(A, (area(C,A), eq(C, countryid(usa)))).

What is the highest point in the state with the capital Des Moines?
answer(C, (high_point(B,C), state(B), capital(B,A),
            eq(A, cityid('des moines',))).)

Figure 1.1: Sample Annotated Corpus

([[capital], capital(_,_)),
([[state], state(_)),
([[biggest], largest(_,_)),
([[in], loc(_,_)),
([[highest, point], high_point(_,_)),
([[area], area(_,_)),
([[population], population(_,_)),
([[capital], capital(_))],

Figure 1.2: Sample Semantic Lexicon

and build the acquisition system itself. This system is then trained on the corpus of interest,
and the system learns to map each input sentence into its desired representation(s). By using
such acquisition systems, an application designer need only decide upon and design a suitable
representation, leaving the difficult issue of constructing a parser (or other grammar formalism) to
the machine learning system. The system described in this dissertation can in turn aid the parser
acquisition system by learning word meanings, thus shortening the inductive leap required.

One knowledge source required by systems that learn to map natural language sentences into
deeper semantic representations is the semantic lexicon. Here we use this term (often shortened to
just lexicon) to refer to a list of words (or multi-word phrases) paired with their meanings. We have
developed a system, WOLFIE (Word Learning From Interpreted Examples), to automatically learn
semantic lexicons from a corpus of sentences paired with their meanings. Experimental results show
that WOLFIE is able to learn correct and useful mappings. The output of the system can be used
to assist a larger language acquisition system; in particular, it is currently used as part of the input
to CHILL (Zelle & Mooney, 1993), a parser acquisition system. CHILL requires a word-to-meaning
lexicon as background knowledge in order to learn to parse into deep semantic representations. By
using WOLFIE, one of the inputs to CHILL is automatically provided, thus easing the task of parser
acquisition. Part of a semantic lexicon for the corpus in Figure 1.1 is shown in Figure 1.2.

Though there are existing computational lexicons (e.g., WordNet, (Beckwith, Fellbaum,
Gross, & Miller, 1991)) and on-line dictionaries (e.g., Longman Dictionary of Contemporary English
automated lexicon acquisition is important for several reasons. First, language is constantly changing: new words are created and additional senses are added to existing words. Second, existing lexicons are often customized to one domain, and new domains require slightly or even radically different meanings for many words. Finally, the organization of a lexicon by hand is pain-staking work, and a more logical or useful representation is likely to be found more quickly by automated methods. The automation of lexicon acquisition would enable lexicons to be updated continuously for each new domain with no need to hand-code the lexicon entries.

Before we proceed, a brief discussion of words and their meanings is warranted. As in Miller, Beckwith, Fellbaum, Gross, and Miller (1993), word form (or just word when the use is obvious) here refers to a “physical utterance or inscription,” and word meaning “to the lexicalized concept that a form can be used to express.” Further, in the context of this dissertation, a word meaning is a symbol, or set of symbols, corresponding to a word form and denoting information about the word that is in some way useful to the performance of intelligent reasoning in the domain in question. In other words, a word’s meaning is what we want the Natural Language Understanding (NLU) system to infer (e.g., add to its current knowledge about the sentence being processed), when it sees the word’s form. From this definition, it is clear that the meaning (or meanings) of a word is dependent on the representation used by the NLU system at hand. In some simple cases, for example, the meaning of man will just be man. In this dissertation, word forms will be set in bold font (e.g., man), and their meanings in tel-type font (e.g., [person,sex: male, age:adult]).

Many past systems designed to automate semantic lexicon acquisition (Fukumoto & Tsujii, 1995; Haruno, 1995; Johnston, Boguraev, & Pustejovsky, 1995; Webster & Marcus, 1995) focus only on acquisition of verbs or nouns, rather than all types of words. Also, these either do not experimentally evaluate their systems, or do not show the usefulness of the learned lexicons. Other approaches (Berwick & Pilato, 1987; Hastings & Lytinen, 1994) to lexicon acquisition assume access to an initial lexicon and parser. A related trend (Boguraev & Briscoe, 1989) is the extraction of lexical knowledge from machine-readable dictionaries such as LDOCE. However, lexicons built in this way typically need further tailoring in order to be useful in the domain of intended application.

The induction of the semantic lexicon is not straightforward for several reasons. First, there are a large number of possible meanings that could be induced for each phrase in a sentence: forcing Chill to also learn a lexicon as it is learning parsers would result in an exponential increase in the number of hypotheses to consider, as the size of sentence representations increases. Second, many words have multiple senses, or meanings. Therefore, knowing the meaning of a word for some sentences does not guarantee knowing its meaning in all situations. Finally, some words may appear very rarely in a corpus, so that the amount of evidence for inferring the correct meaning for them is quite sparse.

This dissertation presents a lexicon acquisition system, Wolfie, that learns a mapping of words to their meanings, that overcomes many of the limitations of previous work, and that circumvents the difficulties just listed. Although the papers mentioned above and others have also addressed the lexicon learning problem, our method differs from these by combining five features. First, the only background knowledge needed for lexicon learning is in the training examples themselves. Second, interaction with a system, Chill, that learns to parse is demonstrated. Third, a simple, greedy algorithm is used for efficiency to acquire word meanings. Fourth, the system is adaptable to different sentence representations. Finally, the mapping problem is eased by mak-
ing a compositionality assumption that states that the meaning of a sentence is composed from the meanings of the individual words and phrases in that sentence, in addition, perhaps to some “connecting” information specific to the representation at hand. This assumption is similar to the linking rules of Jackendoff (1990).

The Wolfie learning algorithm exploits the compositionality assumption. By assuming that a sentence meaning is composed from word meanings, we also assume that each component of the sentence representation can be mapped back to the meaning of only one word or phrase in the sentence: the one that has that component as part of its meaning representation. Thus, during lexicon acquisition, word meanings are derived from the components of the representations of sentences in which each word appears. The compositionality assumption decreases the number of meanings that need to be considered for each word. While in broad linguistic domains, compositionality may not always hold, we are primarily interested in acquiring language processing information about one domain at a time, rather than a broad-coverage parser. The goal is not to simply obtain a parse, but to obtain one that is useful for achieving some task, such as answering a question or performing a command. In these situations, compositionality often does hold. In addition, since we are able to learn the meanings of multiple-word phrases, some of the arguments against compositionality (e.g., idioms) may be overcome.

We have tested the utility of the lexicon acquisition algorithm on one real-world corpus and two artificial corpora, with positive results. The ability of the learned lexicons to aid a parser acquisition program were measured for the real corpus and for one of the artificial corpora. The real-world corpus contains natural language questions about U.S. geography paired with their logical query representations, that can then be used to extract answers to the questions from a database. We also show the applicability of the technique to three diverse natural languages other than English: Spanish, Turkish, and Japanese, using translations of the same corpus. We compared these results to the results obtained with lexicons learned by a comparable system developed by Siskind (1996), showing a significant improvement for our system in most circumstances. One of the artificial corpora was based on that of McClelland and Kawamoto (1986), and augmented with Conceptual Dependency (Shank, 1975) information. Finally, we used a set of artificially generated corpora to test the scalability of our algorithm, with encouraging results.

While building an annotated corpus is arguably less work than building an entire NLP system, it is still not a simple task. Redundancies and errors can creep into the data. A goal should be to also minimize the amount of data that is annotated, yet still reach a reasonable level of generalization performance with the system learned from training on that data. In the case of natural language, there is frequently a large amount of unannotated text available. We would like to automatically, but intelligently, choose which of the available sentences to annotate.

We do this here using a technique called active learning. Active learning is an emerging research area in machine learning that features systems that automatically select the most informative examples for annotation and training (Cohn, Atlas, & Ladner, 1994). The primary goal of active learning is to reduce the number of examples that the system is trained on, thereby reducing the example annotation cost, while maintaining the accuracy of the acquired information. To demonstrate the usefulness of our active learning techniques, we compared the accuracy of parsers and lexicons learned using examples chosen by active learning to those learned using randomly chosen examples, finding that active learning can save significant annotation cost over training on
randomly chosen examples. This savings is demonstrated in the U.S. geography domain.

In summary, this thesis demonstrates a machine learning technique for inducing semantic lexicons; and by building on previous research an entire natural language interface can be acquired from one training corpus. Further, this thesis demonstrates the application of active learning techniques to minimize the amount of data required to annotate as training input for the integrated learning system.

The remainder of the dissertation is organized as follows. Chapter 2 gives more background information on CHILL, and introduces Siskind’s lexicon acquisition system, which we will compare to WOLFIE in Chapter 4. Chapter 3 formally defines the learning problem and describes the WOLFIE algorithm in detail. In Chapter 4 we present and discuss experiments evaluating WOLFIE’s performance in learning lexicons in a database query domain, and Chapter 5 contains experimental results on two artificial corpora. Next, Chapter 6 describes and evaluates our use of active learning techniques for both CHILL and WOLFIE. Chapters 7 and 8 discuss related research and future directions. Finally, Chapter 9 summarizes our research and results.
Chapter 2

Background

Our approach was developed in the context of learning lexicons to support semantic parsing. As we mentioned in Chapter 1, we use the word semantic lexicon to refer to a mapping from words to representations of their meanings. The particular meaning representation is determined by the domain at hand and the desired form for sentence representations. The initial motivation for learning lexicons was so they could be used to bootstrap a parser acquisition system, CHILL (Zelle, 1995). After CHILL acquires a parser, it also uses the lexicons learned to map novel sentences into representations of their meanings. Section 2.1 describes CHILL in more detail.

The most closely related previous research into automated lexicon acquisition is that of Siskind (1996). As we will be comparing our system to his in Chapter 4, we describe the main features of his research in Section 2.2.

2.1 CHILL

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 the system CHILL. CHILL uses inductive logic programming (Muggleton, 1992; Lavrač & Džeroski, 1994) to learn a deterministic shift-reduce parser written in Prolog. The input to CHILL is a corpus of sentences paired with semantic representations, the same input required by WOLFIE. The output is a deterministic parser that maps sentences into parses.

Figure 2.1 shows the basic components of the CHILL system. First, during Parsing Operator Generation, the training examples are analyzed to formulate an overly-general shift-reduce parser that is capable of producing parses from sentences. Next, in Example Analysis, the overly-general parser is used to parse the training examples to extract contexts in which the generated parsing operators are actually useful. The third step is Control-Rule Induction which employs a general ILP algorithm to learn rules that characterize these contexts. Finally, Program Specialization “folds” the learned control-rules back into the overly-general parser to produce the final parser.

The component of CHILL that will be most relevant to our discussion is the initial overly-general parser. This parser varies from one domain and representation language to the next, but it is a straight-forward matter to determine the type of parsing operators needed to parse sentences into a given representation language. The parser is capable of translating a given sentence into many parses representing its meaning, including the correct one(s). Therefore, to increase efficiency,