Relational Learning Techniques for Natural Language Information Extraction

Mary Elaine Califf

Report AI98-276 August 1998

mecalif@ilstu.edu
http://www.cs.utexas.edu/users/mecaliff/

Artificial Intelligence Laboratory
The University of Texas at Austin
Austin, TX 78712
Relational Learning Techniques for Natural Language Information Extraction

by

Mary Elaine Calif, B.A., M.A., M.S.

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

August 1998
Relational Learning Techniques for Natural Language Information Extraction

Approved by
Dissertation Committee:

RAYMOND MOONEY
RISTO MIKKULAINEN
BRUCE PORTER
BENJAMIN KUIPERS
ELLEN RILOFF
To Michael and Jamie.
Acknowledgments

There are many people whom I wish to thank for their support and for their contributions to this research. First, I would like to thank my advisor, Ray Mooney, for all of his help and guidance throughout my time as a graduate student. Working with Ray has been a continuous learning experience. He was always available to help point me in the right direction and I am very grateful for all of his advice and teaching. I also I would also like to thank the other members of my committee, which include Bruce Porter, Risto Miikkulainen, Ben Kuipers, and Ellen Riloff, for their helpful suggestions and comments.

Next, I would like to thank others who have provided help and encouragement, including William Cohen, Robert van de Geijn, and many of my fellow graduate students, among them John Zelle, James Lester, Tara Estlin, and Cindi Thompson. A special acknowledgement is due to Sue Buchele, with whom I shared all of the trials and joys of being a graduate student and a mother. I would also like to thank Ann Hawkins and Shellynne Wucher for their friendship and support.

Finally, I thank my family: my parents, who raised me to believe I really could do anything I truly wanted to do; my husband, who has followed me, supported me, and taken up the slack left by a harried and distracted wife; and my children, who have not always understood why Mommy had to work on the computer all evening, but have tried to be patient and understanding.

This research was supported by a fellowship from AT&T and by the National Science Foundation under grant IRI-9310819.

Mary Elaine Califf

The University of Texas at Austin
August 1998
The recent growth of online information available in the form of natural language documents creates a greater need for computing systems with the ability to process those documents to simplify access to the information. One type of processing appropriate for many tasks is information extraction, a type of text skimming that retrieves specific types of information from text. Although information extraction systems have existed for two decades, these systems have generally been built by hand and contain domain specific information, making them difficult to port to other domains. A few researchers have begun to apply machine learning to information extraction tasks, but most of this work has involved applying learning to pieces of a much larger system. This dissertation presents a novel rule representation specific to natural language and a relational learning system, Rapier, which learns information extraction rules. Rapier takes pairs of documents and filled templates indicating the information to be extracted and learns pattern-matching rules to extract fillers for the slots in the template. The system is tested on several domains, showing its ability to learn rules for different tasks. Rapier’s performance is compared to a propositional learning system for information extraction, demonstrating the superiority of relational learning for some information extraction tasks.

Because one difficulty in using machine learning to develop natural language processing systems is the necessity of providing annotated examples to supervised learning systems, this dissertation also describes an attempt to reduce the number of examples Rapier requires by employing a form of active learning. Experimental results show that the number of examples required to achieve a given level of performance can be significantly reduced by this method.
Contents

Acknowledgments vi

Abstract vii

Chapter 1 Introduction 1
  1.1 Organization of Dissertation ........................................ 4

Chapter 2 Background 5
  2.1 Information Extraction ............................................. 5
  2.2 Symbolic Relational Learning .................................... 8
    2.2.1 General Algorithm Design Issues ........................... 8
    2.2.2 FOIL ........................................................ 10
    2.2.3 GOLEM ...................................................... 11
    2.2.4 CHILLIN ................................................... 13
    2.2.5 PROGOL .................................................. 14
  2.3 Natural Language Processing Resources ........................... 14
    2.3.1 Part-of-speech tagger ..................................... 15
    2.3.2 Lexicon ................................................... 15

Chapter 3 The Rapier System 17
  3.1 Rule Representation ............................................. 17
  3.2 Learning Algorithm ............................................. 18
    3.2.1 Creation of the Initial Rulebase ......................... 19
    3.2.2 Compression of the Rulebase ............................ 19
    3.2.3 Rule Evaluation Metric ................................ 22
    3.2.4 Computing the Generalizations of Two Patterns ....... 24
    3.2.5 The Specialization Phase ............................... 30
    3.2.6 Example ................................................ 32

Chapter 4 Experimental Evaluation 35
  4.1 Domain and Task Descriptions ................................. 35
  4.2 Data Pre-processing ........................................... 36
  4.3 Ablations ..................................................... 36
  4.4 Systems Compared Against ................................... 36
  4.5 Experimental Methodology .................................... 38
4.6 Job Postings Results .................................................. 39
4.7 Seminar Announcement Results ....................................... 42
4.8 Corporate Acquisitions Results ....................................... 45
4.9 Discussion .............................................................. 46
4.10 Sample Learned Rules ................................................ 48

Chapter 5 Using Active Learning with Rapier .......................... 50
5.1 Selective Sampling ....................................................... 50
5.2 Incorporating Selective Sampling into RAPIER ....................... 51
5.2.1 Uncertainty in RAPIER .............................................. 52
5.2.2 Making RAPIER Incremental ....................................... 52
5.3 Experimental Evaluation .............................................. 54

Chapter 6 Related Work .................................................. 56
6.1 Early Machine Learning Systems for Information Extraction ........ 56
6.2 Learning Information Extraction for Structured Text ............... 58
6.3 Recent Learning for Information Extraction from Semi-structured and Free Text .. 59

Chapter 7 Future Work .................................................... 60
7.1 Additional Test Domains ................................................ 60
7.2 Enhancements to RAPIER’s Representation and Algorithm ........... 61
7.2.1 Additional Constraint Types ....................................... 61
7.2.2 Negated Constraints ............................................... 61
7.2.3 Syntax ............................................................. 61
7.2.4 Domain-Specific Semantics ....................................... 62
7.2.5 Named-Entity Extraction ......................................... 62
7.2.6 Multiple Slot Rules ............................................... 62
7.2.7 Precision and Recall Trade-offs .................................. 63
7.2.8 Boosting and Bagging ............................................. 64
7.3 Experimenting with Active Learning .................................. 64
7.4 Open Issues in Information Extraction ................................ 65
7.4.1 Set-valued Slot Fillers ............................................ 65
7.4.2 Document Relevance .............................................. 65
7.4.3 Recognizing Multiple Case Frames ............................... 66
7.4.4 Template Merging and Coreference Issues ....................... 67
7.5 Extension to Other Natural Language Processing Tasks ............. 67

Chapter 8 Conclusions ..................................................... 69

Appendix A Sample Data ................................................... 71
A.1 Computer-related Jobs ................................................. 71
A.2 Seminar Announcements .............................................. 74
A.3 Corporate Acquisitions .............................................. 76
Chapter 1

Introduction

There has been an explosive growth in the amount of information available on networked computers around the world, much of it in the form of natural language documents. An increasing variety of search engines exist for retrieving such documents using keywords; however, answering many questions about available information requires a deeper “understanding” of natural language. One way of providing more “understanding” is with information extraction. Information extraction is the task of locating specific pieces of data from a natural language document, and has been the focus of DARPA’s MUC program (Lehnert & Sundheim, 1991). The extracted information can then be stored in a database which could then be queried using either standard database query languages or a natural language database interface. An example of the information extraction task which was the focus of MUC-3 and MUC-4 appears in Figures 1.1 and 1.2. The goal was to extract information about Latin American terrorist incidents from news reports.

Information extraction systems seem to be a promising way to deal with certain types of text documents. However, a difficulty with information extraction systems is that they are difficult and time-consuming to build, and they generally contain highly domain-specific components, making porting to new domains also time-consuming. Thus, more efficient means for developing information extraction systems are desirable.

Recent research in computational linguistics indicates that empirical or corpus-based methods are currently the most promising approach to developing robust, efficient natural language processing (NLP) systems (Church & Mercer, 1993; Charniak, 1993; Brill & Church, 1996). These methods automate the acquisition of much of the complex knowledge required for NLP by training on suitably annotated natural language corpora, e.g. treebanks of parsed sentences (Marcus, Santorini, & Marcinkiewicz, 1993).

Most of these empirical NLP methods employ statistical techniques such as n-gram models, hidden Markov models (HMMs), and probabilistic context free grammars (PCFGs). There has also been significant research applying neural-network methods to language processing (Reilly & Sharkey, 1992; Miikkulainen, 1993). However, there has been relatively little recent language research using symbolic learning, although some recent systems have successfully employed decision trees (Magerman, 1995; Aone & Bennett, 1995), transformation rules (Brill, 1993, 1995), and other symbolic methods (Wermter, Riloff, & Scheler, 1996).

Given the successes of empirical NLP methods, researchers have recently begun to apply learning methods to the construction of information extraction systems (McCarthy & Lehnert,
NAWswire text

DEV-MUC3-0011 (NOSC)

LIMA, 9 JAN 90 (EFE) -- [TEXT] AUTHORITIES HAVE REPORTED THAT FORMER PERUVIAN DEFENSE MINISTER GENERAL ENRIQUE LOPEZ ALBUJAR DIED TODAY IN LIMA AS A CONSEQUENCE OF A TERRORIST ATTACK.

LOPEZ ALBUJAR, FORMER ARMY COMMANDER GENERAL AND DEFENSE MINISTER UNTIL MAY 1989, WAS RIDDLED WITH BULLETS BY THREE YOUNG INDIVIDUALS AS HE WAS GETTING OUT OF HIS CAR IN AN OPEN PARKING LOT IN A COMMERCIAL CENTER IN THE RESIDENTIAL NEIGHBORHOOD OF SAN ISIDRO.

LOPEZ ALBUJAR, 63, WAS DRIVING HIS OWN CAR WITHOUT AN ESCORT. HE WAS SHOT EIGHT TIMES IN THE CHEST. THE FORMER MINISTER WAS RUSHED TO THE AIR FORCE HOSPITAL WHERE HE DIED.

Figure 1.1: A sample message from the Latin American terrorism domain used in MUC-3 and MUC-4.

1995; Soderland, Fisher, Aseltine, & Lehnert, 1995, 1996; Riloff, 1993, 1996; Kim & Moldovan, 1995; Huffman, 1996). Several different symbolic and statistical methods have been employed, but most of them are used to generate one part of a larger information extraction system. Our system RAPIER (Robust Automated Production of Information Extraction Rules) learns rules for the complete information extraction task, rules producing the desired information pieces directly from the documents without prior parsing or any post-processing. We do this by using a structured (relational) symbolic representation, rather than learning classifiers.

Using only a corpus of documents paired with filled templates, RAPIER learns Eliza-like patterns (Weizenbaum, 1966) that make use of limited syntactic and semantic information, using freely available, robust knowledge sources such as a part-of-speech tagger or a lexicon. The rules built from these patterns can consider an unbounded context, giving them an advantage over more limited representations which consider only a fixed number of words. This relatively rich representation requires a learning algorithm capable of dealing with its complexities. Therefore, RAPIER employs a relational learning algorithm which uses techniques from several Inductive Logic Programming (ILP) systems (Lavrač & Džeroski, 1994). These techniques are appropriate because they were developed to work on a rich, relational representation (first-order logic clauses). Our algorithm incorporates ideas from several ILP systems, and consists primarily of a specific to general (bottom-up) search. We show that learning can be used to build useful information extraction rules, and that relational learning is more effective than learning using only simple features and a fixed context.

Experiments using RAPIER were performed in three different domains of varying difficulty. In a task of extracting information about computer-related jobs from netnews postings, RAPIER performed quite well, achieving recall of 63% and precision of 89%. RAPIER was compared to a Naive Bayes-based system which looks only at a fixed window before and after the filler (by default,
## Filled Template

<table>
<thead>
<tr>
<th>Slot</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>0. MESSAGE: ID</td>
<td>DEV-MUC3-0011 (MICOSC)</td>
</tr>
<tr>
<td>1. MESSAGE: TEMPLATE</td>
<td>1</td>
</tr>
<tr>
<td>2. INCIDENT: DATE</td>
<td>09 JAN 90</td>
</tr>
<tr>
<td>3. INCIDENT: LOCATION</td>
<td>PERU: LIMA (CITY): SAN ISIDRO (NEIGHBORHOOD)</td>
</tr>
<tr>
<td>4. INCIDENT: TYPE</td>
<td>ATTACK</td>
</tr>
<tr>
<td>5. INCIDENT: STAGE OF EXECUTION</td>
<td>ACCOMPLISHED</td>
</tr>
<tr>
<td>6. INCIDENT: INSTRUMENT ID</td>
<td>-</td>
</tr>
<tr>
<td>7. INCIDENT: INSTRUMENT TYPE</td>
<td>GUN: ‘’-’’</td>
</tr>
<tr>
<td>8. PERP: INCIDENT CATEGORY</td>
<td>-</td>
</tr>
<tr>
<td>9. PERP: INDIVIDUAL ID</td>
<td>‘’THREE YOUNG INDIVIDUALS’’</td>
</tr>
<tr>
<td>10. PERP: ORGANIZATION ID</td>
<td>-</td>
</tr>
<tr>
<td>11. PERP: ORGANIZATION CONFIDENCE</td>
<td>-</td>
</tr>
<tr>
<td>12. PHYS TGT: ID</td>
<td>-</td>
</tr>
<tr>
<td>13. PHYS TGT: TYPE</td>
<td>-</td>
</tr>
<tr>
<td>14. PHYS TGT: NUMBER</td>
<td>-</td>
</tr>
<tr>
<td>15. PHYS TGT: FOREIGN NATION</td>
<td>-</td>
</tr>
<tr>
<td>16. PHYS TGT: EFFECT OF INCIDENT</td>
<td>-</td>
</tr>
<tr>
<td>17. PHYS TGT: TOTAL NUMBER</td>
<td>-</td>
</tr>
<tr>
<td>18. HUM TGT: NAME</td>
<td>‘’ENRIQUE LOPEZ ALBUJAR’’</td>
</tr>
<tr>
<td>19. HUM TGT: DESCRIPTION</td>
<td>‘’FORMER ARMY COMMANDER GENERAL AND DEFENSE MINISTER’: ‘’ENRIQUE LOPEZ ALBUJAR’’</td>
</tr>
<tr>
<td>20. HUM TGT: TYPE</td>
<td>FORMER GOVERNMENT OFFICIAL / FORMER ACTIVE MILITARY: ‘’ENRIQUE LOPEZ ALBUJAR’’</td>
</tr>
<tr>
<td>21. HUM TGT: NUMBER</td>
<td>1: ‘’ENRIQUE LOPEZ ALBUJAR’’</td>
</tr>
<tr>
<td>22. HUM TGT: FOREIGN NATION</td>
<td>-</td>
</tr>
<tr>
<td>23. HUM TGT: EFFECT OF INCIDENT</td>
<td>DEATH: ‘’ENRIQUE LOPEZ ALBUJAR’’</td>
</tr>
<tr>
<td>24. HUM TGT: TOTAL NUMBER</td>
<td>-</td>
</tr>
</tbody>
</table>

Figure 1.2: The filled template corresponding to the message shown in Figure 1.1. The slot fillers include both strings found in the documents and other types of values.

4 words), and does not take into account the order of the words, but only their presence or absence. Tests of this Naïve Bayes-based system on the jobs task produced much worse results: 32% recall at 14% precision. These results demonstrate the value, in some information extraction tasks, at least, of a richer, relational representation, capable of focusing on a single previous word, if appropriate, or of considering six or more context words when needed, and also capable of taking into account the order of the tokens in both the filler and the context. On an easier task of extracting information from seminar announcements, Rapier also performed well, achieving 92% precision and 71% recall overall, although the Naïve Bayes system was more competitive in this domain. Rapier did not perform as well on the third task, which is extracting information about corporate acquisitions from newswire articles. This probably indicates that more syntactic knowledge than Rapier has available is required to successfully handle this type of domain and may also indicate a need for domain-specific semantic information. In all three domains, Rapier is competitive with other state-of-the-art learning systems for information extraction which have been tested on the tasks.
The one difficulty of using machine learning to build systems is the necessity of providing examples to the learning system. In the case of learning information extraction rules, this requires that a human perform the information extraction task on a number of documents. While this is significantly less labor intensive than producing an information extraction system by hand, it is clearly desirable to limit the number of examples required as much as possibly. Therefore, we have experimented with the application of active learning, specifically selective sampling, to limit the number of examples required to achieve a given level of performance. Selective sampling is a method for allowing a learning system to select examples to be annotated and used for training from a pool of unannotated examples. The method attempts to select the most useful examples in order to reduce the number of annotated example required. Experiments with selective sampling in the computer-related jobs domain show that it is possible to greatly reduce the number of examples required to reach the level of performance achieved with the full set of 270 random examples.

This research has focused on two primary goals. First, we show that learning, and, in particular, relational learning, can be used to build practical information extraction systems. Second, we show that selective sampling can be effectively applied to learning for information extraction to reduce the human effort required to annotate examples for building such systems.

1.1 Organization of Dissertation

The rest of this dissertation is organized as follows. Chapter 2 presents background knowledge on information extraction, relational learning, and the natural language processing tools which the Rapier system can use. Chapter 3 describes the representation used by Rapier and presents the learning algorithm. Chapter 4 discusses the experimental evaluation of Rapier on several domains and presents the results of this evaluation. Chapter 5 describes the application of active learning to Rapier and the extensions required to the algorithm to allow for active learning and discusses the experimental evaluation of Rapier using active learning. Chapter 6 describes related work in the area of learning rules for information extraction. Finally, Chapter 7 suggests ideas for future research directions, and Chapter 8 reviews the ideas and results presented in this dissertation and discusses relevant conclusions.
Chapter 2

Background

The first part of this chapter discusses the information extraction task and characteristics of traditional information extraction systems. The second section discusses relational learning and some the design choices involved in developing a relational rule learning system. That section also describes several previous relational rule learning systems, all of them inductive logic programming systems, including the three systems that directly influenced the development of Rapier. The final section of the chapter briefly describes the natural language processing resources used in this research.

2.1 Information Extraction

Information extraction is a shallow form of natural language understanding useful for certain types of document processing, which has been the focus of ARPA’s Message Understanding Conferences (MUC) (Lehmert & Sundheim, 1991; DARPA, 1992, 1993). It is useful in situations where a set of text documents exist containing information which could be more easily used by a human or computer if the information were available in a uniform database format. Thus, an information extraction system is given the set of documents and a template of slots to be filled with information from the document. Information extraction systems locate and in some way identify the specific pieces of data needed from each document.

Two different types of data may be extracted from a document: more commonly, the system is to identify a string taken directly from the document, but in some cases the system selects one from a set of values which are possible fillers for a slot. The latter type of slot-filler may be items like dates, which are most useful in a consistent format, or they may simply be a set of terms to provide consistent values for information which is present in the document, but not necessarily in a consistently useful way. An example of this is in the Latin American terrorism domain used in MUC-3 and MUC-4 (see Figures 1.1 and 1.2), where an incident may be “THREATENED,” “ATTEMPTED” or “ACCOMPLISHED.”

The data to be extracted may be specified in either of two ways. The system may fill a template with the values from the document, or, in the case where all slots are filled by strings directly from the document, the system may annotate the document directly.

Information extraction can be useful in a variety of domains. The various MUC’s have focused on tasks such as the Latin American terrorism domain mentioned above, joint ventures,
Position available for Software Programmer experienced in generating software for PC-Based Voice Mail systems. Experienced in C Programming. Must be familiar with communicating with and controlling voice cards; preferable Dialogic, however, experience with others such as Rhetorix and Natural Microsystems is okay. Prefer 5 years or more experience with PC Based Voice Mail, but will consider as little as 2 years. Need to find a Senior level person who can come on board and pick up code with very little training. Present Operating System is DOS. May go to OS-2 or UNIX in future.

Please reply to:
Kim Anderson
AdNET
(901) 458-2888 fax
kimander@memphisonline.com

Figure 2.1: A sample job posting from a newsgroup.

Microelectronics, and company management changes. Others have used information extraction to track medical patient records (Soderland et al., 1995) and to track company mergers (Huffman, 1996). More recently, researchers have applied information extraction to less formal text genres such as rental ads (Soderland, 1998) and web pages (Freitag, 1998a; Hsu & Dung, 1998; Muslea, Minton, & Knoblock, 1998).

Another domain which seems appropriate, particularly in the light of dealing with the wealth of online information, is to extract information from text documents in order to create easily searchable databases from the information, thus making the wealth of text online more easily accessible. For instance, information extracted from job postings in USENET newsgroups such as misc.jobs.offered can be used to create an easily searchable database of jobs. Such databases would be particularly useful as part of a complete NLP system which supported natural language querying of the system. The work on information extraction reported in this dissertation is part of an ongoing project to develop such a system. The initial system handles computer-related jobs only. An example of the information extraction task for the system appears in Figures 2.1 and 2.2.

The architecture of the complete system is shown in figure 2.3. In addition to the information extraction rules learned by RAPIER, the system requires a query parser and a semantic lexicon for the query parse. The query parser is being developed using CHILL (Zelle & Mooney, 1996), a system which learns parsers from example sentences paired with their parses, and the semantic lexicon is